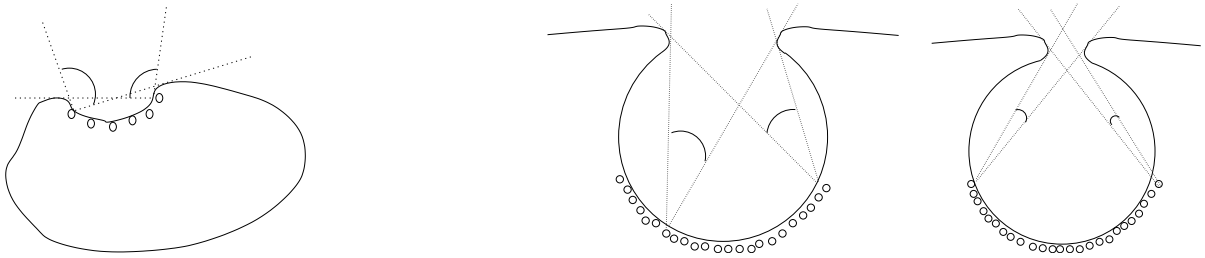
[These lecture notes do not include the introductory lecture (0).]

The origins of spatial vision

Our eyes are very sophisticated optical devices, and it took millions of years of evolution to reach this level of sophistication.[1] It is not known exactly how eyes evolved, but it is believed that the earliest eyes consisted of a small number of light sensitive cells distributed over a small region on the outer surface of an animal (and hooked up to a primitive nervous system). Let's suppose the cells are distributed over a concave pit such as in the figure below. Six light sensitive cells are shown. Because the pit is concave, each cell will receive light from a limited set of directions. For the leftmost and rightmost cells, the range of directions of light coming from the scene is shown in the figure on the left. (See slides for updated figures.)



Now suppose that something to the left of this animal were to move towards the animal and block the partially block the sky (casting a shadow). The leftmost cell only receives light from above/right and so the light received by the cell would not be affected. (See sketch in lecture slides). The rightmost cell receives light from above/left and so the approaching animal would block the skylight and the rightmost cell would received less light. The changing measurement would tell the animal that something dark is now presen on the left. A defensive response of the animal therefore might be to move toward the right, i.e. away from the approaching animal. (Alternatively, the animal could move to the left, which would be a more aggressive response.) Such a response might allow the animal to survive. If it produced offspring with similar concave light-sensitive regions which produced similar actions to changing light measurements, they might also have a better chance of surviving.

One way to improve this vision system would be to make the eye more cavelike, as in the two figures above to the right. Here we are reducing the *aperture* by which light can enter the concavity. This reduces the angle of incident light that reaches each photoreceptor cell of the eye.

The advantage of reducing the aperture is that each cell receives light from a more restricted set of directions which provides the eye with more detailed information about the directional distribution of light arriving at the aperture. The understand this, consider the figures below. The pattern of light in the scene is an alternating dark grey and light grey, and each cell averages over some dark regions and some light regions.

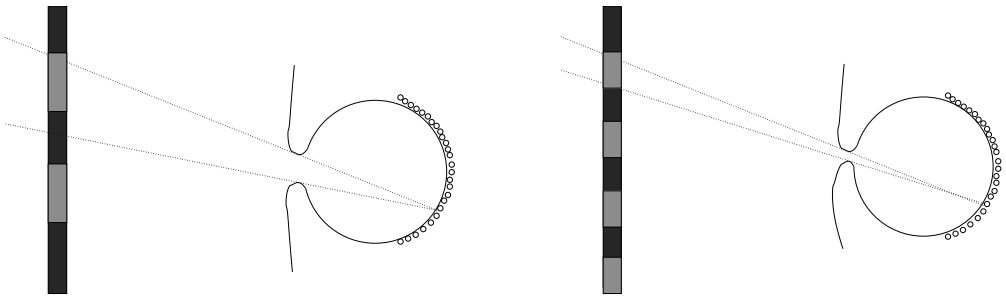
In the figure on the left, the aperture is so big that each cell would see part of a light grey region and part of a dark grey region. If the aperture were slightly larger, so that each cell received light

[1] See interview with Richard Dawkins (outspoken evolutionary biologist at U. Oxford) `https://www.youtube.com/watch?v=bwX3fx0Zg5o`

 ©Michael Langer

from an equal amount of light and dark regions, then each cell would receive the same total amount of light. In this case, we would say that the light and dark greys had been *blurred* away completely.

In the figure on the right, the aperture has been decreased, and the pattern of light and dark grey has a higher frequency. This figure is drawn such that the angular width of the aperture is exactly matched to the width of the pattern on the surface. Notice that the retinal image would still be blurred, since most cells would receive a mix of light from light and dark regions and only a few cells would see only a light gray or only a dark grey. The retinal image would look more like a sinusoid with smooth transition from dark to light, rather than the piecewise constant intensity (light,dark,light,..) in the scene.



The disadvantage of using a smaller aperture is that it reduces the amount of light reaching each cell. Eventually, if the aperture becomes a "pinhole" and the image will be very dark indeed. Next lecture I will discuss how lenses avoid this problem.

Units of angle

We will talk about angles in various ways today. As you know from Calculus, it is common to define an angle in units of degrees or radians. Recall that 2π radians is 360 degrees, or

$$\frac{360 \text{ degrees}}{2\pi \text{ radians}} = \frac{180 \text{ degrees}}{\pi \text{ radians}} \approx 57 \text{ degrees per radian.}$$

When doing vision calculations, it is common to make a small angle approximation.

$$\theta \approx 2 \tan(\frac{\theta}{2}).$$

This approximation essentially says that the length of an short arc of a circle is approximately equal to the length of the line segment joining the end points of that arc of a circle.

In astromony, it is necessary to refer to angles that are much smaller than a degree. In particular, there are 60 *minutes* in one degree, and there are 60 *seconds* in one minute. Minutes of arc comes up often in vision. Rarely do we need to talk about seconds, but it does come up.

Aperture angle and f-number

Let's returning to our discussion of apertures. The notion of an aperture should be familiar to those of you who dabble in photography. Let's consider a camera rather than an eye. The camera is a hollow box and light enters the box through a hole.

 ©Michael Langer

Ignore the lens for today and just think about the hole or aperture. Let A be the diameter of the aperture and let f be the distance from the center of the aperture to the sensor surface. Then, if $A \ll f$ which is usually the case, then we can make a small angle approximation, namely A/f is approximately the angle subtended by the aperture as seen from a point on the sensor on the image plane at the back of the camera box.

The amount of light reaching a point on the image plane depends on what is visible in the 3D scene in directions "seen" by that pixel. It also depends on the angular size of the cone of light rays that reaches this image point. To think about angles, it is best for now to just consider a 2D scene rather than 3D scene. So, the camera is a square and the aperture is gap in the square and the sensor is a line. (See figure in slides.) The angle subtended by the aperture is approximately $\frac{A}{F}$ radians, where A is the width of the aperture and F is the distance from the aperture to the sensor. Its inverse, $\frac{F}{A}$ is called the *F-number*.

If you have done any photography using an SLR camera, then you are familiar with f-number, as it is one of the main parameters you can manipulate. When you change the f-number, in fact you are just changing the aperture since f is fixed. The effect of course is to change the amount of light that reaches the lens, making the image brighter or darker. There are other effects as well, as we'll see once we consider lenses.

What are some typical f-numbers? Camera's often have f-numbers[2] ranging from about 2 to 16. For example, if $A = 5mm$ and $F = 50mm$, then the f-number $\frac{F}{A}$ is 10. A small angle approximation works quite well here. We can also define an f-number for a human eye. Typical values of the aperture (pupil) diameter are $A = 5mm$ and $F = 25mm$, for an f-number of 5. A small angle approximation still holds for these values too.[3]

Visual angle

A second fundamental angular quantity is the angle subtended by an object in the 3D world as seen from a position in space. This angle is called the *visual angle* subtended by this object. Assuming that the object's height (or width) is small compared to the distance to the object, we can make a small angle approximation and define:

$$\text{visual angle (radians)} = \frac{\text{height of object}}{\text{distance to object}}$$

Let's suppose that the aperture angle is very small (large f-number) and treat the aperture as a point in space. This is usually called a *pinhole camera.* From high school geometry reasoning – namely, opposite angles are equal – we know that the visual angle subtended by the object can be written equivalently as:

$$\text{visual angle (radians)} = \frac{\text{height of image (of object) on sensor}}{\text{distance from pinhole to sensor}}$$

For example, consider your thumbnail which is about 1 cm wide. Suppose you view your thumb at an arm's length distance say about $57 = \frac{180}{\pi}$ cm. The thumbnail would have a visual angle of

[2]Technically it is the camera and the lens together that define the f-number since the aperture is defined in the lens body.

[3]You may be asking yourself how you know when a small angle approximation is "good enough". There is no single answer to this. It really depends on the precision you need.

 ©Michael Langer

$\frac{1}{\frac{180}{\pi}} = \frac{\pi}{180}$ radians. Converting to degrees by multiplying by $\frac{180}{\pi}$ degrees/radian gives us 1 degree, i.e. a thumbnail at arm's length subtends about 1 degree of visual angle.

Here is a second example. Consider a person's head a large distance, say 18 m. Suppose the person's head is about 30 cm high. To make the calculation easier, say it is 31.4 cm high, or $\frac{\pi}{10}$ m. Then the visual angle subtended by this person's head would be about $\frac{\pi/10}{18}$. Converting to degrees by multiplying by $\frac{180}{\pi}$ gives 1 degree, i.e. the visual angle of the person's head at that distance would be 1 degree.

A third example of a visual angle is the moon which subtends about half a degree, or 30 minutes (arcmin). You may have heard of the moon illusion, which is that the moon appears much bigger when it is near the horizon than when it is overhead. This is not an optical effect due to bending of light through the atmosphere, as some people assume. Rather, it is a perceptual effect. This illusion is very strong and has been studied in great detail literally for centuries. (Read the wikipedia article if interested.)

Image position

We would like to define positions of points in an image and relate these image positions to positions in the 3D scene. Define a coordinate system with axes XYZ such that the origin $(X, Y, Z) = (0, 0, 0)$ is at the center of the camera/eye aperture. We'll just ignore the aperture itself for the rest of today and assume a pinhole camera. Let $(\hat{X}, \hat{Y}, \hat{Z})$ be the coordinate axes. Let the Z be depth variable and XY be the axes parallel to the image plane. Typically X is the right right and Y is up. The $\hat{Z}$ axis is called the *optical axis*.

Let's next relate positions XYZ in the 3D scene with positions on the image plane. We begin by reviewing some of the basic geometry of image formation. Consider a 3D scene point (X_0, Y_0, Z_0) in this coordinate system. Suppose that the image plane is behind the camera at a distance f. The line through this scene point and through the origin (pinhole) intersects the *image plane* $Z = -f$ at position $(x, y, -f)$. The point of intersection is the *image position*. So, what we have just done is project the 3D point onto the image plane.

Using high school geometry (similar triangles), we can see that

$$\frac{x}{f} = \frac{X_0}{Z_0}, \qquad \frac{y}{f} = \frac{Y_0}{Z_0}$$

Note that if (x, y) is close to the center of the image, then we can make a small angle approximation and talk about $\frac{x}{f}$ and $\frac{y}{f}$ as angles (in radians).

Notice that the image will be upside down and backwards, in the sense that if the object points has $X_0 > 0$ then $x < 0$ and if $Y_0 > 0$ then $y < 0$. It can be confusing to think of images that are upside down and backwards. Typically instead one thinks of the images as upright, as we discuss next.

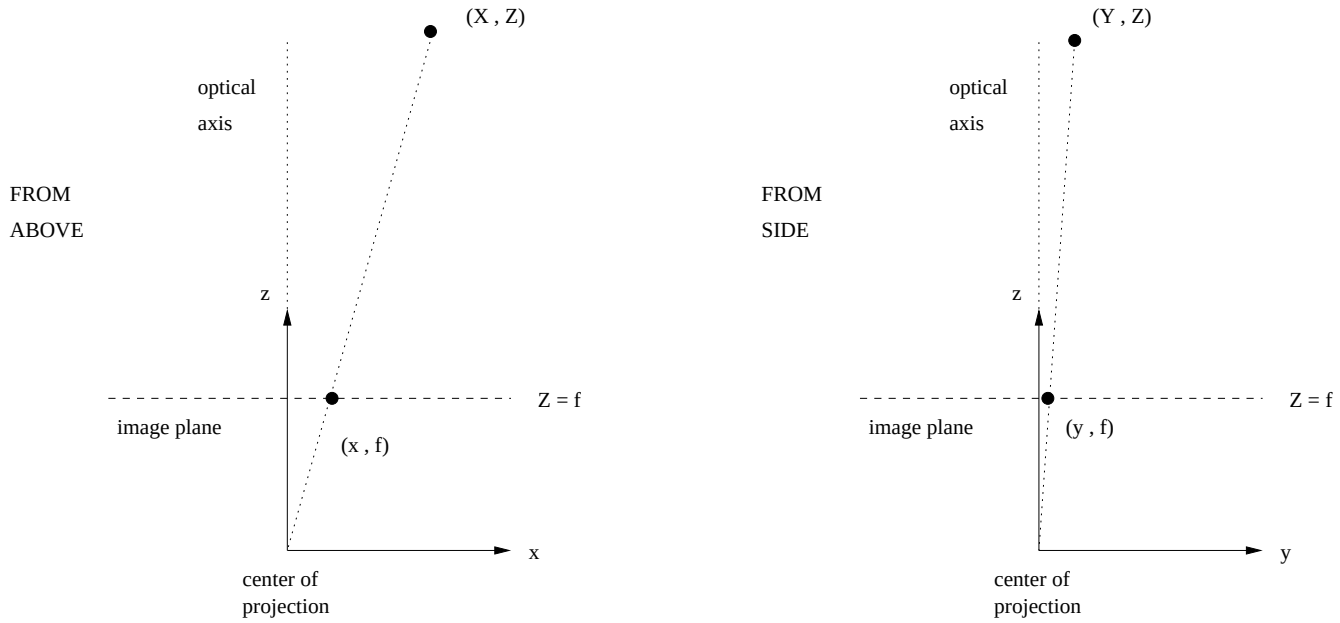
Visual direction

Consider a plane at $Z = f$ in front of the eye or camera. This is not a real image plane, but it is still useful to think about. Just as we projected the 3D scene though a pinhole and onto an image

 ©Michael Langer

plane behind the camera, we can project the 3D scene towards a pinhole and consider where the image point intersects the plane $Z = f$ in front of the camera.

$$(x, y) \;=\; f(\frac{X_0}{Z_0}, \frac{Y_0}{Z_0}). \tag{1}$$



Such a point defines a visual direction from the eye/camera out to the scene. With these (x, y) coordinates, if $X > 0$ then $x > 0$ and $Y > 0$ then $y > 0$. We will typically use this coordinate system (x means to the right, y means up). Moreover, we are typically interested in visual direction rather than some position in an arbitrary plane $Z = f$. We will refer to visual direction $(\frac{x}{f}, \frac{y}{f})$. If these values are small, then they are approximately angles in radians.

Depth map

For every image position (x, y, f) in this abstract image plane in front of the camera, there is typically one surface point that is visible along the ray from the center of projection through that position. The function $Z(x, y)$ maps each position in the (x, y) projection plane to a depth in the world. This function is called the *depth map*.[4]

Notice that the *depth* is not the Euclidian distance $\sqrt{X^2 + Y_2 + Z^2}$ to the 3D point (X, Y, Z). Rather we are only considering the Z value. If we are looking a wall that directly in front of us, then all points on the wall would have the same depth, even though the Euclidian distance would vary along the wall.

[4]Note that we could alternatively define the depth map to be a function of visual direction $Z(\frac{x}{f}, \frac{y}{f})$.

Example: Ground plane

Consider a specific example of a depth map. Suppose the only visible surface is the ground, which we approximate as a plane. Suppose the camera/eye is height h above this *ground plane*. That is, the ground plane is

$$Y = h$$

where $h < 0$. We are still assuming the camera is pointing in the Z direction.

What is the depth map of the ground plane? From Eq. (1), we substitute $-h$ for Y:

$$(x, y) \;=\; f(\frac{X_0}{Z}, \frac{h}{Z}).$$

In particular,

$$Z(x, y) \;=\; \frac{-hf}{y} \tag{2}$$

Thus, the depth map $Z(x, y)$ does not depend on x. It only depends on y. When $y = 0$, we have $Z = \infty$. This is the *horizon*. When $y < 0$, we have $Z > 0$. (Note that $f > 0$.) These are the visible points on the ground and we see that closer points to the eye (smaller Z) have more negative y. What about points where $y > 0$. These are not points on the ground, and their depths are not defined in Eq. (2). If there is nothing in the scene other than a ground plane, then the points where $y > 0$ would be the sky. We could take the depths at $y > 0$ to be infinity.

Note that the depth map for a groud plane only depends on y, and not on x. For any fixed y, all visible points along that horizontal image line have the same depth (independent of x). Also, points of a fixed depth $Z = Z_0$ all project to the same y value. Again, $\frac{y}{f}$ is the angle of a point below the visual horizon. This angle varies inversely with depth.

Binocular disparity

Having two eyes gives us two slightly different views of the world, and the slight differences provide information about depth. Let's begin by assuming that we have two eyes or cameras and that the optical axes of the two eyes are parallel, i.e. the eyes have the same Z direction. Let the right eye be positioned at point $(0, 0, T_X)$ in the left eye's coordinate system. The distance T_X is sometimes called the *interocular distance*. With these assumptions, a 3D point with coordinates (X_0, Y_0, Z_0) in the left eye's coordinate system would have coordinates $(X_0 - T_x, Y_0, Z_0)$ in the right eye's coordinate system. As such, this 3D point would project to a different x value in the left and right images. The difference in x position of that point is called the *binocular disparity*. In human vision, it is more common to define it in terms of visual direction, so that's what we will do.

$$\text{disparity (radians)} \;\equiv\; \frac{x_l}{f} - \frac{x_r}{f} \;=\; \frac{X}{Z_0} - \frac{X - T_x}{Z_0} \;=\; \frac{T_x}{Z_0}.$$

Note that we are assuming the eyes are parallel and pointing forward. In this case, the y values of the projected points are the same in the two eye images.

 ©Michael Langer

Vergence

To visually explore the world around us, we rotate our eyes namely we point the optical axes of each eye at a particular position in 3D space. The 3D point that we "look at" is typically not at infinity, and so the two eyes and the 3D point form a triangle. We say that the eyes *converge* on this 3D point, and the angle of the triangle at the 3D point the eyes are looking at is the *vergence angle*.

Rotating (verging) the eyes changes the binocular disparity. If we look at a point with the two eyes, it means that we are rotating the eyes such that we set (x_l, y_l) and (x_r, y_r) to $(0, 0)$. In particular, the 3D point that the eyes are looking at will have zero disparity. Other points will have positive or negative disparity depending on whether these points are in front of or behind the 3D point we are looking at. We will discuss this in more detail in a few weeks. For now let's sketch out just the basics of how this works. See the accompanying slides which show a sketch of two people, a ground plane, and a horizon.

Let the left and right eyes rotate to the left or right by angles θ_l and θ_r radians. Note that we assume the rotation is left and right, i.e. about the Y axis. Suppose that this 3D point had horizontal angular directions x_l and x_r in the left and right eye when the eye's were pointing straight ahead. Since the eyes have been rotated by θ_l and θ_r respectively, these rotations bring the binary disparity of this 3D point to $\frac{x_l}{f} - \theta_l$ and $\frac{x_r}{f} - \theta_r$ in the two eyes, respectively. This would change the disparity to

$$\begin{aligned} \text{disparity (radians)} &= (\frac{x_l}{f} - \theta_l) - (\frac{x_r}{f} - \theta_r) \\ &= (\frac{x_l}{f} - \frac{x_r}{f}) - (\theta_l - \theta_r). \end{aligned}$$

Thus, the effect of rotating the eyes horizontally (left/right) is to shift *all* the points in each image by angles θ_l and θ_r in the left and right eyes respectively. This changes the disparity of all points by a constant $\theta_l - \theta_r$.

 ©Michael Langer

Image sampling and resolution

The visual world around us contains an enormous amount of detail. We can see individual blades of grass at a distance of several meters, and we recognize faces at distances of tens of meters. The photoreceptor cells in our eyes *sample* the images, and our ability to see details depends in part on this sampling. It also depends on the defocus blur in the images, which we will discuss shortly.

What is the number of samples per degree of visual angle of our eyes or of a camera? To calculate this sampling rate, we need to know the distance s between samples on the image sensor, and the distance f of the image sensor from the center of the aperture. The angular distance between samples is then s/f radians. The sampling rate is the number of samples per angle which is the inverse of the distance between samples, i.e. f/s.

For example, consider some camera that has a given number of pixels on its sensor, say 3000×2000 and suppose the sensor area were $30mm \times 20mm$. One could calculate the distance between pixels from these values. If the distance from the sensor to a small aperture were given (and pretending there was no lens) then one could calculate the number of samples (pixels) per radian or per degree. See the Exercises for some examples.

Blur due to finite aperture

Up to now we have only considered eyes that are formed by having a concave surface. In the case of an extreme concavity, light enters only through a small aperture. The image that is formed will have limited sharpness because each sensor point will receive light from a cone of directions and these directions come from different 3D points in the world. The image sensor will average together the intensity values of the rays coming from these different 3D points.

Another way to think about blur is to note that each 3D point in the world will send rays of light to different points on the image sensor. That is, rather than thinking about each sensor receiving light from many different points in the 3D scene, we can think of each 3D scene point as sending light to many different sensor points. These two ways of thinking about blur just differ in what we are holding "fixed", either a single 2D sensor point or a single 3D scene point. Both are valid ways of thinking about blur. We will see this concept pop up several times as we study vision.

Thin lens model

Let's now consider lenses in the eye. A lens changes the direction of incoming light rays. Lenses allow some 3D scene points to produce a focussed image despite the eye having a finite size aperture. If you need a refresher on how lenses work, see
`https://www.khanacademy.org/science/physics/geometric-optics/`
If you want to learn more about how lenses evolved, see `http://www.youtube.com/watch?v=mb9_x1wgm7E` (Richard Dawkins video)

We restrict our discussion to the *thin lens* model. You may have seen the derivation of this model in your high school or your freshmen physics course. A key assumption of the thin lens model is that, for any 3D object point (X_o, Y_o, Z_o) in the world, the light rays that diverge from this point and that pass through the lens all will converge at some image point (X_i, Y_i, Z_i) behind the lens. Such points are called *conjugate pairs*. Using simple geometric arguments, one can derive relationships between the X, Y, Z variables of conjugate pairs. For example, you may recall that

 ©Michael Langer

rays that are parallel to the optical axis (Z axis) and that pass through the lens will then pass through the optical axis at the point $Z = f$. The constant f is called the *focal length* of the thin lens. The focal length depends on the curvature of the two faces of the lens and on the material of the lens i.e. the "index of refraction" which has to do with whether it is made of water, glass, etc. The inverse of f, i.e. $1/f$ is called the *power* of the lens. [ASIDE: Note we have changed the definition of the variable f and the term "focal length". The definition from the last lecture (which is used in computer vision) was based on a pinhole camera model.]

The above property about parallel rays allows one to derive (details omitted) the following, called the *thin lens equation*:

$$\frac{1}{Z_o} + \frac{1}{Z_i} = \frac{1}{f}$$

which you should have seen in your Physics 1xx courses. The case of parallel rays is a special case in which the object is very far away from the lens, i.e. $Z_o \approx \infty$). Taking $Z_o = \infty$, and plugging into the thin lens formula gives $Z_i = f$. This holds for any set of incoming parallel rays, as long as the direction is not too far from the optical axis.

One way to think of the thin lens model is that if we have an object in the scene at some distance Z_o, then the image of that object will be at some distance Z_i behind the lens. But we can think of the thin lens equation in the opposite way too. Suppose we have an image sensor plane that is a distance Z_{sensor} from the center of the lens. The thin lens models that say points on the sensor plane have a set of conjugate points on a scene plane, called the *focal plane* which is at depth $Z_{focalplane}$ such that:

$$\frac{1}{Z_{focalplane}} + \frac{1}{Z_{sensor}} = \frac{1}{f}.$$

Example

Suppose your eye is focused on an object that is a distance of 10 m away and you hold up your finger at arm's length. Assume Z_{sensor} is 2 cm (the length of your eye) and suppose the aperture ("pupil") is 3 mm. What will be the blur width of your finger?

We apply the thin lens equation twice – once for the focal plane at 10 m :

$$\begin{aligned} \frac{1}{f} &= \frac{1}{Z_o} + \frac{1}{Z_{sensor}} \\ &= \frac{1}{10} + \frac{1}{.02} \end{aligned}$$

and once for the finger:

$$\frac{1}{f} = \frac{1}{.57} + \frac{1}{Z_i}$$

This gives $\frac{1}{Z_i} = 48.1$ and so $Z_i \approx .0207$. Thus the image of the finger is focussed slightly beyond the sensor, which causes blur on the sensor.

To compute the blur width w, we use similar triangles:

$$\frac{A}{Z_i} = \frac{w}{Z_i - Z_{sensor}}$$

so

$$\frac{.003}{.0207} = \frac{w}{.0007}$$

which gives $w \approx .0001$ m.

The blur width w spans some distance on the sensor surface. What is the visual angle covered by this blur width? It might not be clear what this question means, since visual angle was defined last lecture for pinhole cameras only and here we obviously don't have a pinhole camera. The way to think about it to ask what the visual angle *would be* for that distance if we *were* to have a pinhole camera. The answer is:

$$\text{blur width (radians)} = \frac{w}{Z_{sensor}} = \frac{.0001}{.02} \text{ radians} \approx \frac{1}{4}\text{degrees}$$

Recall that the angular width of your finger at arm's length is about 1 degree. So in this example, the blur width is about $\frac{1}{4}$ of the finger width. [**April 23 (edit). This seems like a lot! I will doublecheck with a camera and report back. We can't always trust our eyes on these things.**]

In the exercises, I ask you to show that the blur width for a point at depth Z_o is

$$A(\ |\ \frac{1}{Z_o} - \frac{1}{Z_{focalplane}}\ |\ ).$$

Note from this expression that the image is in perfect focus when $Z_o = Z_{sensor}$ and that the blur increases linearly with the distance in diopters from the focal plane.

Depth of Field

If a scene has a range of depths, then it is impossible for all points in the scene to be in perfect focus. Only one depth is in perfect focus. That said, our vision systems are limited in how well they can *detect* defocus blur, since we have a finite grid of photoreceptors. Some points that are out of focus will still appear perfectly focussed to us. So, a range of depths appears in perfect focus, and it is useful to give this range a name: the *depth of field* is the range of depths that are *perceived* to be in focus. The term is most often used in photography where one is describing a captured image, but it can be used in vision too. Note that this range of depths straddles the single depth that is in perfect focus. Some points closer than the focal plane and further from the focal plane appear to be in focus.

The typical depth of field that is quoted for human vision is 0.3 diopters (D). This means that the range of depths $[Z_{near}, Z_{far}]$ that *appears* in perfect focus at any one time typically satisfies about

$$\frac{1}{Z_{near}} - \frac{1}{Z_{far}} = 0.3.$$

The value 0.3 is only ballpark, however, and it depends on the pupil size, the individual person, , and the scene. Different scenes produce different image patterns, and for some of these patterns blur might be easier or harder to detect.

For example, the following depth intervals each have a difference of about 0.3 diopters (see slides for illustration):

- $[3.3m, \infty m]$ or $[.3D, 0D]$
- $[2m, 5m]$ or $[0.5D, 0.2D]$
- $[1m, 1.43m]$ or $[1D, 0.7D]$

Because blur increases linearly with diopter distance from the focal plane, the focal plane for these examples is at $0.85D, 0.35D$, and $0.15D$. Note that these focal plane distances are not the halfway point in the above interval in meters, but rather they are the halfway point in inverse meters (diopters).

Accommodation

The *power* of a lens is defined as the reciprocal of the focal distance f of that lens. When an optical system has a sequence of lenses that have focal lengths say f_1, f_2, etc, the combined power of the lenses is approximately the sum of the powers, $\frac{1}{f_1} + \frac{1}{f_2} + etc$. In the eye, there are two refracting elements: the lens and the cornea. The cornea (the hard protective surface that interfaces with air) has more power than the lens. The high power of the cornea is due mainly to the large difference in index of refraction between the cornea and air. A typical power for the cornea is 40, i.e. a focal length 1/40.

The cornea is hard and doesn't change shape, and so it is lens that allows the focal plane (or power of the eye) to vary. There are muscles in the eye that can squeeze the lens, causing the lens curvature to increase, which decreases the focal length. So how does changing the lens power affect the focal plane? Consider the thin lens equation

$$\frac{1}{f_{cornea}} + \frac{1}{f_{lens}} = \frac{1}{f_{focalplane}} + \frac{1}{Z_{sensor}}$$

where the left hand side is the combined power of the cornea and lens. The power of the corner and the distance Z_{sensor} are fixed, and so there is only one varying element on the left and right side, namely by changing the power of the lens, you change the focal plane distance.

As we age, our lens becomes more rigid. The effect is huge. A ten year old child can change the power of the lens over a range of 15 diopters, but this range steadily decreases as one ages up to about 50 where the range is reduced to a mere 1 diopter. So beyond the age of 50, one still can accommodate a bit, but not much. This is a well known and universal effect of aging, and it is called *presbyopia*.

Another problem which many of you are more familiar with is short sightedness (*myopia)* versus long sightedness (*hyperopia*). Myopia means that the lens is too powerful relative to the size of the eye, and so rays coming from distant objects tend to be focussed in front of the sensor surface, and the rays then diverge before they reach the surface, creating blur. To counter myopia, one wears glasses that have negative power. For example, I am myopic and my prescription is about -3 diopters. When I wear my glasses, the power of my eye is,

$$\frac{1}{f_{cornea}} + \frac{1}{f_{lens}} + \frac{1}{-3}.$$

How does my optometrist decide what prescription I need? Since my problem is that I cannot see distant object clearly, I need glasses that correct my vision to allow me to see objects at a

distance up to infinity. Note that this does not mean that I need to be able to focus at infinity. It is enough that I can focus up to about 0.15 diopters from infinity i.e. $1/0.15 \approx 7$ meters, since my depth of field will allow me to see clearly those last 0.15 diopters beyond 7 m.

People who are hyperopic have the opposite problem. They need optical corrections to see objects that are nearby. They need to add power, rather than subtract power. For example, if they want to see something clearly at a distance of say 20 cm or 5 D (diopters), they need the near end of the depth of field to be at that distance. Strictly speaking, this means that is enough for them to be corrected a distance of slightly greater than 20 cm since their depth of field will give them clear vision in an interval around 20 cm. However, note that at this distance, tiny changes in depth lead to relatively large changes in blur (for fixed focal length). For example, a 0.3 diopter range 'centered' at 20 cm is $[4.85D, 5.15D]$ which corresponds to a mere $[20.6cm, 19.4cm]$ only!

Open questions

We have reviewed some of the basic models and concepts of blurring that results from defocus. But we have just talked about image formation here, not about vision. Here are few questions that we would like to be able to answer about vision:

First, How does a visual system determine if an image is in focus ? When I take my glasses off while lecturing the scene in front of me is very blurry. This is obvious, and there is nothing I can do about it other than put my glasses on. But objects that are close to me might be only slightly out of focus at any given time. But how can I tell that? How can I characterize the image I am measuring as blurred? Similarly, when you look at a photograph, how can you decide if it is sharp or blurred? What properties of the image make it sharp?

Second, how does the visual system accommodate ? If a scene around us is slightly blurred, then we need to adjust the focal length of our lens, i.e. we need to accommodate. But how? Should we focus at a farther or closer distance? (That is, is the object we are looking at blurred because it is too close or too far from where we are focusing?)

Third, a related question: how does accommodation interaction with binocular vergence ? If I rotate my eyes in to look at something closer to me, then I should increase the power of my lens too, so that the object will be in focus. Do the accommodation and vergence systems "talk to" each other. It turns out that they do.

Fourth, is defocus blur a depth cue ? When I change my focus to make the image of some object sharp, am I getting depth information about that object? Or am I just making it more sharp? We'll return to these questions later in the course.

Today we will examine the measurment of light by photoreceptor cells in the eye. We will look at how sensitive photoreceptors are to different wavelengths of light and also how sensitive they are too different levels of light e.g. night versus day. We will examine how different 'colored' lights can be distinguished, and touch on phenomena such as color blindness.

Light spectra

A good place to start is with Isaac Newton and his prism experiment (late 17th century). Newton observed that when a beam of sunlight is passed through a prism, the beam is spread out into a fan of different color lights – like a rainbow. He argued based on this experiments that light from common radiating source such as the sun or a candle flame is composed of a mixture of colors. The theory that explains Newton's experiments and many other optics experiments has come a long way since then. In a nutshell, we now know that light just is electromagnetic waves, with wavelengths ranging from 400-700 nanometers. (A nanometer is 10^{-9} meters. Thus, you need about 2000 wavelengths of light to extend a distance of one millimeter.)

For any beam of light, we can write the distribution of power in that beam as a function of wavelength. More generally, any function of wavelength can be referred to as a *spectrum*. The light emitted from a source (sun, light bulb, candle) has an *emission* spectrum. A surface that is illuminated has a *reflectance* spectrum, which specifies for each wavelength what is the fraction of light that arrives at the surface that is reflected. Note that light that arrives at a surface doesn't change its wavelength upon reflection. Rather, for each wavelength, some is reflected and the rest is absorbed or transmitted through the medium.

Transmission and absorption spectra are both important in vision. Transmission spectra arise in the context of filters, for example, red and cyan filters that are a cheap way to view 3D images. (More on this later.) Such filters also reflect light and absorb light. Typically we are concerned with how well they transmit light, rather than how much they absorb versus reflect. *Absorption* spectra are especially relevant for understanding photoreceptors, which we discuss next.

Photoreceptors: Rods and cone

There are two general classes of photoreceptors in the human retina. One class is specialized for discriminating between very low levels of light (night vision). The receptors are long and thin, and are called *rods*. The second class is specialized for discriminating between high light levels (day vision), and also between different spectra. These are called *cones* since their shape is conical.

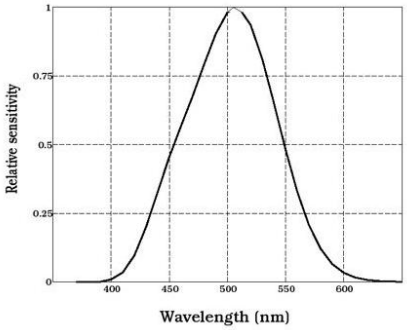
At very low light levels, namely at night when there is only moonlight, only the rod system is functioning. All rods have the same spectral sensitivity, and so there is no way to compare the spectral distributions of light at two different parts of the retina. Thus at night, we only see shades of grey from black to white. During the day when light levels are high, our rod system shuts down and only the cones are operating, and we can see color. Of course, since level of light is a continuum, there must be some in between levels in which both rods and cones are operating. In these levels (twilight, or night with some artificial light), one can still see in color but not as well as at levels in which the cones are fully operating.

There are three subclasses of *cone* cells, called L, M, and S where L is for long wavelength (red), M is for medium wavelength (green) and S is for short wavelength (blue). Each cone type is defined by a "pigment" (a protein) that absorbs certain wavelengths of light better than others.

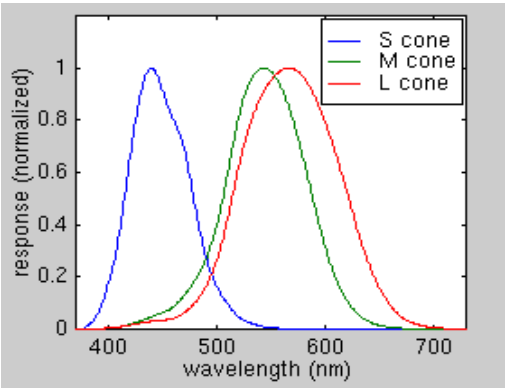
 ©Michael Langer

The curves below shows how the response (or sensitivity) of the rods and cones depends on wavelength. Each of the curves has been normalized so that its maximum is 1. That is, for each photoreceptor type, there is a certain wavelength for which that photoreceptor type responds best. These curves are called *spectral sensitivity functions*.

Let's just consider the cones for now. The L and M cones have quite similar spectral sensitivities and that the range of wavelengths for L and M is almost non-overlapping with the range for S. This has consequences for how these three "channels" are encoded in subsequent processing stages, which I'll discuss next lecture.



rods



cones

Given some spectrum $E(\lambda)$ of the light arriving at an L, M, or S cone, we can write out a measure of the light absorbed by the photoreceptor as:

$$I_{LMS} = \int C_{LMS}(\lambda)\ E(\lambda) d\lambda$$

which is effectively over 400 to 700 nm since the C functions are non-zero only there.

We can loosely think of this quantity as the responses of L,M, or S cells to the spectrum $E(\lambda)$. But we should keep in mind that the response of a photoreceptor is a complicated thing: it involves changes in membrane potential as well as release of neurotransmittors. One can measure the former but not the latter, and even measuring the former is quite difficult to do.[5] The main idea I want you to get across here is that, for a given tiny local neighborhood on the retina where all three cones are present, there will be triplet of values I_{LMS} that we can associate with the response of the cell. It is the *values* that matter, since our goal here is to build computational models (starting next week).

Let's returning to the above equation. Although λ is a continuous variable, in practice one represents such a spectrum by breaking the interval 400 to 700 nanometers into N_λ bins, for example, 30 bins each having a 10 nanometer range. These bins are small enough and the functions are smooth enough that the functions are approximately constant within each bin. Let E_λ be this spectral intensity function, where now λ is discrete.

[5] Such measurements have been made both in live animals as well as in photoreceptor cells that have been isolated and kept alive.

Similarly, we can characterize the response of the three cone types to this beam of light using a $3 \times N_\lambda$ matrix $\mathbf{C}$ whose rows are C_L, C_M, and C_S, respectively. Each row specifies the relative sensitivity of the cone at each wavelength. i.e. each curve has been normalized to have a maximum value of 1. Therefore, we model the responses of the three cones with a discrete approximation of the integral:

$$\begin{bmatrix} I_L \\ I_M \\ I_S \end{bmatrix} = \begin{bmatrix} C_L \\ C_M \\ C_S \end{bmatrix} E(\lambda) \tag{3}$$

A key implication of this model is that information is lost when a cone measure the light. The discretized spectrum $E(\lambda)$ has dimension N_λ whereas a cone only has one response. The fact that each cone cell's response is just one variable (one dimensional), and all information about the spectrum collapses to that one variable is so important that it is given a name: the *principle of univariance*. The same principle applies to rod cells too, of course.

Metamers and color blindness

One implication of the linear model is that if two spectra $E_1(\lambda)$ and $E_2(\lambda)$ produce the same integrating intensity triplets at a point

$$\mathbf{C}\ E_1(\lambda) = \mathbf{C}\ E_2(\lambda)$$

then these two spectra will be visually indistinguishable. In this case, these two spectra are called *metamers.* Metamers occur often, especially in scenes with many surfaces and different reflectances, but we are (by definition) unaware when they occur.

One important example of metamerism, which we do notice, is color blindness. Many people (2 % of males) are missing a gene for one of the three cone pigments. This leads to three types of "color blindness", depending on which type is missing. "Color blind" doesn't mean the person can't see any colors. Rather, it means that they cannot distinguish some spectra that color normal people can distinguish. Such spectra are metamers for the color blind person.

The model of color blindness follows immediately from the above matrix model. With color blindness, one has only two classes of cones and so the matrix is $2 \times N_\lambda$ rather than $3 \times N_\lambda$. One only has two variables by which spectra can be distinguished rather than three. Many professions do not allow color blindess (police officer, baggage handler, electrician, pilot or driver)[6]

Another type of color blindness – and indeed a very common one – is that one of the pigments for the three cones has a different spectral absorption than normal. This typically occurs with either L or M cone. For example, the abnormal (anomolous) cone, say L, has an absorption spectrum that is closer to the M's absorption spectrum than a normal person's L cone is. Such a person still has a three dimension color vision (*trichromacy*) but has trouble distinguishing red from green. Such a person is said to be an *anomolous trichromat*. Notice that if the absorption spectrum of the L cone happens to be very similar to that of the normal M cone, then such a person is essential a *dichromat* i.e. having just two cone types (M and S).

[6] `http://wereadbetter.com/7-jobs-that-you-are-prohibited-from-with-colorblindness/`

In case you wish to read more on this, here is some of the basic terminology. A person who is missing one type of cone is said to have _____*anopia* where the prefix to fill the underline specifies which of the three cones is *missing*: *prot* for L, *deuter* for M, or *trit* for S which are Greek roots for first, second, third. So, for example, a person missing the S cone is said to have tritanopia. A person who has *abnormal* cone absorptivity is said to have a _____*anomaly*. So, for example, someone with an abnormal L is said to have a protanomaly.

One student asked in class what the term *red green color blindness* referred to. It refers to any type of problem with the L or M cones, namely one could be either missing or just anomolous. Problems with the L or M cones are much more common than problems with the S cones. Another student asked if some people are missing two of the three cones. I looked it up and the answer is yes, but it is very rare.

Rod vision is an extreme case of metamerism. In sufficiently dim conditions in which the cones are not operating, one no longer perceives color. One does still perceive shades of gray though. We would say that two surfaces that are placed side by side and that produce the same rod response levels would be metameric. Note that rods are most sensitive to wavelengths in the middle of the spectrum, which we roughly associate with say green. This does not mean that the world at night looks green. Rather, it means that if you have red, blue and green objects that appear roughly equally bright during the day, then the green object will appear brighter at night.[7]

Color displays

One type of spectrum where this theory finds an application is electronic color displays (projectors, computer monitors, TVs, cell phones). Is it easy to characterize the spectra of light coming from each pixel of a display by adding together the three spectra that are determined by the RGB values of that pixel. More precisely, let the spectra of light emitted by each of the RGB color elements of a display be represented by the columns of an $N_\lambda \times 3$ matrix $\mathbf{P}$. (For old TVs, $\mathbf{P}$ stood for "phosphor".) Let $\mathtt{e}$ be a 3×1 vector that specifies a scalar weight for each spectra. So the spectra $E(\lambda)$ that results from a pixel can be written as the following weighted sum:

$$E(\lambda) = \mathbf{P}\ \mathbf{e}$$

For simplicity,[8] let's take $\mathbf{e}$ to be the RGB values in [0,1] at a pixel. In the slides I just wrote RGB instead of $\mathbf{e}$.

What is the set of LMS values that can be produced by such a color display ? If we let the three components of $\mathbf{e}$ be in [0,1], then we can look at how those $\mathbf{e}$ vectors map to triplets of intensities absorbed by the LMS cones:

$$I_{LMS} = \mathbf{C}\ E(\lambda) = \mathbf{C}\ \mathbf{P}\ \mathbf{e}$$

The matrix product $\mathbf{C}\ \mathbf{P}$ is a 3×3 matrix which maps from the unit cube of $\mathbf{e}$ values to LMS space.

[ASIDE: I did not mention this in the lecture but ... from basic linear algebra, we can see that the three columns represent respectively the LMS coordinates of the R,G, and B emitters (on maximum intensity, i.e. value 1). According to this simple model (which is basically correct,

[7] There is a technical sense in which we can compare the brightness of a red versus green or blue object, but we don't have the tools for explaining that yet in the course.

[8] I say "for simplicity" because usually there is also a non-linear transformation called *gamma* from the RGB value to the $\mathbf{e}$ value.

ignoring the issue of monitor gamma), the LMS triplets that can be reached are a distorted cube in LMS space, namely the linear transformation of the points in the unit cube in RGB space.]

Transmission spectra, and anaglyphs

An interesting example in which transmission spectra matter is the case of colored glass or plastic. One vision application is *anaglyph images* which can be used to produce perception of 3D. Anaglyphs are composed of a pair of grey level images that are presented in the different color channels. For example, one image might be presented in the R channel only with R having some value ρ that varies across the image, and the other image might be presented in the G and B channels with value ψ that varies across the image.

The key idea of 3D stereo using anaglyphs to film (or photograph, in the case of a still image) a scene from two neighboring camera positions. Then, when presenting the scene as an image as described briefly above, place color filters in front of each eye that will only let the light from one of the two images through. Typically anaglyph glasses have a red filter over the left eye and a cyan filter over the right eye, so the left eye will see the $(\rho, 0, 0)$ red image and the right eye will see the $(0, \psi, \psi)$ cyan image, where ρ and ψ will vary with position in the image. This gives a 3D effect since the images that reach the left and right eye correspond to the images that were capture by a left and right camera in a 3D scene – namely the binocular disparities are consistent with that 3D scene. We will discuss this again later. In the meantime, see the example in the slides and see the exercises.

Temporal effects on image measurement

Let's briefly discuss the *temporal* properties of the cell response to light. If you flash a pulse of light briefly on a photoreceptor, it doesn't respond instantly but rather it takes several milliseconds to respond and then the response continues. The membrane potential decreases (becomes more negative – see slide) for a short time and then climbs back to its resting state. The magnitude of the response (size of the potential drop) and the duration of response will depend on the length and magnitude of the pulse of light that was used. Not surprisingly, a longer duration and higher magnitude pulse will produce a greater response.

There is another factor that determines the response of a photoreceptor and that is the intensity of the light over the recent past, which affects the current state of the cell. If the cell was continuously exposed to a bright light for a few seconds or even minutes and then it was stimulated with the pulse mentioned above, it will have less of a response to that pulse than if the cell was exposed to darkness in the previous several minutes before the pulse. The main concept here is that, at any time, the cell will have some operating range over which its response depends on the brightness of a short pulse of light. If the brightness of the pulse is too low, there will be no measurable response. If the brightness is too high, then the response will max out. (One refers to the response as being saturated.) Often the response obeys a sigmoid (S) shaped curve. This response curve itself will shift as the background level of light changes.

Camera's also have this sigmoid shaped response curve. For any camera setting, if the image captured is too dark (e.g. because the scene is dark or because the exposure time is too short) then the image will have RGB values of 0. If the scene is too bright, then it will have values of 255 (maximum 8 bit value). There is some operating range in the middle in which the camera will

measure distinct image intensities in each RGB channel. Indeed part of the technical challenge of photography is choosing the camera settings so that you don't have too many 0 or 255 values.

Getting back to photoreceptors and vision, we refer to the shifting of the response as *adaptation*. Adaptation occurs not just in the photoreceptors; it occurs throughout the vision system and indeed throughout all sensory systems. I presented an example in the lecture slides of how we adapt to a white or black square on a grey background. If you look at the dot between the two squares for say 30 seconds and then you look to the right, you will see a blurry black and white square in the visual direction where the white and black square were, respecively. Roughly what is happening here is that the cells that encode for those parts of the image are adjusting their "code" for what is dark versus light. The part of the image that adapted to the black square is now procssing grey, and so that part of the visual field looks brighter, since grey is brighter than black. Similarly, the part of the image that adapted to the white square is now procssing grey, and so that part of the visual field looks darker, since grey is darker than white.

What's happening here is that your visual system does not just provide information about what is out there. It also (simultaneously as part of the its "code") provides information about *changes* in what was out there. While the visual system obviously makes mistakes in judging brightness, the visual system seems to have an overall benefit in adapting because it allows us to move our eyes from dark parts of scenes to bright parts of scenes and adjust our operating range. The adjustments can occur not just at times scales of seconds, but rather the adjustments can occur over minutes. For example, when we walk from the bright sunlight into a cave, the drop in intensity can easily be a factor of one million or more. This extreme adaptation is handled by more than just shifting the operating range of photoreceptions; it is handled by switching from the cone system to the rod system.

Yet another form of adaptation is the pupillary response. If the overall light level in a scene goes up suddenly or if you look at a brighter region of the scene – then your pupil may shrink to reduce the amount of light to come in. This provides a *global* adaptation, which is different from the *local* adaptation for the squares above. Note that the diameter of the pupil can range from say 2 mm up to say 8 mm. Considering that the area grows like diameter squared, the changing pupil size can lead to a roughly factor of 16 range of intensity of light reaching the retina.

Next lecture I will finish up the discussion of photoreceptors and then we will move on to other parts of the retina.

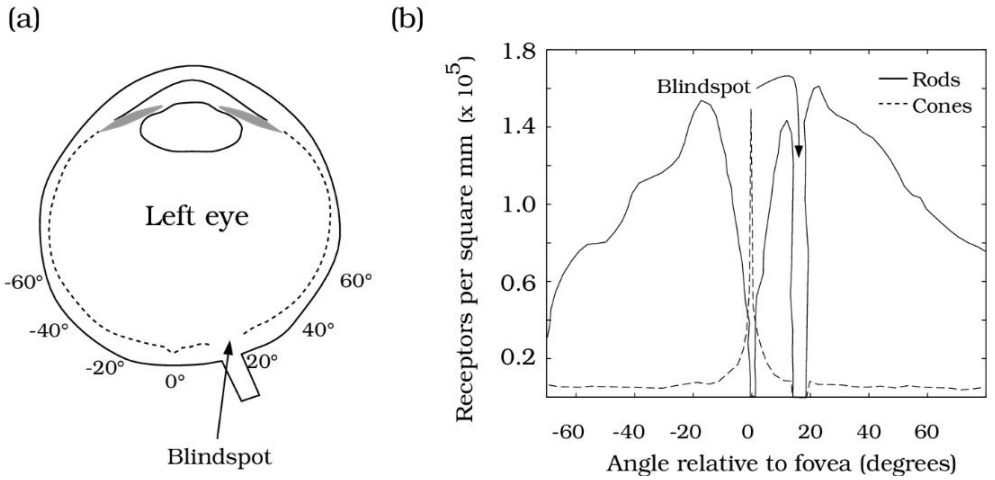
 ©Michael Langer

Retina

We didn't quite finish up the discussion of photoreceptors last lecture, so let's do that now.

Let's consider why we see better in the direction in which "we are looking" than we do in the periphery. There are a few reasons for this. One of the main reasons is that the density of cones is so much greater in the center of the field of view. More pixels per degree of visual angle means more image information. Cone density falls off quickly to about 10 degree away from the center and byond that it remains roughly constant.

In the periphery, the density of rods is much higher than the cones, and indeed the density of rods in the periphery is comparable to the density of cones in the fovea. Does this mean that we can see as well in the periphery at night as we can in the fovea during the day? Obviously that can't be, and the reason is that the rods are much less reliable sensors. They are noisy since they operate under low lighting conditions where the image "signal" is relatively small and so any noise has a more significant effect.



The cell densities are plotted in cells per mm^2. You should be able to say what a mm^2 corresponds to (roughly). See Exercises. The exercises also discuss the blindspot shown in the figure.

Ok, we are done with the photoreceptors for now. Let's consider other cells in the retina. The retina consists of several layers of cells. The first layer contains the photoreceptor cells, and is followed by three layers which perform computations to encode the image. The cells in these initial four layers have continuous responses. See slides.

The cells in the fifth layer are called the retinal *ganglion cells*. They are quite different from the other cells in the retina since they need to transmit their response to the brain. They do so by sending spikes. I discussed spikes in the introductory lecture (0). Let me return to them briefly now and repeat some of the points I made back then.

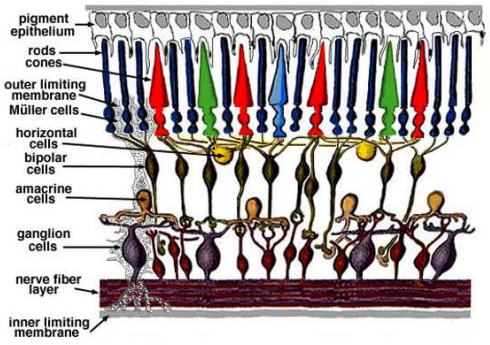


Fig. 2. Simple diagram of the organization of the retina.

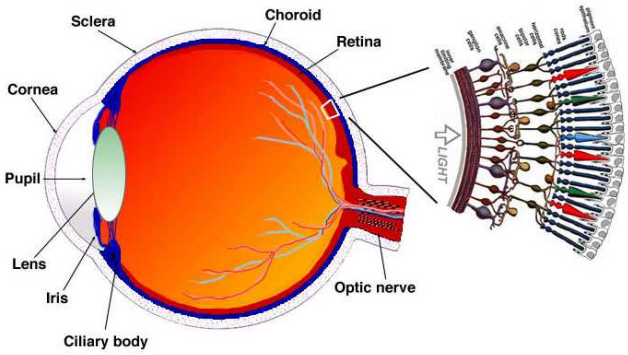


Fig. 1.1. A drawing of a section through the human eye with a schematic enlargement of the retina.

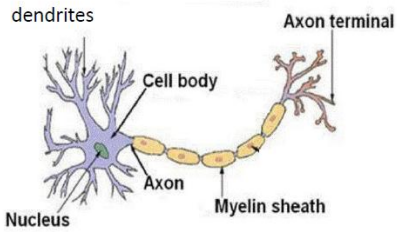
Responses of a neuron: continuous vs. discrete

Cells in the retina respond to images and they encode (in some sense) what is in the image. But what does it means to say that a neuron in the retina has some response to an image? Let's distinguish to aspects of "response". The first is what an experimenter can measure from a single neuron, namely an electical potential difference across the neuron's cell membrane (more on that below). The second is what a neuron communicates to a neighboring neuron, namely it releases hormones (neurotransmitters) that are picked up by the neighbor; these neurotransmitters in turn affect the responses of neighboring neurons.

Here is a bit more detail. First, the figure below shows the basic structure of nerve cell (or neuron) such as a retinal ganglion cell. It has a cell body, and the cell body has branches coming out of it. These branches are called *dendrites*. When neurotransmitters are released from neighboring cells in the retina, these neurotransmitters may bind to the dendrites, which causes a change in the electrical potential across the neuron's cell membrane. As in the case of photoreceptors, the change is that membrane channels will open or close, allowing ions such as potassium and sodium to travel in or out of the cell. The net effect is that the concentration of ions inside versus outside the cell will vary over time, and thus there may be a difference in electrical potential across the cell membrane. This is what an experimenter typically measures, when studying the state and response or a single neuron.

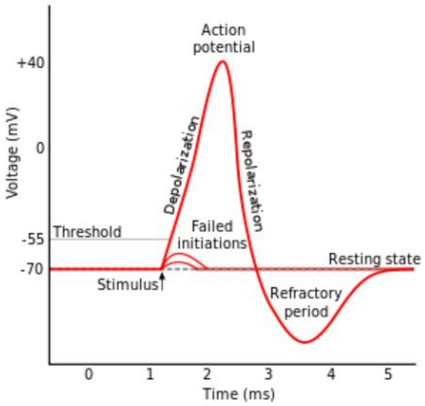
The resting (average) potential of a cell is typically about -70 mV (millivolts), namely the inside of the cell has more negative charge than the outside. If the potential difference is above -70 mV, then we say that the cell is *depolarized* (i.e. closer to 0), and if the potential differences is below -70 mV then we say that the cell is *hyperpolarized* (i.e. further from 0).

The communication between cells occurs at a location called a *synapse*. The cell releasing the neurotransmitter is the pre-synaptic cell and the cell receiving (binding) the neurotransmitter is the post-synaptic cell. These neurotransmitters can be either depolarizing or hyperpolarizing. As we will see next, a depolarizing neurotransmitter is *excitatory* and a hyperpolarizing neurotransmitter is *inhibitory*.



Spikes (Action Potentials)

How do cells communicate over long distances? For example, how does a retinal ganglion cell communicate its response to the rest of the brain? The basic mechanism for long distance signally is called a *spike* or *action potential*, which is sudden and large depolarization of the cell membrane. See figure below.



An action potential is triggered when the cell membrane reaches a certain depolarization threshold, which causes it to depolarize further and even become positive. The action potential is propagated over a distance as a single wave (spike) along a special part of the cell called the cell *axon*. Think of an axon as a long wire. (In fact it is tube wrapped in a fatty insulator.) A cell typically can "spike" up a rate of up to 200 times per second.

There is much to say about spikes but let's just consider a few important facts for now. First, for a given cell, every spike has the same shape. (See sketch above.). The information carried by spikes is purely in the timing of the spikes, not the shape. There has been much effort in the past few decades to understand exactly how much the timing matters. On the one hand, the initiation of a spike depends on a somewhat noisy signal (namely, binding of neurotransmitters from neighboring cells) and so it is difficult to imagine how the exact timing could be reproducable and hence reliable. On the other hand, some computations do require precise timing, as we'll see later in teh course when we study the auditory system.

Receptive Field of Retina Ganglion Cell

Cells in the visual system (at least those in the early processing stages) typically respond to images in a restricted region. For a given cell, we refer to the set of visual directions that the cell is sensitive to as its *receptive field*. Photoreceptors obviously have a very small receptive field, since they pretty much only respond when light strikes them directly. For other cells in the retina, the response of one cell will affect another neighboring cell. The result is that cells can respond over a wider range of visual directions, by being indirectly affect by responses from other cells.

Just like the density of rods and cones varies over the retina, there is variation in the retinal ganglion cells across the retina. In particular, the sizes of retinal ganglion cells is smallest in the fovea and increases in the periphery. (See plot in slides.) This increase in receptive field size roughly follows the decrease in density in the cones.

What information do the spikes from each retinal ganglion cell encode about the retinal image? The ganglion cells do not simply encode a pixel by pixel copy of the LMS photoreceptor image. Rather, they pre-process the image to make some aspects of the image more explicit. Indeed all layers of the retina contribute to this pre-processing. Rather than looking at the detailed circuits in the various layers of the network, let's look at some simple models of the what image transformations are being computed.

The simplest model is that the visual system encodes sums and differences of LMS (cone) response values in local neighborhoods. That is, after the LMS cones measure the light arriving at the retina at each location (x, y), subsequent layers of cells in the retina compute weighted sums and differences of the LMS responses. We'll look at a few types of these sums and differences: spectral, spatial, temporal, and combinations of all these. Today we'll just discuss spectral and spatial.

Spectral sums and differences, and color opponency

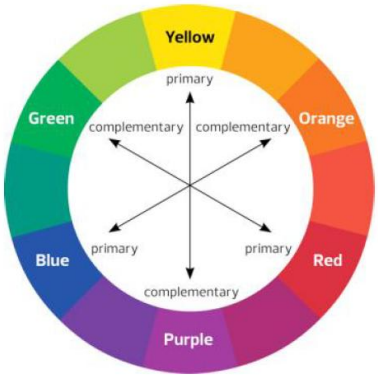
From many experiments over the years, neuroscientists have learned that the retina transforms the LMS measurement as follows: "L + M" measures the overall physical brightness in the medium and long wavelengths, "L-M" compares the long and medium wavelength response, and "L + M - S" compares the medium/long overall intensity to the short wavelength intensity. These arithmetic expressions should not be taken literally right now, since I haven't defined what exactly L, M, S mean here in terms of numerical values. (e.g. What are the units?) Rather, for now, just think of them symbolically: e.g. L+M-S means that there are cells who response increases when the image in the receptive field of the cell has stronger L or M components and the response decreases when the cell's receptive field has a stronger S component.

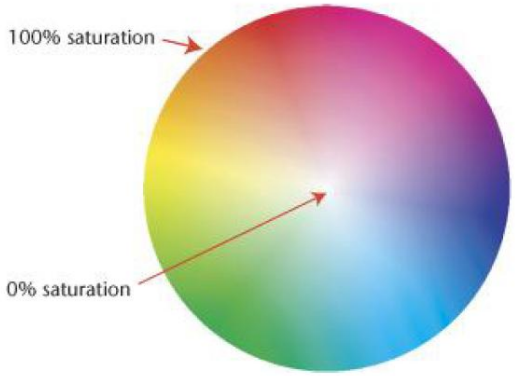
Measuring *differences* in cell responses is called *opponency*. L - M is called red-green opponency. L + M - S is called yellow-blue opponency. The reason $L + M$ is called "yellow" is that if you mix together two lights that appear red and green, then you get a light that appears yellow. e.g. an image pixel with RGB value (1,1,0) appears yellow.

Color opponency is a very old idea and can be expressed in many ways. For example, in school you may have learned about primary colors and secondary colors and how to use them. (See ASIDEs below.) In vision science, the idea of opponency goes back to Hering in the late 1800's. One of the key observations is that some colors seem to be in-between other colors, e.g. we perceive orange as reddish yellow, as if both red and yellow are both *in* orange. Similarly, we perceive cyan as blueish green, and we perceive purple as reddish blue. However, we cannot perceive a color to

be blueish yellow, or reddish green. These pairs of colors oppose each other in some fundamental sense. These observations are believed to be the direct perceptual consequency of an underlying opponency circuitry, namely computing the LMS differences mentioned above.

If you took art class in school, then you are familiar with the idea of color opponency already. You learned about primary and secondary colors and how they related to color mixing. You also learned about "complementary" colors" and how colors can be arranged on a wheel (red, orange, yellow, green, blue, purpose) and how there are special relationships between colors that are opposite each other on the wheel. (See below left.) I am not going to attempt to explain color art theory in this course; I just want to mention that there are connections to color opponency.





Hue, saturation, value (HSV)

The color signal in an image is a 3D vector (LMS) and there are many ways to encode these 3D vectors. One of the common ways to distinguish colors from each other is based on the *relative* amounts of the spectrum at different wavelengths versus the *total* amount of light in the spectrum. In LMS theory, the former concerns the two difference channels (L-M, L+M-S) and the latter concerns the L+M channel. If one thinks of a color circle, then the points on the edge of the circle define colors that are as pure as can be, and points in the interior of the circle (see right above) correspond to a mix of pure colors with a neutral color (white or grey). By using a polar coordinate system for points in the circle and its interior, one can sweep out a range of colors. The angle or direction from the center of the circle defines the (maximally) pure color – often called the *hue*. The distance from the center is the purity – often called the *saturation*.

The polar coordinate system accounts for two of the three dimensions of LMS color space. The third color dimension is often called the `value`, or lightness, or luminosity. (These terms all have specific technical definitions in color science, but the details don't concern us.) The specific case of saturation equal to 0 is the center of the color circle. In this case, the values can range from black to grey to white. Think of this third dimension as coming out of the page in the figure above right.

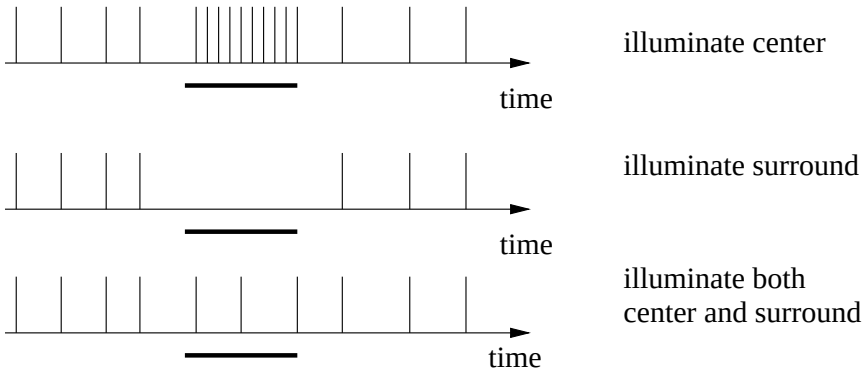
If you have used color pickers in MS paint or Powerpoint to select colors, then you will be familiar with these terms. I encourage you to experiment for a few minutes and see how RGB values gives rise to different HSV (or HSL) codes. At the very least, see the slides for an example.

Spatial sums and differences: Lateral inhibition

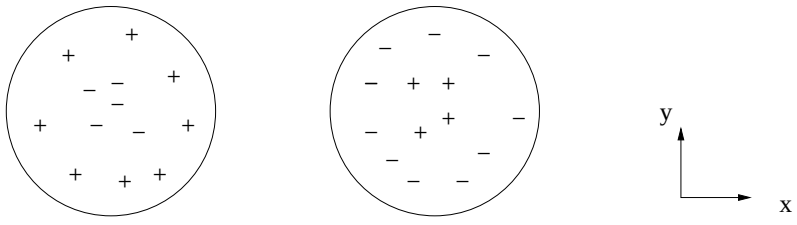
As mentioned earlier today, because interactions in the retina are spatially localized, each ganglion cell can respond to only a spatially restricted region in the retina: the receptive field. Interestingly, a cell does not have a uniform response to light over its receptive field, but rather it computes sums and differences of the light intensity over different parts of the receptive field. They also compute temporal sums and differences, but I won't mention that today.

In 1950s, a researcher named Steve Kuffler for the first time measured spike trains from single ganglion cells of the cat retina. He recorded from single cells over time, while shining a tiny spot of light on the retina. He carried out these experiments in a very dark room, so that the only light shining on the retina was the tiny spot. He found that for each retinal ganglion cell there was small region of the retina that affected the spike rate of that cell, i.e. the receptive field.

Kuffler found many ganglion cells for which the firing rate increased when the tiny spot of light shone on a particular region. This is called *ON region* for that cell. He also found that surrounding this ON region was an annulus (ring) shaped region in which the tiny spot of light *decreased* the firing rate of the cell. This surrounding region is now called the *OFF region*. Because these cells were excited by light in the center and inhibited by light in the surround, these cells are called *ON center/OFF surround*.



Kuffler also found retinal ganglion cells that had the opposite property, namely there was a central round region in which the the cell's response decreased when the tiny spot of light was shone there, and a surrounding annulus region in which the response increased when light was shone there. These cells are called *OFF center/ON surround*.



One can model the cell's response behavior by assigning weights to the different points in the receptive field. The ganglion cell's response is the sum of the weighted intensities over the receptive

fields. For now, we just to think of the L+M channel. In Assignment 1, you will think about difference channels too.

Note that because the intensities in the surround have the opposite effect as the intensities in the center, you can think of the image in the surround as inhibiting the response to the image in the center. This local spatial inhibition (or opponency) is often called *latera inhibition*.

DOG model

One model for achieving the center-surround effect is to suppose that there is one mechanism for excitement over a neighborhood and that the effect falls off with distance from the center of the receptive field, and that there is a different mechanism for inhibition that also falls off with distance. If the excitation were to come from a small neighborhood and be strong in that neighborhood and if the inhibition were to come from a larger neighborhood and be weaker over that neighbhorhood, then this would naturally lead to an ON-center/OFF-surround receptive field.

Rodieck and Stone (1965) proposed a specific model which was based on the 2D Gaussian function:

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \tag{4}$$

A 2D Gaussian is just the product of two 1D Gaussians,

$$G(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2}$$

$$G(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-y^2/2\sigma^2} \ .$$

The 2-D Gaussian is *radially symmetric* in the sense that it only depends on the squared radius x^2+y^2. Note that this Gaussian is centered at $(0,0)$ but more generally it could be centered at any (x_0, y_0) by shifting. Also, note that we are ignoring the time dimension.

The *difference of Gaussian* function is then defined:

$$DOG(x,y,\sigma_1,\sigma_2) \ = \ G_1(x,y,\sigma_1,\sigma_2) \ - \ G_2(x,y,\sigma_1,\sigma_2)$$

and again it is centered at $(0,0)$. Here 1 and 2 are the center and surround, i.e. $\sigma_1 < \sigma_2$. This center would be ON-center and OFF-surround. To obtain OFF-center ON-surround, one would use $\sigma_1 > \sigma_2$.

Finally, the response of a retinal ganglion cell whose receptive field is centered at (x_0, y_0) depends on the inner product of the DOG with the image

$$L(x_0,y_0) \equiv \sum_{x,y} \ DOG(x-x_0, y-y_0) \ I(x,y)$$

where L stands for "linear". However, the response of the cell (e.g. firing rate of a retinal ganglion cell) isn't exactly modelled by L. For example, cells have a maximum firing rate, so if we were to increase the image intensity, then eventually the response would saturate. Also, cell's cannot have negative responses. So if the image were positive only in the "negative part" of the DOG function, then the model would give a negative number for L, which wouldn't make sense as a response. To convert the L into a meaningful response, we would need to set the response to 0 when the L values

 ©Michael Langer

are negative. One can model these non-linear mappings from L to a response in several ways, for example, using a sigmoidal shape curve, or by *half-wave rectifying*, namely setting all negative L values to 0. See slides.

A related point is that we need both ON-center OFF-surround cells and an OFF-center ON-surround cells. Depending on the image, an ON-center OFF-surround cells can have an L value that is either positive or negative. In the case it is negative, the cell would have no response and so the information about the image would be lost. Having an OFF-center ON-surround cell at that same location would have a positive L value, namely the negative of the negative value of L of the first cell. So as long as both types of cells are around, no information will be lost. (Of course, we still have the issue of saturating to bright images. The only way to deal with that is adaptation, as discussed last lecture.)

Cross-correlation

To understand retinal processing of images, we want to know not just the response of a single cell to the images, but also also the responses of a family of cells that all have the same receptive field shape. For this, one defines the cross correlation of two functions, in this case DOG and I by:

$$DOG \otimes I(x_0, y_0) \equiv \sum_{x,y} DOG(x, y)\ I(x_0 + x, y_0 + y)\ = \sum_{u,v} DOG(u - x_0, v - y_0)\ I(u, v)$$

where I used a change of variables $x_0 + x = u$ and $y_0 + y = v$.

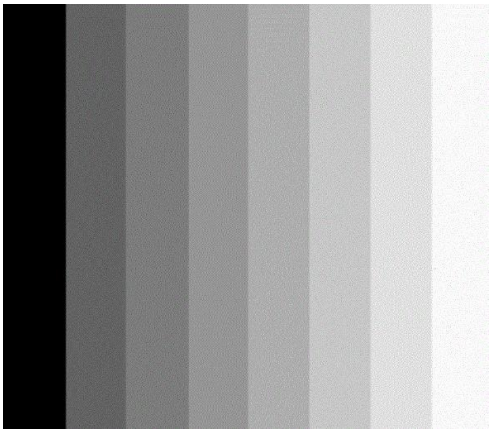
Think of the DOG as a template, and imagine sliding that template across the image. See slides. The formulas above says the template is at (x_0, y_0). But you should think of (x_0, y_0) one of many positions. So we are thinking of cells at many different positions and we are thinking of the responses of a *population* of cells that all have the same receptive field weighting function, namely a DOG of some particular σ_1 and σ_2 that define a center and surround size.

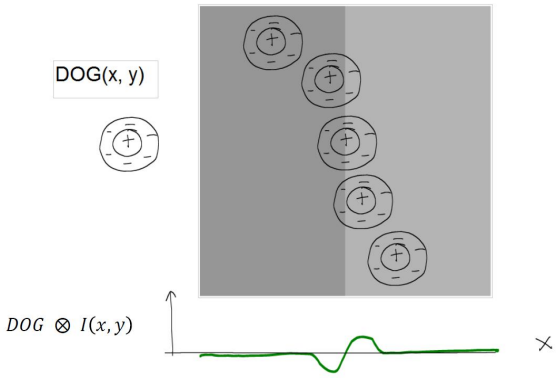
 ©Michael Langer

Today we'll examine how orientation information such as edges and lines is encoded in early visual processing. There is much to say about this topic before one gets to the stage in visual processing where individual cells are sensitive to oriented structures though. I'll keep this preliminary discussion short and just discuss a specific phenomenon called Mach Bands[9].

Mach Bands

If you look at the image on the left which consists of a set of stripes, each of constant shade of gray, you will notice that the boundaries between the stripes appear to have slight rise (when the stripe goes from light to dark) or fall (when it goes from dark to light). This *edge enhancement* effect is believed to be an artifact of how our eye and brain codes the image. It causes us to fail to perceive the intensities as they really are.





Many have argued that Mach bands are the result of the center-surround coding mechanism, in particular, the DOG "filtering" that happens in the retina. This idea is illustrated in the sketch above right. As we move the DOG template across the edge, it begins in a uniform region and gives 0 response since the ON and OFF regions cancel. Then it encounters a rise in intensity in an OFF region, which leads to an overall negative linear response. When the DOG straddles the edge, the left and right halves of the DOG each have uniform intensities and because of symmetry the ON and OFF regions in each half are balanced just as the cell's overall ON and OFF regions are balanced, so again the cell give no response. As the DOG template continues beyond the edge, the intensity rises because the tailing edge of the OFF region falls on the lower intensity region – less OFF contribution leads to an increase in response.

Mach bands are not just a curiosity. They have practical applications, for example, when people examine images and need to make subtle distinctions between grey levels. The best example of this is dentists or radiologists who examine radiographs. Such professionals are well acquainted with the effects of Mach bands. They cannot change their visual systems, but they can learn when and when not to believe what their eyes are telling them.

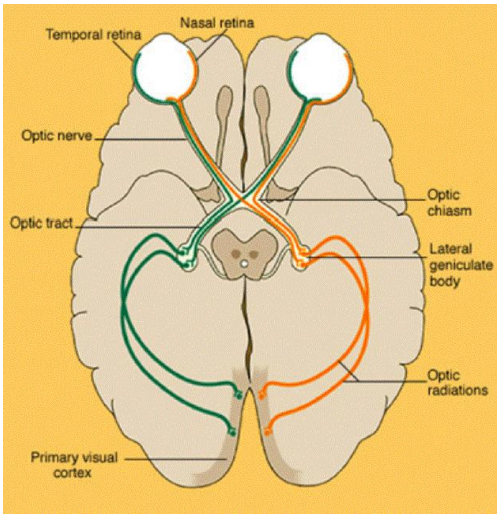
[9]named after Ernst Mach who was a 19th century scientist

 ©Michael Langer

Early visual pathway: retina to cortex

Let's move further into the brain. The axons of the retinal ganglion cells of each eye are bundled together into the optic nerve which sends the signals to the lateral geniculate nucleus (LGN) on each side of the brain. The two LGNs which are in the thalamus are located near the center of the brain. Note that the optic nerve from each eye needs to split into two in order to send signals to both halves of the brain. See figure below.

Cells in the LGN relay the signals to the surface of the back end of the brain. The surface of the brain in general is called the *cortex*, and the surface of the brain at the back of the head is called the *primary visual cortex* (V1) because this is the first area of the cortex to receive visual inputs.



[ASIDE: I say the LGN cells "relay" the retinal signals to the cortex, but there is more going on than that. The LGN receives axons from the retinal ganglion cells (about 10^6 of them), but it receives far more axonal inputs (about 10^7 of them) from the visual cortex – that is, there is a feedback loop between the visual cortex and the LGN.]

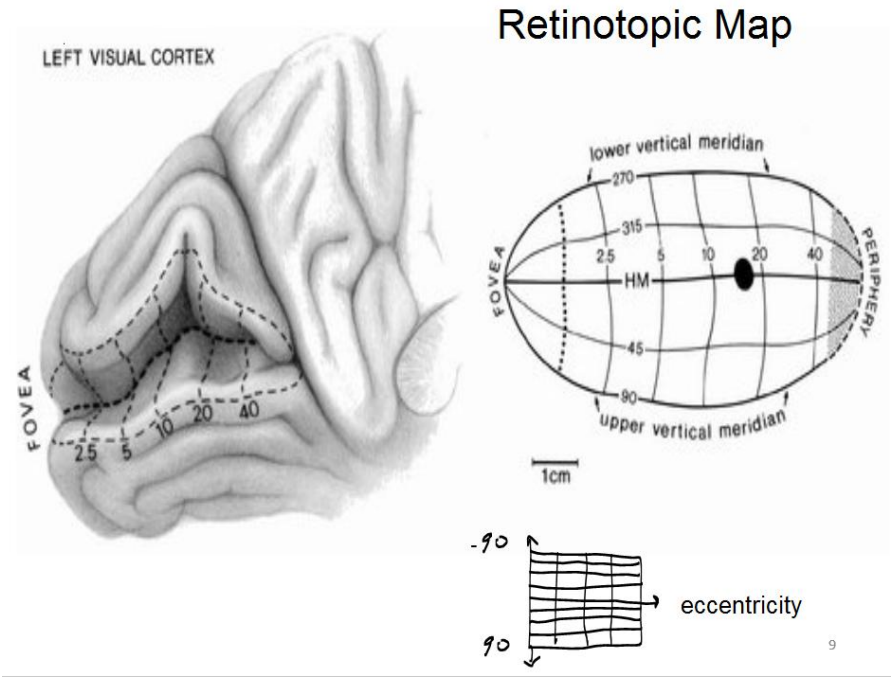
The figure above illustrates how the two halves of the visual field are coded by the two halves of the brain. The left half of each retina codes the right visual field and the axons from these retinal cells terminate in the LGN on the left side of the brain. Relay cells in the LGN then send their axons to the left half of the primary visual cortex (V1). Thus, the left half of V1 receives the image code from both eyes for the right visual field only. Similarly, the right half of V1 receives the image code of left visual field. The cells that encode the "seam" along a vertical meridian between the left and right halves of the visual field can be found in both halves of the brain. So there is an overlap in the representation of the vertical meridian.

Retinotopic maps

When the axons from the retina are bundled into the optic nerve, their spatial arrangement is preserved (to some extent). These cells terminate in the LGN. When you measure the receptive fields from neighboring cells in the LGN, you typically find that they encode the intensities of nearby visual directions, or equivalently, nearby retinal positions. In this sense, the LGN is said have a

retinotopic map: nearby points on the retina map to nearby points in the LGN. Similarly, cells in the LGN project to V1 and if you measure the receptive fields of nearby cells in V1, you generally find that they encode intensities of nearby visual directions (i.e. nearby positions on the retina). So V1 also has a retinotopic map.

Here I give just a few details about the retinotopy in LGN and V1. There are six layers in each LGN, and each relays information from just one eye. There are also differences in the receptive field properties of different LGN layers. In some layers, the cells have relatively large receptive fields but are not sensitive to color differences, and these cells respond to the time variations in the stimuli. These cells are involved in motion processing which I'll get to in a few lectures. In other LGN layers, the cells have smaller receptive fields which encode color and intensity differences. These cells do not seem to be involved in motion processing. The details of the different LGN layers are not crucial for our understanding. My main point here is that within each layer of the LGN, the cells are arranged in a retinotopic map. They then relay signals to V1.



Because the receptive fields[10] of retinal ganglion cells in the fovea are so much smaller than in the periphery, it is possible to pack many more retinal ganglion cells per mm^2 in the center of the retina. The signals get relayed from LGN to V1, and so the inputs to V1 are dominated by the cells near the center of the visual field. This requires a deformation of the retinotopic map.

One simple way to think about this deformation of the retinotopic map in V1 is to use polar coordinates (r, θ) for visual direction instead of (x, y): one coordinate r is eccentricity and the other coordinate θ is an angle away from say the x axis. This polar coordinate system (r, θ) captures

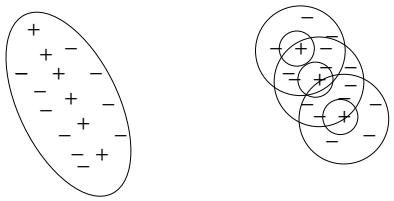
[10]The receptive field size doesn't just depend on the cell body. It also depends on the width span of the cell dendrites (branches) that the cell uses to "read" signals from its neighbors.

the distortion of the retinotopic map, namely that the number of cells that represent a given visual angle increases toward the center of the image (toward $r = 0$.) The figure above illustrates the distortion. V1 in the left cortex is flattened out into an elongated ellipsoid (close to a rectangle). The directions θ from -90 to 90 degrees (or 270 deg to 90 deg) are represented in the map. These cover half of the visual field.

The slides show another example which is based on fMRI images. The point there is that the central part of the visual field is coded using a relatively large part of V1. We still have a retinotopic map, but it is distorted.

Orientation selectivity in primary visual cortex

What are the receptive field properties of cells in the primary visual cortex? The first experiments to successfully address this question were carried out in the late 1950's by David Hubel and Torsten Wiesel. (For this and subsequent work, these two researchers were awarded the Nobel Prize.) Hubel and Wiesel examined the responses of single cells in primary visual cortex of anaesthetized cats. They found that each cell responded to an small area of the visual field but, unlike in the retina and LGN, the receptive fields in V1 were not radially symmetric. Instead the cells were tuned to a particular orientation, such as in the sketch below. The response of this cell can be thought of as a weighted average of the image intensity over the ellipsoidal region shown. The weights are positive along a center stripe parallel to the elongation and negative along the two flanking stripes. Such cells might be thought of as *line detectors*. Cells are also found of the opposite sign, namely negative along the center stripe and positive along the flanking regions.



Hubel and Wiesel discovered the orientation properties quite accidently. Hubel describes the discovery here: `https://www.youtube.com/watch?v=IOHayh06LJ4`
For a longer video showing the mapping of the receptive field, see
`https://www.youtube.com/watch?v=Cw5PKV9Rj3o`.

Simple cells

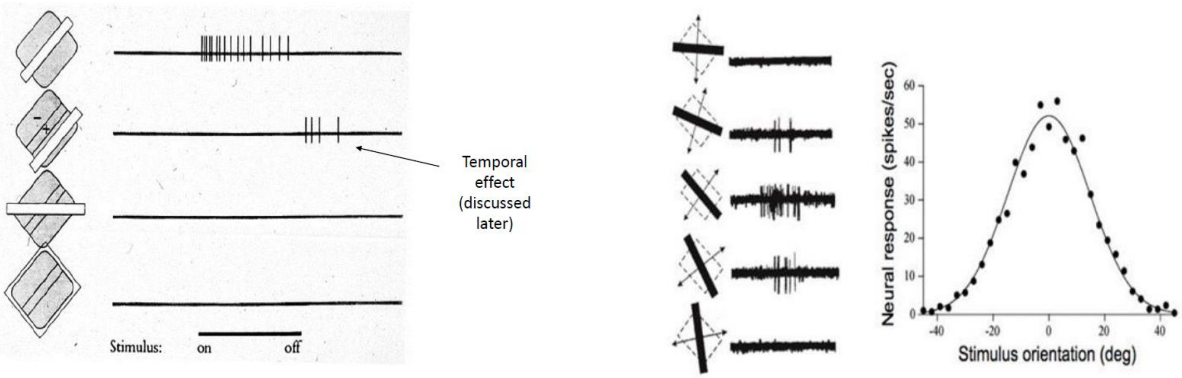
Hubel and Wiesel discovered a number of different types of cells in the primary visual cortex. The type I described above are called *simple cells*. These cells have well defined ON and OFF regions that are elongated such that an oriented bright line can either excite the cell (in the ON region) or inhibit the cell (in the OFF region). I will often call the ON and OFF regions "excitatory" and "inhibitory" respectively. See figure below left. The first shows a white line on the elongated excitory region of the cell. When this white line stimulus is turned on, the cell spiking rate goes up and when the line turns off the cell stops spiking. The second example is more subtle. The line is placed over an inhibitory region. There is no response shown until the white line stimulus turns off,

 ©Michael Langer

as if a removal of inhibition acts as an excitation. The models that we will discuss later today do *not* handle these temporal effects. Next week when we discuss motion processing, we will consider temporal effects.

The third example on the left shows a white line of the wrong orientation and the fourth example shows a very thick white bar, the same width as the receptive field. In both of these cases, there is no response from the cell.

The figure on the right is called an *orientation tuning curve*. It shows how *one cell*'s response varies as the orientation of a line varies. (Note that this has a different meaning than saying that a fixed line stimulus produces responses to *different cells* that have the same receptive field position and size and are tuned to different orientations.)



Hubel and Wiesel proposed that simple cells are formed by summing the inputs from a set of center-surround LGN cells whose receptive field centers fall along a line (see slides). It has also been found that simple cells have a large variety of profiles. Some are ON center OFF surround (with orientation preference, as always); others are OFF in the central elongated region and ON in the flanking regions. Still others have an edge like receptive field structure so they are ON on the left side and OFF on the right side, or vice-versa. Finally, simple cells are also sensitive to color. For example, there are double opponent simple cells that might be R+G- on one half of their oriented edge profile and R-G+ on the other half.

As with the DOG functions from last lecture, simple cell receptive field profiles define either positive or negative linear responses, depending on whether the white line stimulus is on the ON or OFF region, respectively. But neurons cannot have negative responses and so a non-linearity must be used to model the negative response e.g. half wave rectification as we discussed last lecture. As long as one has both ON center cells and OFF center cells, one will not lose information because of half wave rectifiation since one of the two will carry any non-zero response.

Gabor model

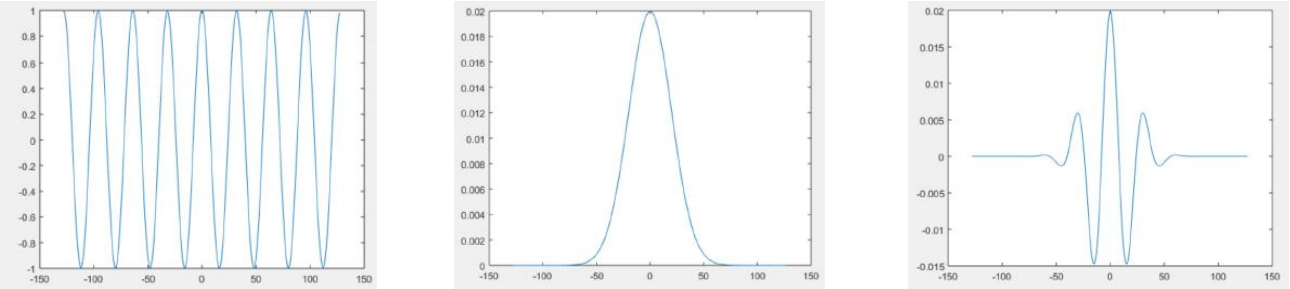
The standard mathematical model of simple cell receptive fields is the Gabor function. Let's define this function first in the 1D case, and then in 2D. Consider a cosine function which is sampled on

a sequence of N uniformly space points

$$\cos(\frac{2\pi}{N}(k_0 x))$$

where k_0 is the spatial frequency which has units of number of cycles per N samples. Typically it is an integer between 0 and $N-1$. Notice that as x goes from 0 to N, the cosine argument goes from 0 to $2\pi k_0$ radians which is indeed k_0 cycles or "times around the circle".

A 1D cosine Gabor is defined by multiplying the cosine function by a Gaussian function of some standard deviation σ. There is no fixed relationship between k_0 and σ. One is free to vary them at will. Increasing σ for a fixed k_0 will increase the number of side lobes. One can define a sine Gabor similarly.



To model the shapes of (2D) simple cell receptive fields, one uses a 2D cosine or sine function and a 2D Gaussian. Consider a 2D cosine function of size $N \times N$,

$$\cos(\frac{2\pi}{N}(k_0 x + k_1 y))$$

where k_0 and k_1 are fixed integers between 0 and $N-1$. This family of 2D cosine functions can define a range of frequencies and orientations. To understand how, note the expression $\frac{2\pi}{N}(k_1 x + k_2 y)$ has a constant value c along a line,

$$\frac{2\pi}{N}(k_0 x + k_1 y) = c.$$

For example, if $c = 0$, the line passes through $(x, y) = (0, 0)$. For different c, one gets different lines and the cosine takes different values. The cosine variation occurs in a direction perpendicular to these lines, namely, in direction (k_0, k_1). One can define a 2D sine function similarly.

Another way to understand 2D sinusoid functions is to note that if you fix x to have a particular value so that you are looking along only a vertical line (column) in the (x, y) domain, then the argument $\frac{2\pi}{N}(k_0 x + k_1 y)$ has k_1 cycles as y goes from 0 to N. Similarly, if you fix x then you are looking along a horizontal line (row) and the argument has k_0 cycles as x goes from 0 to N.

To define a 2D Gabor function, we multiply a 2D cosine function by a 2D Gaussian:

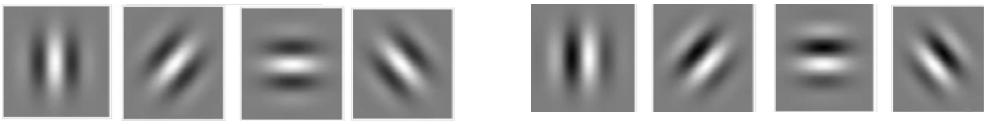
$$cosGabor(x, y, k_0, k_1, \sigma) \equiv G(x, y) \cos(\frac{2\pi}{N}(k_0 x + k_1 y).$$

We define a *sine Gabor* similarly:

$$sinGabor(x, y, k_0, k_1, \sigma) \equiv G(x, y) \sin(\frac{2\pi}{N}(k_0 x + k_1 y).$$

 ©Michael Langer

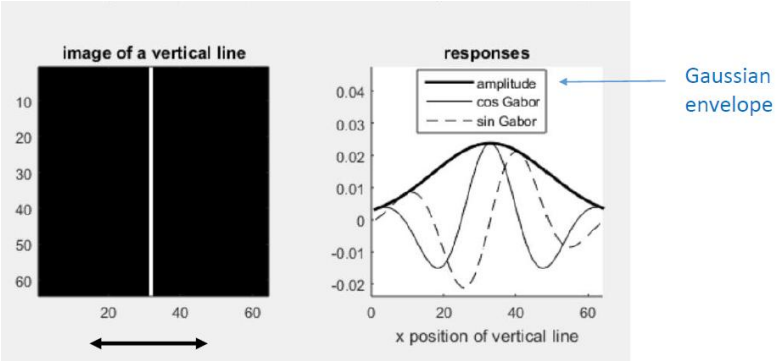
Four examples of cosine (left) and sine (right) Gabors are shown below.



Let's examine how a sine or cosine Gabor cell responds to the position of a line across the receptive field. This is similar to the orientation tuning curve shown above but now we vary the position rather than orientation of the line stimulus. The figure below shows the linear response of a cell as a function of the x position x_{line} of the line, namely the inner product of the Gabor template with the image:

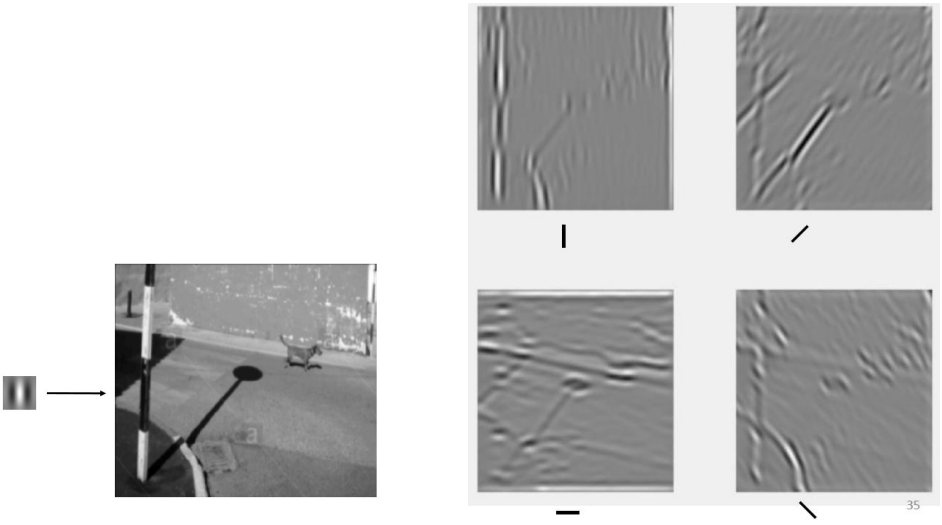
$$< cosGabor(x, y, ...), I(x, y; x_{line}) > \quad = \sum_{(x',y')} cosGabor(x', y', ...) I(x', y'; x_{line})$$

where $I(x', y'; x_{line})$ has value 0 everywhere except on $x = x_{line}$, and the $<\quad>$ notation here is for inner product. Note the response follows the shapes of a 1D sine and cosine Gabor in x. (See Exercises.) The figure below also indicates a "Gaussian envelope". I will discuss this next lecture.



We next examine the response of a family of sine or cosine Gabors to a single image. Here I show just the response to a family of cosine Gabors. (See the slides for responses to a family of sine Gabors.) By "response", I mean the cross correlation of the cosine Gabor and the image. Here are the results for four different cosine Gabors. The filtered image on the upper left shows the results for vertical Gabor.

Notice how the vertical Gabor gives a good response along the pole on the left side of the image, but the details of where the response is a large and positive number (white) versus large and negative (black) vary along the pole. The right diagonal Gabor (top right) picks out the diagonal shadows in the image. Do you understand why the diagonal shadow of the pole is black and flanked by two bright white diagonal regions? (If not, then try to think it through and ask me if you don't get it.) Examine the other two filtered images and identify which parts give a large response.


35

Last lecture we discussed simple cells in V1. We considered sine and cosine Gabor models for such cells. The linear response of these cells to an image was defined by the inner product of a Gabor function with that image.

One technical point: As we saw with retinal ganglion cells, simple cell cannot have a negative response (real cells cannot have a negative number of spikes per second), and so we need two versions of each cell where the weights of one version are just the negative of the other, and so the two versions of each cell have the same linear response magnitude but opposite sign. In the model, the two responses are half wave rectified so one is positive and the other becomes 0. This allows the model effectively to represent both positive and negative responses of the linear cell.

Complex cells

Hubel and Wiesel found a second class of cells in V1 that are also sensitive to oriented intensity patterns (lines and edges), but these cells were quite different from the simple cells. Whereas a simple cell has a well defined excitatory and inhibitory region, this second class of cell does not. These cells are not sensitive to the precise position of the oriented pattern (edge, line) within the cell's receptive field. (Many of these cells are sensitive to motion of an oriented pattern as well. I will discuss this in an upcoming lecture.) Hubel and Wiesel called this second class of cells *complex cells*.

There are many ways one can model a complex cell. One way is to take a set of simple cells that have the same orientation and are distributed over a range of shifted positions. The responses of each simple cell are half wave rectified, and the complex cell could be defined by taking the sum of the these half wave rectified values. The complex cell's receptive field would be the union of the receptive fields of the simple cells that it reads from. It would respond to lines or edges at various positions within that receptive field, but you couldn't say that one position was excitatory or inhibitory. I I sketched out such an example cell in the slides (model 1).

The second and third models that I mentioned in class is defined by a sine Gabor and cosine Gabor pair that have the same frequency (k_0, k_1) and envelope size σ and are both centered at (x_0, y_0). That is, the receptive fields now coincide. The linear response of the sine and cosine Gabors are defined by the inner products of each with the image, and form a pair:

$$(< cosGabor(x - x_0, y - y_0), I(x, y) >, \;\; < sinGabor(x - x_0, y - y_0), I(x, y) >)$$

Think of the pair as a vector in a 2D space. Model 2 and 3 differ in what they do with these two responses.

Model 2 defines the complex cell response as follows:

$$\mid < cosGabor(x - x_0, y - y_0), I(x, y) > \mid + \mid < sinGabor(x - x_0, y - y_0), I(x, y) > \mid$$

that is, the sum of the absolute values. This computation was illustrated in the slides in a slightly different way, namely by taking the sum of two pairs of sine and cosine Gabors, which are each *half wave* rectified. Mathematically, we have the following (for the cosine Gabor). Letting $[\;]_+$ be the half-wave rectification operator, we write:

$$\begin{aligned} & \mid < cosGabor(x - x_0, y - y_0), I(x, y) > \mid \\ = \;\; & [\; < cosGabor(x - x_0, y - y_0), I(x, y) > \;]_+ \;\; + \;\; [\; < -cosGabor(x - x_0, y - y_0), I(x, y) > \;]_+ \end{aligned}$$

and similarly for the sine Gabor.

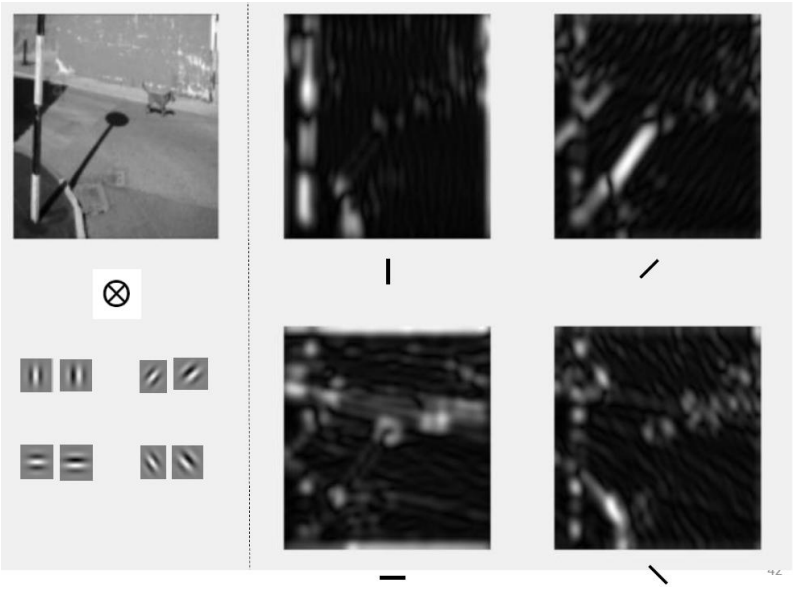
Model 3 is similar except that we take the squared values of the linear cosine and sine Gabor responses, rather than the absolute values. Taking the square value might seem to buy us nothing at first glance, but in fact it does make the math a bit cleaner which hopefully you'll appreciate soon. The basic idea is to treat the linear responses of the cosine and sine Gabor as a 2D vector (we have a pair of values), and to consider the Euclidean length of this vector:

$$\sqrt{< cosGabor(x - x_0, y - y_0), I(x, y) >^2 + < sinGabor(x - x_0, y - y_0), I(x, y) >^2}.$$

This is the third model of the complex cell's response, and it is the most commonly used model.

Example

Consider again the example image from the previous lecture. Now we take the responses of both the sine and cosine Gabors of each orientation, and we compute the complex cell responses at each pixel. We compute eight maps – four orientations of a sine Gabor and four orientations of a cosine Gabor – and we cross-correlate each of the Gabors with the image. For each of the four orientations and for each image position, we define a complex cell response, which is what the images show. The four images below right now represent non-negative values only, so zero response is black. Roughly the same regions as last lecture give a large response. For example, the vertically oriented complex cells gives good responses on the pole, and the right diagonal oriented cells gives a good response on the cast shadow. The main difference between the responses below and what we saw last lecture with the simple cells is that the position information in the complex responses is less detailed. This is exactly what complex cells encode: they encode that there is some oriented structure in a local neighborhood but they don't indicate exactly where.



Estimating binocular disparity

Recall that the left and right eye images are send to the LGN but the signals are not combined there. The left LGN carries the signals from the right visual field and the right LGN carries signals from the left visual field. However, the left and right eye's signals for each field are fed into different LGN layers and then relayed separately to V1 where they provide the inputs to binocular simple and complex cells. I will not discuss binocular simple cells. Instead I will discuss just binocular complex cells. Before I do, let's consider what computational problem these cells are solving.

Consider a visual direction (x_0, y_0) in retinal coordinates, that is, relative to each eye's coordinate system. Suppose that the left and right images near this direction have similar intensities, except for a horizontal shift (see analglyph slide), that is, a binocular disparity. This shift might vary over the image, because the depths will vary and the shift depends on depth.

Near (x_0, y_0), the visual system could attempt to estimate the shift to be the value d that minimizes

$$\sum_{(x,y)\in\mathcal{N}(x_0,y_0)} (I_{left}(x+d, y) - I_{right}(x, y))^2$$

where the sum is over (x, y) coordinates in a neighborhood of (x_0, y_0). Note that $d > 0$ corresponds to a *leftward* shift of I_{left}. For the correct d, this would remove the disparity between the left and right images so they would be properly registered and their point-to-point difference would be 0.

The idea of this computation is that if you shift the left image by the correct disparity d, then the shifted left image should correspond pixel-by-pixel to the right image – at least in the local patch where the disparity is roughly constant. In that case, the above sum of squared differences should be 0 for the correct d. For other values of d, sometimes the left image will be brighter at a pixel than the right image and sometimes it will be darker, so the intensity difference at that pixel will be non-zero. We square the intensity differences because we only case how much it is different from 0, not whether it is positive or negative.[11] The idea for estimating disparity d near (x_0, y_0) for a particular left-right image pair is to choose the d value that minimizes this sum of squared differences.

While the above computational model works well (and is the basis for many computer vision methods for binocular stereo, the model is not biologically plausible. In the brain, binocular disparity estimation occurs in V1 which analyzes images using Gabor-like cells. We next consider a model based on such cells. We restrict ourselves to vertical oriented cells. (In Assignment 2, you will explore why.)

Up to now we have considered monocular complex cells in V1 which were constructed rom simple cells. We now consider binocular complex cells which are constructed from simple cells, namely Gabor cells for the left and right eyes. Using a similar idea as the computer vision method above, we could estimate the disparity d by finding a d shift that minimizes the follwing sum of squared differences:

$$(< cosGabor(x - x_0 - d, y - y_0), I_{left}(x, y) > - < cosGabor(x - x_0, y - y_0), I_{right}(x, y))^2$$

$$+ (< sinGabor(x - x_0 - d, y - y_0), I_{left}(x, y) > - < sinGabor(x - x_0, y - y_0), I_{right}(x, y))^2$$

Here the d shift is for the sine and cosine Gabor for the left eye, that is, the Gabors for the left eye are centered at $(x_0 + d, y_0)$ and the Gabors for the right eye are centered at (x_0, y_0).

[11] We could have alteratively taken the absolute value, and indeed some computational models do that.

 ©Michael Langer

The idea is that if we place a cosine Gabor template at $(x_0 + d, y_0)$ in the left image and at (x_0, y_0) in the right image and if d is the true disparity in the images – then the linear responses of the left and right eye cosine Gabors will have the same value, so if we subtract one from the other then we get 0. Similarly, the linear responses of the sine Gabors will have the same value, so if we subtract one from the other we get 0. The shift d that minimizes the sum of squared differences in the above expression would the best estimate of the disparity.

The above model works well in theory. Unfortunately, it doesn't describe binocular complex cell responses in V1. Rather, complex cells in V1 that are tuned to a disparity d have a *maximum* response (not a minumum response) at that disparity. So we need to change the model slightly so that it has this property. We do so by summing rather than taking a difference:

$$(< cosGabor(x - x_0 - d, y - y_0), I_{left}(x, y) > + < cosGabor(x - x_0, y - y_0), I_{right}(x, y))^2$$

$$+ (< sinGabor(x - x_0 - d, y - y_0), I_{left}(x, y) > + < sinGabor(x - x_0, y - y_0), I_{right}(x, y))^2$$

The intuition here is that when the shifted distance d corresponds to the correct disparity then the two cos Gabor responses will be identical and the two sine Gabor responses will be identical, as above. Now, when we sum them and square them, rather than getting a perfect cancellation, we get a big response. Several models[12] along these lines were proposed in the 1990's. These cells have peak responses to images of some disparity d to which the cell is tuned, and they are sensitive to particular orientation (usually vertical), and they don't care about the specific position of the (vertially) oriented structures within their receptive fields, just like the monocular complex cells we discussed earlier.

In particular, note what the above model of a binocular complex cell predicts would happen if the visual system were shown only one image – for example, if one eye were closed. That eye would have $I = 0$, and so it would not contribute to the response and the model would predict that the cell behaves just as a monocular cell – namely for the image given to the other eye. For example, if $I_{left} = 0$ everywhere, then the response would be

$$\sqrt{(< cosGabor(x - x_0, y - y_0), I_{right}(x, y) >)^2 + \ (< sinGabor(x - x_0, y - y_0), I_{right}(x, y) >)^2}$$

In the slides, I give an example of a response of binocular complex cell that is "tuned" to zero disparity ($d = 0$). This cell has its four Gabor receptive fields (sine and cosine, left and right eye) centered at the same position (x_0, y_0). I showed the responses of this cell to a single vertical line in the left and right eyes as a function of the x position of the line. Three different plots showed the responses for three image disparity values namely 2, 10, and 18. You can generate those plots yourself by running the code in

`http://www.cim.mcgill.ca/~langer/546/MATLAB/complexCells.zip`

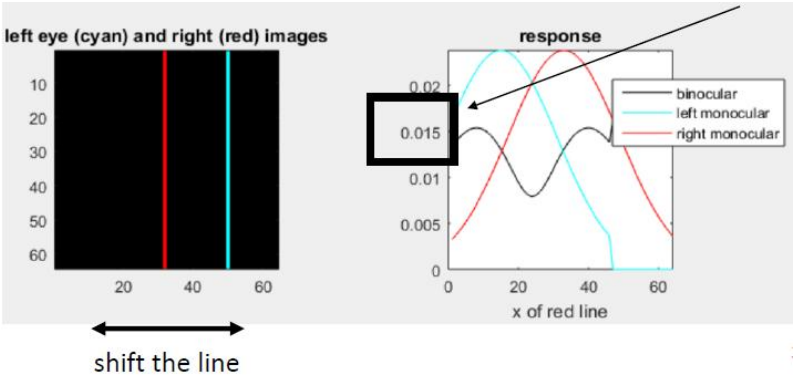
in particular, run `binocularComplexCell.m`. You will need to input a disparity value for the vertical line. You might also run the code `monocularComplexCell.m` . You don't need to inspect what the code is doing; indeed I suggest you *don't* since it is implemented using Gabor formulas that look different (but mathematically equivalent!) from what I wrote above.

[12] e.g. Ohzawa, Freeman, DeAngelis, Qian, Fleet and others

Below I show the responses for the case that the disparity between the left and right image is 18 pixels. The cell is tuned to zero disparity, so we would not expect the cell to have a good response to an 18 pixel disparity.

The arrow points to the value of the peak response value for this binocular complex cell, namely about 0.015. This is about one third the value of the response for the case of disparity = 2 pixels. (See the slides for that plot).

Also note that the peak response for the binocular cell shown for the 18 pixel disparity images is *less than* the peak monocular responses (green and cyan curves) to the same shifted line images. The monocular responses are what we would get if one eye were closed. To understand why the binocular response is so poor, note an image disparity of 18 pixels corresponds to roughly half a wavelength of the Gabor. (The Gabor's sinusoid is defined to have 2 cycles for 64 pixels, so half a wavelength is 16 pixels). So when the line's position in the left eye sits on a maximum of (say) of the sine Gabor, the shifted line in the other eye's image will fall on a minimum of that Gabor, and so there is a cancellation of values when the left and right sine Gabors are summed. (The cancellation may not be exact because the value of the Gaussian window will typically not be the same at the line and the shifted line in the other image.) The same argument can be made for the responses of the left and right cosine Gabor to the shifted line in the left and right image, namely the responses will be of opposite sign and will roughly cancel.



[I ended the class here. I will finish up the rest of this material next lecture, and then move on to motion processing in V1.]

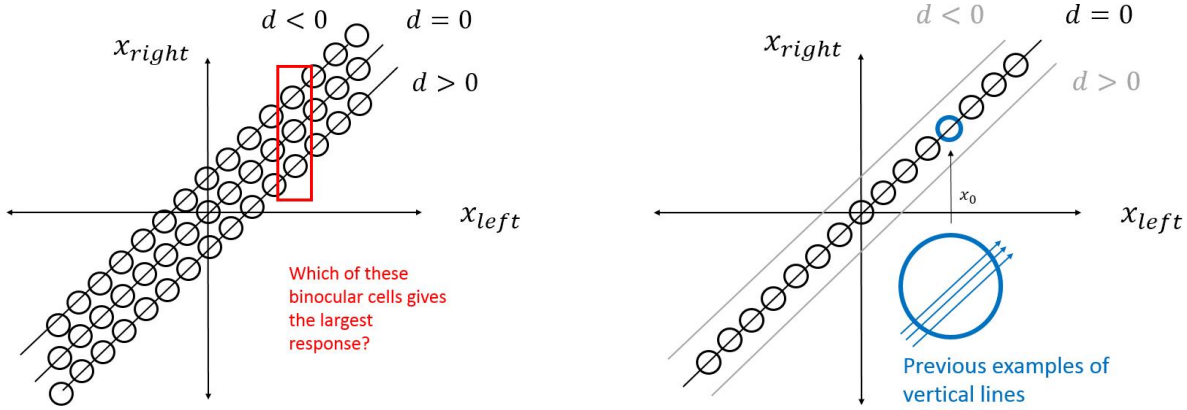
Disparity Space

The above example showed how one binocular complex cell – which was defined to be tuned to disparity of 0 – responds as a function of two parameters: the disparity of the vertical image line, and the location of this line. Let's next consider a slightly different question: how does a family of binocular complex cells (say each with peak tuning to a *different* disparity) respond to a single stimulus? That is a very important question, since the visual system estimates the disparity at an image location (x_0, y_0) by comparing the responses of this family of cells and choosing the disparity the cell that gives the biggest response.

Let's not deal with a numerical example here. (I'll save that for Assignment 2.) Instead let's just sketch out conceptually what it means to have a family of cells that are tuned to different

disparities.

The figure(s) below considers just a 1D case where the variable is x. Each binocular cell has two monocular receptive fields, centered at x_{left} and x_{right} and so we can indicate the binocular receptive field with a disk (or square, if you prefer – I use disk because I'm thinking of a Gaussian in each dimension and the product of two 1D Gaussians is circularly symmetric). If the monocular receptive field centers are at the same position in the two eyes, $x_{left} = x_{right}$, and then this cell would be tuned to a disparity of 0. This is just the case of the example above. If the monocular receptive field center for the left eye is to the right of the monocular receptive field center for the right eye, then this cell would be tuned to a positive disparity. Similarly, if the monocular receptive field center for the left eye is to the left of the monocular receptive field center for the right eye, then this cell would be tuned to a negative disparity. See the $d > 0$ and $d < 0$ zones in the plot below.



The idea for the figure on the left is that, for each x_{left} (say) the visual system "considers" the set of binocular complex cells whose left monocular receptive field is centered there. See cells highlighted in red. The best estimate of the disparity would correspond to that of the binocular cell in the (red) set that gave the largest response.

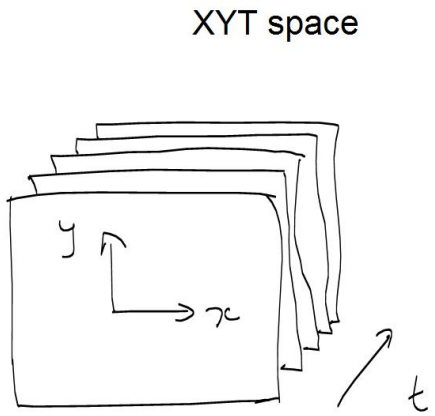
Finally, to explore this disparity space representation a bit more, consider again the vertical white line example. As the position of the line is shifted in both the left and right eyes, it's x value sweeps out a diagonal line in the disparity space. See blue arrows in the big disk. Three different line disparities are sketched there (say 2, 10, 18). Think of these as the three examples given in the slides.

We will return to these ideas again a few lectures from now, and in Assignment 2.

 ©Michael Langer

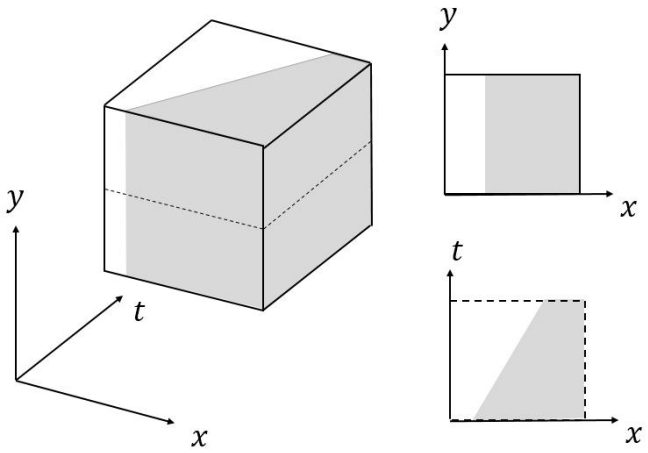
Time varying images

Up to now we have said very little about how images vary over time. But of course they often do. Let's think of an image as a function of x, y and t, namely $I(x, y, t)$.



A video is a sequence of image frames.

An an example image motion, consider a vertical intensity edge drifting to the right over time. The figure below shows a small space-time cube through which the edge passes, and it shows an XY slice and an XT slice through the cube. This edge drifts to the right with speed v_x so v_x is the slope of the edge in the XT slice (where slope is measured $\frac{dx}{dt}$, not $\frac{dt}{dx}$). As an aside for now, note that there could be a motion component in the y direction. However, this component would be impossible to *measure* since the image intensity does not vary in the y direction.

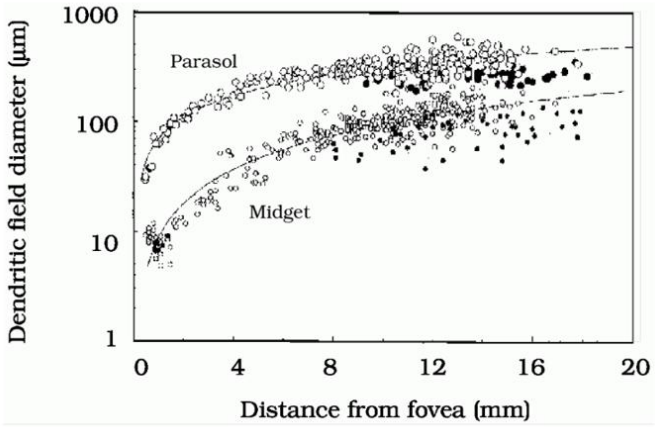


See the lecture slides for two other examples of $I(x, y, t)$. One is just a moving bar instead of a moving edge. The second is more interesting and shows a real video of a person walking from left to right. An XT slice reveals the motion pattern of the person's legs.

Retinal receptive fields and time-varying images

To model how the visual system estimates image motion, we need use model components that build on cells that respond to time varying images. Let's begin in the retina. Photoreceptors measure light intensity continuously over time (unlike digital video cameras which take discrete samples). A photoreceptor does not respond instantaneously, however. Rather there is a delay in the response. There is also temporal blurring, namely if we shine a very brief pulse of light on a photreceptor then the duration of its response will be longer than the pulse.

Retinal ganglion cells also have a temporal dependent response. It turns out there are there are two classes of ganglion cells. These two classes differ in several ways. One is the size of their receptive fields. As the figure below shows, the first class ("midget") of cells have dendrite (bush) diameters that are roughly factor of 10 smaller than the second class ("parasol") of cells. Notice that the sizes of both classes of cells increase steadily as we goes from the center of the field of view into the periphery. Think of the σ of the DOG functions as increasing with eccentricity. Both the difference between the sizes of midget versus parasol and the increase in size with eccentricity are big effects. Note the "x axis" (abscissa) in the figure is on a linear scale whereas the "y axis" (ordinate) is on a log scale.



The response (rate) of a ganglion cell at any time t will depend on the image in some local spatial neighborhood and on some local time interval *in the past.* Consider the XT slice for the cell shown below. Its temporal receptive field lies in the range $t < 0$ and this is meant to illustrate the receptive field weights for determining the response (firing rate) at time $t = 0$. The receptive field can be positive only for $t < 0$ since the cell's response cannot depend on something that hasn't happened yet.

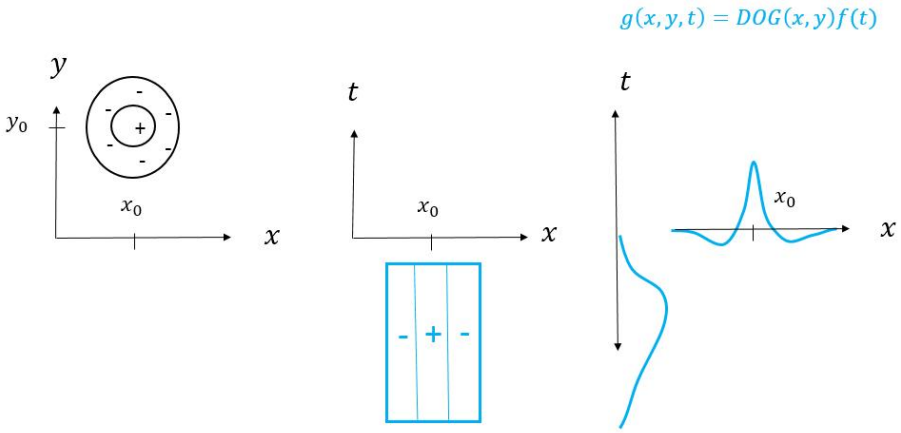
Note that the XT slice for this cell is shown for the slice through the center of the cell. A YT slice through the center of the cell would look similar. Think of rotating the cell's receptive field around its central vertical axis parallel to the T dimension. The cell has a cylinder shape in XYT.

I have given this cell a *separable* response function, namely a DOG in XY and a function $f(t)$ to describe the temporal dependence. Retinal cells do not have separable responses, in general. Intuitively, think of the DOG(x,y) profile as resulting from an excitatory effect of one spatial diameter and an inhibitory effect of a difference spatial diameter, and think of the excitory effect as

having some temporal dependence and the inhibitory effect as have a different temporal dependence. In that case, we might have instead

$$g(x, y, t) = G_{excite}(x, y) f_{excite}(t) - G_{inhib}(x, y) f_{inhib}(t)$$

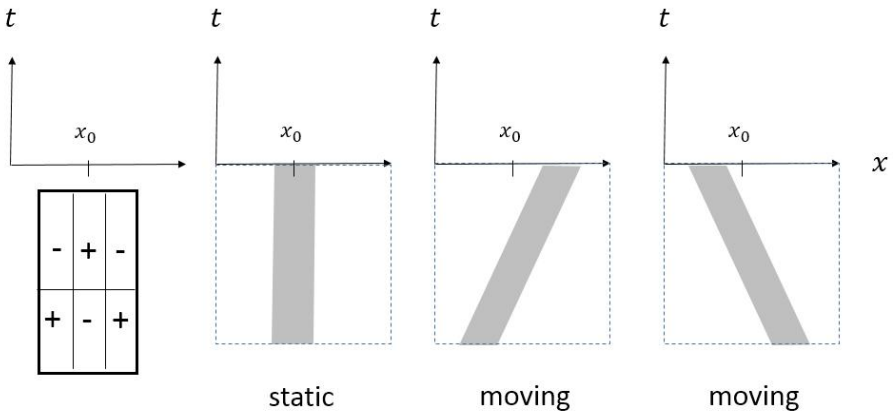
and in this case $g(x, y, t)$ would not be separable, even though the two terms that define it would be separable.



The cell below on the left (only XT shown) *is* sensitive to changes in the intensity over time. As shown in the slides, this cell also could be defined as a separable function. Here the dependence on time could have an excitatory part and an inhibitory part.

This cell would not respond well to a static intensity pattern since at each (x, y) position the pattern would be constant over time (by definition) and the cell's negative and positive weights would cancel. But notice that the cell would give a response to patterns that move over time, and the motion could be either to the left or right. For example, if the motion is at a particular slope in XT, it could cut across the + regions, or it could cut across only the - regions. If the cell's receptive field were stretched or shrunken over time, then the cell would be more sensitive to slow speeds or fast speeds, respectively (see slide).

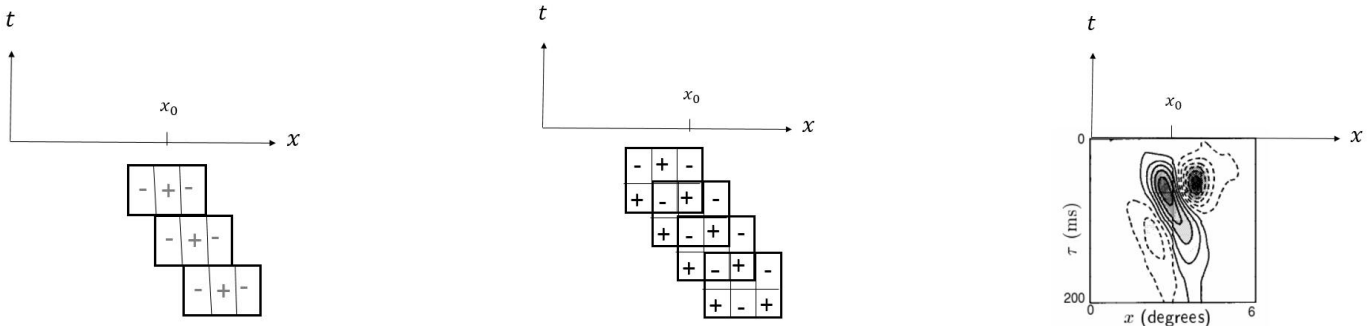
The arguments I am making here are in XT space only. If we consider the Y dimension also, then the arguments require a bit more work to understand and visualize. Let's not go there, since at this point I just want to make the basic point that variation in temporal sensitivity over time can result in sensitivity to motion direction. To really understand the motion system, we need to go beyond the retina (and LGN). Let's do that next.



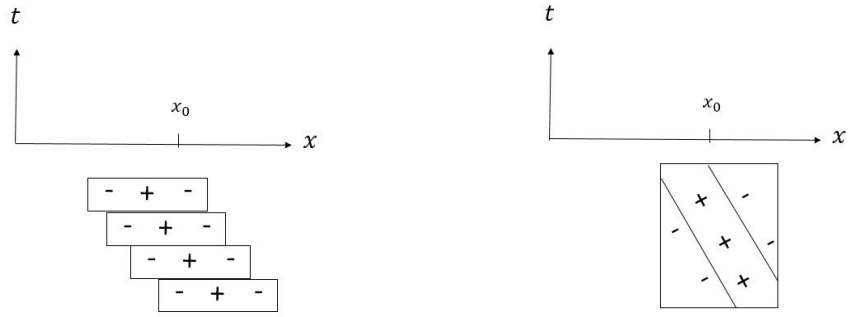
Directionally selective cells in V1

Many cells in V1 – both simple and complex – are sensitive to motion direction, and these cells are also sensitive to orientations (XY). How can the responses of such cells be modeled?

The first idea is that you can define a motion selective cell by summing up the outputs of cells in a time dependent way, namely by delaying the inputs of some cells relative to others. This is the idea of the *Reichart* motion detector that was proposed in the 1950's. The basic idea is illustrated below. The example on the left shows an XT slice through three DOGs that each have a short temporal sensitivity (relative to the previous plots – but keep in mind that the scales of these plots are arbitrary). The idea is that this illustrates one motion selective cell whose receptive field profile is defined by three DOG cells that are delayed in time. The example in the middle is a bit more complicated. Here the receptive field profile is composed from four cells that each have a temporal sensitivity (similar to the one shown at the top of this page, but compressed in time). Overlapping the receptive fields in space-time by delaying the cell inputs will again give rise to a cell that is motion selective. The example on the right is from a real cell. The + and - regions are indicated by iso-contour plots with solid curves indicating + and dashed curves indicating -. *Note that all these cells prefer motion to the left only,* not like the cell's on the previous page which responded to motion either to the left or right.



Let's next consider how to model cells that are both motion direction selective and orientation selective. One way is to stack together XY slices such that the receptive field is shifted by some amount (v_x, v_y) per time slice. (The slice can be thought of as have some duration Δt.) Another way is for the receptive field to be fixed over time, but have the + and - regions within the receptive field shift over time. See below.



One can show (and I will do in a future lecture) that cells whose XY receptive field slices are selective for particular orientations can only detect motion that is perpendicular to that orientation. For example, suppose a cell is sensitive to vertical orientations – e.g. either a cosine or sine Gabor whose underlying sinusoid varies in x only. If the image at that receptive field location contained a vertical line or edge and if that line or edge were moving vertically, then there would be no change in the image across the receptive field regardless of the speed of the line. As such, the cell would be blind to the vertical component of the motion. If, however, the line were to move horizontally instead, then the cell's response would *depend* on the speed of that horizontal motion, in particular, it would depend on how the line or edge fell on excitatory and inhibitory regions in the various XY slices over different times t.

One can model such orientation and motion sensitive cells using Gabor functions. As in the figure above, we could stack together identical Gabors that are shifted over time or we could stack together Gabors that have the same spatial receptive field over time but shift the phase of the Gabor over time, that is, gradually go from a sine to a cosine Gabor over time. (This is a new idea, which I did not mention in the lecture. But hopefully you see the intuition of the idea from the figure above right.)

Another way to define a Gabor is in terms of a sine or cosine function in XYT. Consider a 3D cosine function

$$\cos(\frac{2\pi}{N}(k_0 x + k_1 y) + \frac{2\pi}{T}\omega t)$$

where k_0 and k_1 are fixed integers between 0 and $N-1$, and ω is an integer between 0 and $T-1$. Note that we are sampling time discretely just as we are sampling space.

To understand this function, note the expression in the cosine's argument has a constant value c along a plane in XYT, namely

$$\frac{2\pi}{N}(k_0 x + k_1 y) + \frac{2\pi}{T}\omega t = c.$$

The value of the cosine changes with c and one goes from plane to other plane. Another way to think of it is in terms of a video. Fixing t corresponds to a single frame, and gives a 2D cosine function of (x, y). This cosine has k_0 cycles per N pixels in the x direction and k_1 cycles per N pixels in the y direction. For fixed pixel (x, y), the video changes like a cosine over time t, with temporal frequency ω cycles per T frames. (As an Exercise, figure out the speed of the wave as it travels over time.)

To make a 3D Gabor function, we multiply the 3D cosine or sine by a 3D Gaussian:

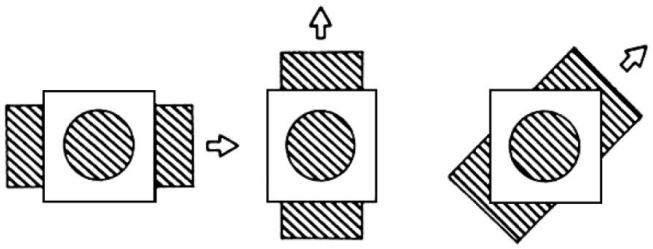
$$G(x, y, t; \omega, \sigma, \sigma_t) \ \cos(\frac{2\pi}{N}(k_0 x + k_1 y) + \frac{2\pi}{T}\omega t)$$

This Gabor is centered at the origin in XYT.

Aperture problem

A space-time Gabor cell will give its best response to an XYT image pattern that matches the Gabor profile. Roughly speaking this will be a moving bar or edge, depending on whether we have a cosine or sine Gabor, respectively. As discussed earlier, such a cell will be most sensitive to line or bar motions that are perpendicular to the spatial orientation of the cell. In particular the cells will be blind to motions that are parallel to the spatial orientation of the cell. I will be technically more precise about these claims later, but for now I just want to familiarize you with it.

The figure below illustrates the issue another way. Suppose we have an image consisting of parallel lines. (The same argument holds for just one line or edge.) Suppose we view that image through an aperture, which we can think of as the receptive field of some cell. In this aperture, we cannot distinguish several different motion vectors (v_x, v_y). We can only "see" the component of motion that is perpendicular to the orientation of the lines. This is known as the *aperture problem*.



The subtlety in the above discussion – and a possible source of confusion – is that we just discussed the orientation both of the cell receptive field and of the underlying image, and those seem like two very different things. They are. However, as we will understand better when we learn about linear image *filtering*, if we are only looking at the outputs (responses) of the Gabor functions then all we get to measure is the image component that has the same pattern as the response: only that component is able to pass through the (Gabor) filter.

 ©Michael Langer

In the first half of the lecture, I'll define a general computational problem of estimating motion in an image, and how to solve it. This abstract formulation of the problem is similar to classical computer vision methods for computing local image motion, and the ideas of these models have been used in many human and non-human vision experiments, to understand how the biological motion estimation systems works.

In the second half of the lecture I will sketch a computational model for motion processing in the brain which is in terms of the XYT receptive fields of the V1 cells which we discussed last lecture.

Image motion constraint equation

The computational problem of *local* image motion estation is to estimate the local image velocity (v_x, v_y), which is the vector describing the local change in position over time as points move across the visual field. We would like to make such an estimate at each image position (x, y) and time t, but we can only do so if there is intensity information present that indicates the motion(s). The intensity information we'll use here is the partial derivatives $\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}, \frac{\partial I}{\partial t}$ at each point. Under certain conditions, this turns out to be enough for a basic formulation of the problem. Note that we do need the intensity to be changing locally over position, since if intensity is constant across a local patch then we cannot say that anything in the patch is moving since all visual directions in the patch look the same.

Suppose that the points in a small local patch have image velocity (v_x, v_y), and I'll say what that means below for points to "move". For now, let's not worry about the units, whether the space units are pixels or photoreceptors, mm on the retina, or visual angle and or whether the time units are seconds, or some frames in a video. With velocity (v_x, v_y), each point by a distance $(v_x\Delta t, v_y\Delta t)$ in a time interval Δt. If the image intensities $I(x, y, t)$ are smooth enough that we can compute local derivatives, and write a Taylor series expansion of the intensities near (x, y, t) as

$$I(x + v_x\Delta t, y + v_y\Delta t, t + \Delta t) \;=\; I(x,\; y,\; t) + \frac{\partial I}{\partial x}v_x\Delta t + \frac{\partial I}{\partial y}v_y\Delta t + \frac{\partial I}{\partial t}\Delta t + H.O.T.$$

where H.O.T. stands for "higher order terms", namely higher than first order derivatives. The partial derivatives are evaluated at (x, y, t).

We now make a key assumption about the motion, namely that the image intensity of a moving point doesn't change over time – this is sometimes called *intensity conservation*. Thus, when a point moves from (x, y) to $(x + v_x\Delta t, y + v_y\Delta t)$ from time t to time $t + \Delta t$, respectively, we have

$$I(x + v_x\Delta t, y + v_y\Delta t, t + \Delta t) = I(x,\; y,\; t).$$

This lets us cancel these two terms in the Taylor series above. If we further ignore the higher order terms, then we have:

$$\frac{\partial I}{\partial x}v_x\Delta t + \frac{\partial I}{\partial y}v_y\Delta t + \frac{\partial I}{\partial t}\Delta t \;=\; 0$$

or

$$\frac{\partial I}{\partial x}v_x + \frac{\partial I}{\partial y}v_y + \frac{\partial I}{\partial t} = 0. \tag{5}$$

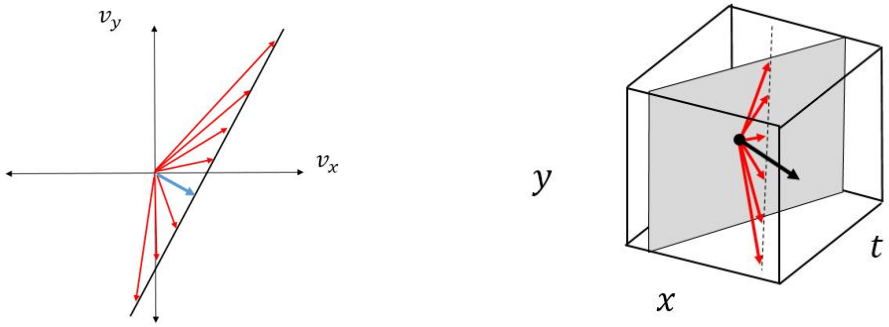
The latter is called the *motion constraint equation*. It expresses the relationship between the spatial and temporal derivatives of the image in terms of the image velocity (v_x, v_y). In particular, it

 ©Michael Langer

expresses a relationship between what we want to estimate – (v_x, v_y) – and the image quantities that we can directly measure, namely partial derivatives of intensity.

The intensity conservation assumption is similar to the assumption we made when discussing how to estimate binocular disparity in lecture 5. With binocular disparity, we assumed that the left and right eye images $I_{left}(x, y)$ and $I_{right}(x, y)$ were the same except for local horizontal shifts by the disparity d which was the quantity that we wanted to estimate. Here with image motion, we assume that image positions of projected 3D scene points are moving over time and that the image intensity of each projected point stays the same over time. Here the quantity we want to estimate is the local velocity (v_x, v_y).

Given a time varying image $I(x, y, t)$, one can compute the three partial derivatives. But can one estimate for (v_x, v_y) from these local derivatives at (x, y, t) alone? Unfortunately not, since Eq. (5) only gives one linear constraint at each point and this equation has two unknowns, namely v_x and v_y. All we can say is that (v_x, v_y) lies on a particular line in the 2D space of (v_x, v_y). See figure below. The shortest such candidate velocity vector (shown in blue) is normal to the line, and hence it is called the *normal velocity*.



Another way to express the same constraint is to consider XYT space and write Eq. (5) as

$$(\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}, \frac{\partial I}{\partial t}) \cdot (v_x, v_y, 1) = 0.$$

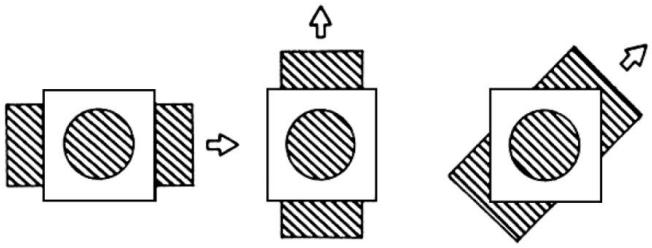
which says that the 3D vectors $(v_x, v_y, 1)$ in red are constrained to be perpendicular to the 3D intensity gradient vector.

Aperture problem

The ambiguity of the motion constraint equation is often called the *aperture problem*, since we can think of viewing the image through a small aperture in space-time such that only the first order partial derivatives can be computed. Note by "aperture" here, I'm not talking about a camera aperture like in lecture 2. Rather I'm just talking about a receptive field– a limited image window.

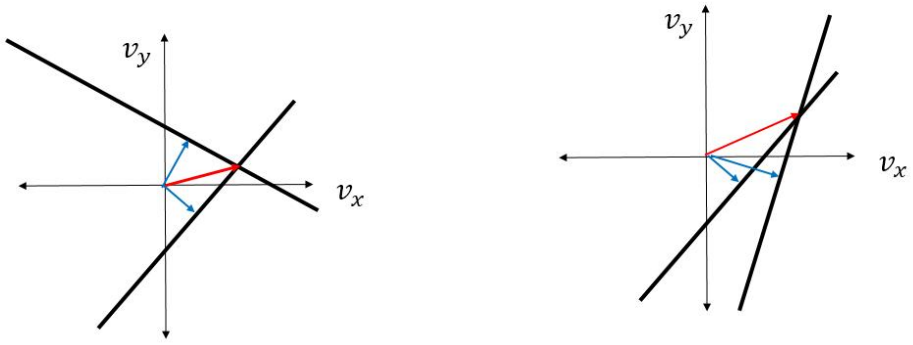
The aperture problem is more general than this, though. It applies anytime one has a moving 1D pattern. For example, the illustration below shows a set of oblique parallel stripes that are moving, either horizontally, vertically, or obliquely. Given only the motion in the aperture, one cannot say what the "true" velocity vector is. ASIDE: This problem is also related to the barber pole illusion:

`http://www.opticalillusion.net/optical-illusions/the-barber-pole-illusion/`



To avoid the aperture problem and estimate a unique velocity vector, one needs two or more such equations. The natural way to do so is to assume that the velocity vector (v_x, v_y) is constant over some local image region, and to combine constraints of Eq. (5) from two nearby points whose spatial gradients $(\frac{\partial I}{\partial x}, \frac{\partial I}{\partial x})$ differ. Since the two points would define two different lines in (v_x, v_y) space and the true velocity must lie on both these lines, one can solve for the true velocity vector by computing the intersection of the two lines. This is called the *intersection of constraints* (IOC) solution.

As examples, see figure below. The red vector is the IOC solution and the blue vectors are the normal velocities. The one on the right is counterintuitive because both lines have a normal velocity that is downward to the right, but the true solution is upwards to the right. This is surprising because one might have expected that the true solution should be "between" the normal velocity motion vectors defined by the two given constraints. For example, one might expect the solution to be the average of the two normal velocities.[13]



[13]There are many experiments in which humans *do* perceive image velocities using a "vector averaging" solution, rather than an IOC solution.

Gabor cells and the motion constraint equation

Recall the XYT Gabor cells that we discussed last class. These cells were defined by taking a sine wave with some spatial frequency (k_x, k_y) and some temporal frequency ω, and multiplying by a Gaussian window. Let's briefly review a few properties of the underlying sine wave. The sine way travels with velocity $(\frac{2\pi}{N}k_x, \frac{2\pi}{N}k_y, \frac{2\pi}{T}\omega)$ in XYT which is the vector perpendicular to constant values of the sine function. Note this is similar to the idea of the motion constraint equation, where we define "motion" to be a path of a point of some intensity. For the moving sine wave, we define the motion to the normal velocity only, namely the velocity normal to the set of points of constant intensity i.e constant value of sine.

The XYT Gabor cell response to a moving image $I(x, y, t)$ is defined in the usual way by taking the inner product of the Gabor function and the image function over XYT. The response at some time t for a Gabor cell located at position (x_0, y_0) will depend on the image in the recent past before t and in a spatial neighborhood of that point. We are modelling the space-time window by a Gaussian, but of course one needs to note that the response can only depend on the past and so a more accurate model would have a hard cutoff for the window. The same is true for space, in fact, as the Gaussian in theory has an infinite extent.

Such a Gabor will generally give its largest response when the image contains spatial variations in the direction $(\frac{2\pi}{N}k_x, \frac{2\pi}{N}k_y)$ and when the image component of the velocity in that direction matches the normal velocity of the Gabor's underlying sine wave. Note that the response depends on both of these factors.

How can a set of XYT Gabors cells be used to estimate the image velocity (v_x, v_y) at a point, when each of the Gabor cells can only detect a normal component of velocity? The answer to this question is similar to intersection of constraint solution described above. Suppose we wanted to design a cell at a further layer of processing, such that this cell would respond best when the image velocity near (x_0, y_0) had some given value (v_x, v_y). Call this a *velocity tuned cell*.

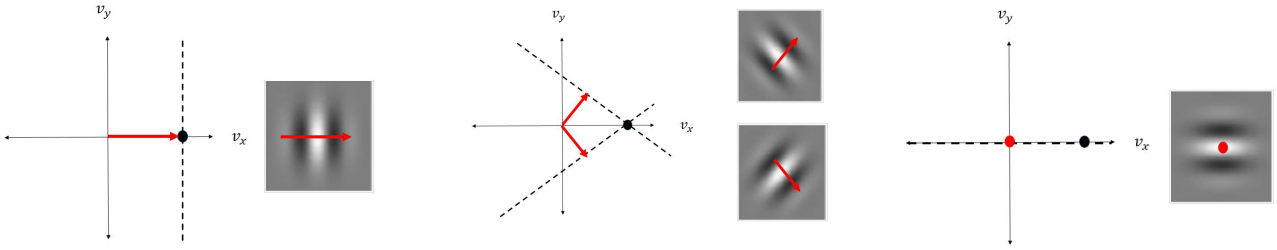
When velocity (v_x, v_y) is present, which Gabor cells would give a good response? The answer is: those Gabor cells whose underlying sine wave has normal velocity $(\frac{2\pi}{N}k_x, \frac{2\pi}{N}k_y, \frac{2\pi}{T}\omega)$ satisfying the motion constraint equation:

$$(\frac{2\pi}{N}k_x, \frac{2\pi}{N}k_y, \frac{2\pi}{T}\omega) \cdot (v_x, v_y, 1) = 0.$$

This defines a family of Gabor functions, namely those whose 2D motion motion constraint line passes through (v_x, v_y). As an example, take a motion $(v_x, v_y) = (v_0, 0)$ at some speed v_0 in the x direction.

The figure above illustrates several Gabors whose motion constraint lines intersect that particular velocity vector $(v_0, 0)$. The Gabor on the left has a normal velocity equal to $(v_0, 0)$. The two Gabors in the middle panel have normal velocities that are 45 degrees away from $(v_0, 0)$ and have a smaller speed, namely $v_0/\sqrt{2}$. The Gabor on the right has an orientation that parallel to the x axis, and it is most sensitive to zero normal velocity. That is, it prefers no motion in its normal velocity direction. (See Exercises.)

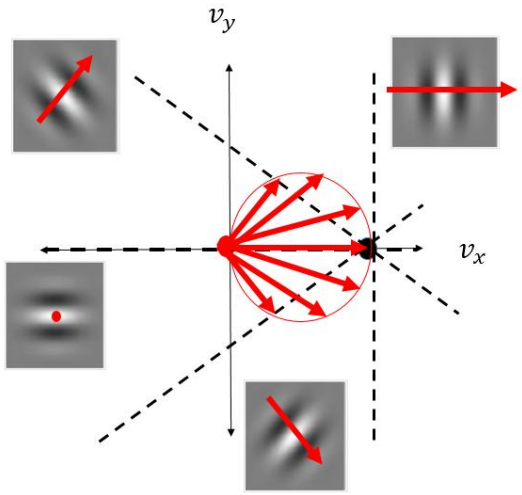
 ©Michael Langer



Velocity tuned cells in area MT (middle temporal lobe)

There are cells in the visual system that are velocity tuned. These cells are found in the temporal lobe in an area known as MT (middle temporal).[14] There are direct connections from V1 to MT.

The basic model for an MT cell that is tuned for motion in some direction (v_x, v_y) is illustrated below. This cell has excitatory inputs from a set of XYT Gabor cells, namely from those V1 cells whose underlying Gabors have normal velocity motion constraint line passing through or close to (v_x, v_y).



I will not give further details because it would take too long. If you are curious, have a look at this paper: `http://www.cns.nyu.edu/pub/lcv/simoncelli96-reprint.pdf`
I put one of the figures in the slides and briefly discussed it during the lecture.

[14] MT also has normal velocity cells, but I won't discuss them here.

 ©Michael Langer

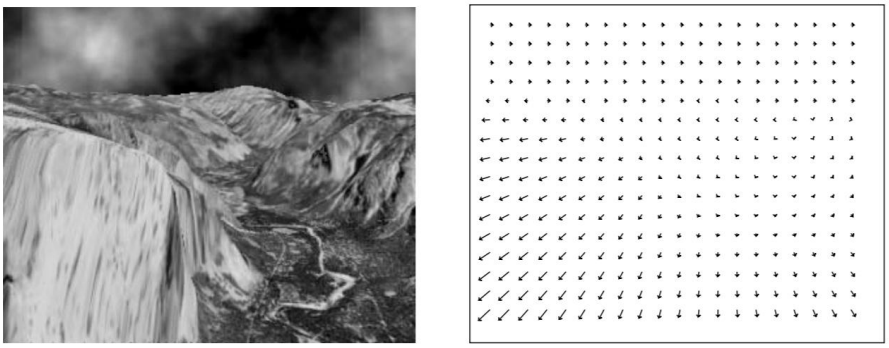
Motion Field

In lecture 7, we examined the computational problem of estimating the motion (v_x, v_y) at a point (x, y) in the visual field. The idea was to measure local derivatives of image intensity, and to use these derivatives to constrain the possible velocity vectors. The main assumption was that moving points do not change their intensity over time, and indeed that was the defining property of a moving point.

Today we are going to consider not just one point, but all the points (x, y) in an image. Lets say we have a depth map $Z(x, y)$ and we would like to know the image velocity (v_x, v_y) for each (x, y). We will assume that the motion is due to movement of the eye/camera. For simplicity, we will assume that the scene itself is static. In this case we can write down simple formulas for how the velocity (v_x, v_y) at each point in the image depends on the motion of the observer and on the depths of the scene points. These velocities define the instantaneous *motion field*.

As an example, consider a single frame from a video known as the *Yosemite sequence*. This was a computer graphics generated video of a fly through through the Yosemite Valley in California[15] Because was computer generated, it had a well defined depth map $Z(x, y)$ and one could compute a vector field — (v_x, v_y) at each pixel – shown on the right.

Today we will look at the motion fields that arise from different observer motions and different scene layouts. I'll first discuss observer motions that consist of a change in observer position, but no observer rotation.



Translation of viewer

We suppose that the viewer changes position over time by moving in a straight line over a short time interval, and does not rotate during this motion. Because the viewer observes the scene from different positions, the projected positions of objects in the image change too.

Suppose the camera translates with 3D velocity (T_x, T_y, T_z). For example, forward camera motion with unit speed is 3D velocity $(0, 0, 1)$. Rightward camera motion with unit speed is 3D velocity $(1, 0, 0)$. Upward camera motion is $(0, 1, 0)$. When the camera translates, the position of any visible point varies over time. In the camera's coordinate system, the position of the point moves with a velocity vector opposite to the camera. If the camera coordinates of a point at time

[15] It was often used in early computer vision research (1980's and 1990's) to test the accuracy of computer vision methods for estimating image motion.

$t = 0$ are (X_0, Y_0, Z_0), then at time t the point will be at $(X_0 - T_x t,\ Y_0 - T_y\ t,\ Z_0 - T_z\ t)$ in camera coordinates.

Now let's project the 3D point into the image plane. How does the image position of this point in the image vary with time? We will use a visual field projection plane $Z = f$ *in front of the viewer* and express the position in radians. The image coordinate of the projected 3D point is a function of t, namely,

$$\frac{1}{f}(x(t), y(t)) \ = \ (\frac{X_0 - T_x\ t}{Z_0 - T_z\ t}, \frac{Y_0 - T_y\ t}{Z_0 - T_z\ t})$$

Taking the derivative with respect to t at $t = 0$ yields an *image velocity vector* (v_x, v_y) in radians per second:

$$(v_x, v_y) = \frac{d}{dt}(\frac{x(t)}{f}, \frac{y(t)}{f})\ |_{t=0} \ = \ \frac{1}{{Z_0}^2}(-\ T_x\ Z_0\ +\ T_z\ X_0, -\ T_y\ Z_0\ +\ T_z\ Y_0). \tag{6}$$

The velocity field depends on image position (x, y) and on the depth Z_0 and on (T_X, T_Y, T_Z). We next decompose the velocity field into a lateral component and a forward component.

Lateral component of translation

Consider the case that $T_z = 0$. This means the viewer is moving in a direction perpendicular to the optical axis. One often refers to this as *lateral motion*. It could be left/right motion, or up/down motion, or some combination of the two. Plugging $T_z = 0$ into the above equation yields:

$$(v_x, v_y) = \frac{1}{Z_0}(-T_x, \ -T_y) \ .$$

Note that the direction of the image velocity is the same for all points, and the magnitude (speed) depends on inverse depth.

A specific example is the case $T_y = T_z = 0$ and $T_x \neq 0$. The motion field corresponds to an observer looking out the side window of a car, as the car drives forward. In the case that the scene is a single ground plane, recall the relation $Z = \frac{h}{y}$ from lecture 1. The image velocity is then

$$(v_x, v_y) = -\frac{T_x}{h}(y, 0).$$

The minus sign is there because the image motion is in a direction opposite to the camera motion. The speed is proportional to y is a result of the depth of the ground plane being inversely proportional to y, e.g. the depth is ∞ for $y = 0$ which is the horizon. See the examples given in the slides 9, 10 which show two frontoparallel surfaces and a ground plane, respectively.

Lateral motion is very important for vision. Our eye position almost always shifts over time. If when we think we are still, in fact we are continously shifting our weight and changing our pose. This is in part to relieve our joints and muscles, but it also provides visual information for maintaining our pose. As we lean to the left, the visual scene drifts slightly to the right, and vice-versa. We rely on this motion field the stabilize ourselves with respect to the surrounding world.

This reliance of the motion field becomes evident when we stand in front of a cliff, so that the ground in front of us is tens or hundreds of metres away. Normally, the ground in front of us moves opposite to us as we sway slightly back and forth. But when we stand in front of a cliff, there is

 ©Michael Langer

essentially no lateral motion (visual) field because Z is so big and $\frac{1}{Z}$ is near 0. This lack of motion is problematic for visually controlling our posture. It is the main reason we get dizzy (vertigo) when we stand at the edge of a cliff. More generally, it is one of the factors that contribute to a fear of heights. It is also why it is more difficult to do fancy balance poses in yoga when you are looking up at the sky or a high ceiling than when you are looking down at the ground in front of you.

Forward translation

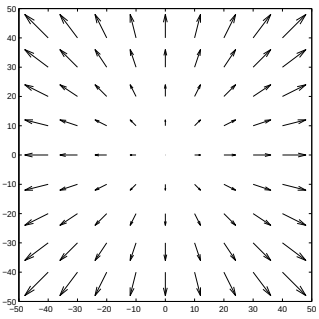
In case of forward translation ($T_x = T_y = 0$ but $T_z > 0$), Eq. (6) becomes

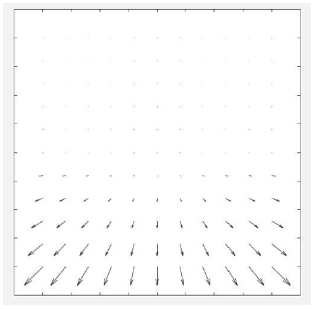
$$(v_x, v_y) \;=\; \frac{T_z}{Z_0}\,(\frac{x}{f}\;,\;\frac{y}{f}). \tag{7}$$

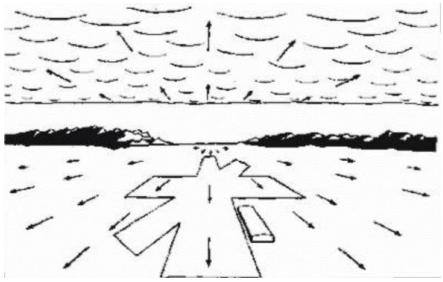
By inspection, this field radiates away from the origin $(x, y) = (0, 0)$. Also, the speed (i.e. the length of the velocity vector) is :

- proportional to the angular distance $\sqrt{(\frac{x}{f})^2 + (\frac{y}{f})^2}$ from the origin
- inversely proportional to the depth Z_0
- proportional to the forward speed of the camera T_z.

See the example on the left.







The middle panel shows the case of a ground plane, which has depth map $Z = h\frac{f}{y}$ and so:

$$(v_x, v_y) \;=\; = \frac{T_z}{h}\;(\;(\frac{x}{f})(\frac{y}{f})\;,\;(\frac{y}{f})^2)$$

Note that in this case the velocities near the horizon $y = 0$ are small. This is a familiar case of walking forward. Another situation in which this arises is what a pilot sees when landing a plane. This scenario was one of the first applications in which psychologists studied this 'direction of heading' problem. (The illustration on the right above is taken from a classic book by J. J. Gibson in 1950.)

General (non-lateral) translation

In the case that we do not have pure lateral translation, i.e. if $T_z \neq 0$, we can write the motion field slightly differently. Putting the lateral and forward components of the motion field together, we get

$$(v_x, v_y) = \frac{1}{Z_0}(-T_x, \ -T_y) \ + \frac{T_z}{Z_0}\ (\frac{x}{f}\ ,\ \frac{y}{f}) \tag{8}$$

$$= \frac{T_z}{Z_0}(-\frac{T_x}{T_z}, \ -\frac{T_y}{T_z}) \ + \frac{T_z}{Z_0}\ (\frac{x}{f}\ ,\ \frac{y}{f}) \tag{9}$$

$$= \frac{T_z}{Z_0}\ (\frac{x}{f} \ -\frac{T_x}{T_z}, \ \frac{y}{f} \ -\frac{T_y}{T_z}) \tag{10}$$

Define the special image direction:

$$(\frac{x_0}{f}, \frac{y_0}{f}) \ = \ \frac{1}{T_z}(T_x, T_y) \tag{11}$$

which is called the *heading direction*. Then,

$$(v_x, v_y) = \frac{T_z}{Z(x,y)}(\frac{x - x_0}{f}, \frac{y - y_0}{f}) \ .$$

Notice that the translation field diverges away from the heading direction. See example in slides.

Computing the heading direction (from MT to MST)

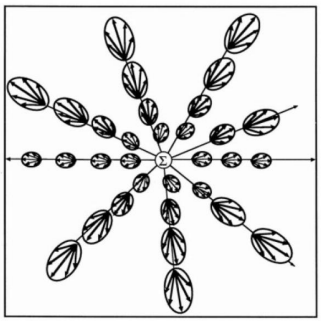
How can a visual system estimate the direction in which it is heading? There are basically two steps. First, estimate the local velocities at as many points (x, y) as possible. Second, estimate a direction from which all velocity vectors point away.

As I discussed last lecture, the brain carries out the first step itself in two parts. (Cells in V1 measure normal velocity components, and cells in MT combines these normal velocity estimates to estimate velocities.) How does one brain compute the heading direction from these velocity estimates? This computation occurs in another area of the brain, known as MST which is close to MT. MST stands for "medial superior temporal". "Medial means inside (as opposed to lateral). Superior means on top. "Temporal" refers to temporal lobe.

Cells in MST receive direct inputs from cells in MT. MST cells have very large receptive fields. Many of these cells are sensitive to expanding patterns within their receptive field. You can think of these cells as getting excitatory input from MT cells whose tuned velocities (v_x, v_y) form an expanding patter. Different MST cells are sensitive to a variety of motion field patterns – not just expanding. I sometimes refer to these as "global" motion patterns because the receptive fields are so big.

The figure below illustrates the receptive field structure of an MST cell. At each location of the receptive field, the cell gets excitory inputs from a (v_x, v_y)-sensitive cell in area MT. Each of the ellipses in the figure illustrates one MT cell. Only about 30 such cells are shown. Each MT cell itself receives excitatory input from a set of V1 cells, namely from those V1 cells whose spatial orientation and normal velocity peak sensitivity is consistent with the velocity of the MT cell. (The MT cell's responses were sketched out last lecture.)

 ©Michael Langer

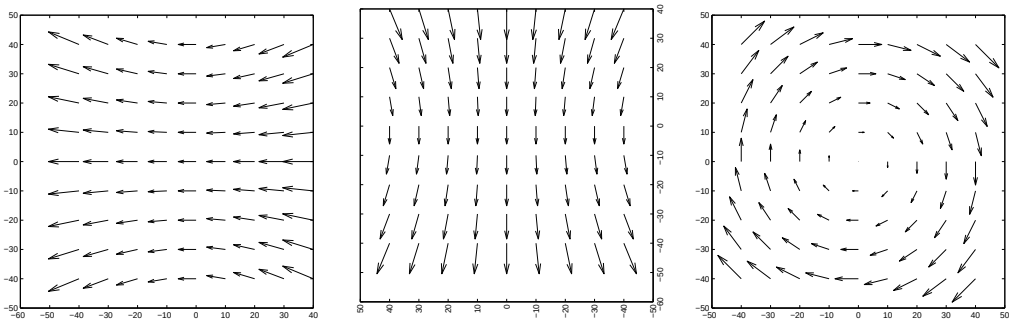


Rotation of viewer

The viewer can not only change position over time. It can also change the direction of gaze over time. This can be done by rotating the head, or by rotating the eyes within the head, or both a the same time. Note that when the viewer's head rotates, this induces both a translation and a rotation, since the viewer rotates the head around some point in the neck. Lets only concern ourselves with pure rotation for the moment.

When the viewing axes rotates smoothly, a smooth motion field is produced on the retina. If one projects onto an image plane as we did for translation, then one can derive equations for the motion field. These equations depend on the axis that the viewer is rotating around, and on the speed of rotation. These equations are a bit more complicated to derive, so I will skip them and just show pictures.

The two fields on the left show the motion for panning (left and right) and tilting (up and down). The velocity vectors within each of these motion fields are not exactly parallel. The slightly curving of the fields is a subtle artefact that is due to projection onto plane $Z = f$. We can ignore this "second order" detail. The roll field on the right occurs when the rotation axis is the axis in which the viewer are looking. In this case, the speed increases radially away from the axis, like a spinning wheel.



$\boldsymbol{\Omega} = (0, 1, 0)$ tilt

$\boldsymbol{\Omega} = (1, 0, 0)$ pan

$\boldsymbol{\Omega} = (0, 0, 1)$ roll

Eye rotations are controlled by muscles that are attached to the side of the eyeball. See the figure in the slides. There is a pair of opposing muscles for each of the three rotation directions. These muscles are signalled directly by motor (output) neurons whose cell bodies are in the midbrain.

The axons from these motor neurons are bundled together into the *oculomotor nerve*. This nerve and other nerves carry information such as blink commands, accommodation controls, and pupil contraction controls. Some of these signals are computed directly in the midbrain and nearby structures, without going to the cortex. This allows very fast feedback to control the eyes. We will discuss an example next.

VOR (vestibulo-ocular reflex)

One fundamental eye movement is called the vestibulo-ocular reflex or VOR. When the head moves – whether it is translation or rotation or both – the motion causes a shift in the retinal image. The role of VOR is to quickly sense this head motion and to rotate the eyes to compensate for it and to keep the retinal image as stable as it can. Look at one of the words on this page (or screen) and then rotate or translate your head left and right and remain fixated on that word. You will find this is very easy to do, and you can move your head quite quickly and maintain your gaze on the object. The VOR plays a central role in this.

The VOR depends on the vestibular system (V) which is part of your inner ear. The vestibular system senses linear and rotational acceleration of the head. There are two parts – see slides. The first part detects rotational acceleration. It consists of three loops called the *semi-circular canals*. These are filled with fluid, and when the head rotates, the fluid moves in the canal and this fluid motion is sensed by little mechanical receptors. (Details omitted.) If the head continues to rotate, the fluid drags along and eventually has the same speed as the canal itself. At that point, if the head *stops* rotating, then the fluid keeps going and again the system senses the fluid motion relative to the canal, which sends a (erroneous) signal that the head is rotating again. This is what happens when you spin around 10 times, and then stop spinning. (And you fall down.)

The second part of your vestibular system measures linear acceleration. How does this work, intuitively? Imagine a grassy surface with stones sitting on it. If the surface is suddenly moved sideways, then the stones will roll relative to the surface. If the surface moves upwards, then the stones will press down on the surface and if the surface noves downward, then the stones will press less (like when the elevator goes up or down). In the vestibular system, the "grass" is a set of mechnical receptors and the stones are just that – small stones (called otoliths).

The VOR is extremely fast, and the reason this is possible is that the circuit is so short (see below). VOR does not depend on a visual signal, and indeed works even when the eyes are closed. You can verify this for yourself. Look at some object in the scene, and close your eyes. Now shake your head back and forth and keep trying to fixate the imagined location of the object. Your eyes will rotate as you do so, but will keep fixation (within say 5 deg of visual angle) on whatever you had been looking at before you closed your eyes.

Note the vestibular system doesn't measure the rotation of head directly, but rather it measures changes in rotational velocity over time (or rotational acceleration), and it doesn't measure the translation (T_X, T_Y, T_Z) directly but rather it measures the change $\frac{d}{dt}$ in the translation velocity over time. The system needs to integrate the changes in rotation or translation over time in order to maintain an estimate of the rotation velocities or the translational velocities themselves.

Smooth pursuit

Another important type of eye movement is *smooth pursuit* eye movements. These are voluntary movements that keep a desired object on the fovea. An example is the eye movements that you make when you visually track something moving the world e.g. when you watch a dog walk by. These eye movements are relatively slow. For example, if I move my finger in front of your eye, you can keep your fovea tracking on my finger, but only up to some limited speed.

The reason for the speed limitation is that this smooth pursuit system needs to process the motion. If the image of the object that you want to track starts slipping from your fovea, it means that you are moving your eye too slowly or quickly. Your visual system need to estimate this slippage. This requires that the signal reaches all the way to area MT. That is a few stages of processing just to detect that the eye movement is too slow! The brain also needs to compute the correction and send that signal back to the midbrain where the motor correction can be computed and send to the muscles that control the direction of the eye. (The various pathways are well known, but I am omitting the details here since I just want to make a general point about why the system is relatively slow.)

[ASIDE: The following was only briefly mentioned in the lecture. I include it here to be more complete.]

Note that eye movements (VOR and smooth pursuit) produce rotational components in the motion field. For VOR, the rotational components are meant to cancel out the rotational componets that are due to head motions. If VOR is working properly, then there is no net rotational motion field from head motion + VOR. However, there may still be a rotational component to the motion field from smooth pursuit eye movements. This rotation motion field is added to the translational field, and so if you are walking (translating) while visually tracking some other object (perhaps stationary, perhaps not) then your motion field will be the sum of a translation and rotation field. See the slides for an example.

Disentangling the translation (walking) and rotation (smooth pursuit) component fields would be a difficult computational problem, if the visual system could only rely on visual input to do so. Fortunately, since the visual system controls the smooth pursuit, the system "knows" how the eye is rotating. This information could help to disentangle the translation and rotation components of the field.

This lecture carries forward some of the topics from early in the course, namely defocus blur and binocular disparity. The main emphasis here will be on the information these cues carry about depth, rather than on how blur and binocular disparity information is coded in the visual system.

Depth from Blur

Lecture 2 examined how blur depends on depth. You learned about accommodation, namely by changing the focus distance of the eye, you can bring surfaces at certain depths into sharp focus and cause surfaces at other depths to become more blurred. Accommodation can be used to judge the 3D depth of points – at least in principle, since the eye controls accommodation. This does not seem to be how people judge the depths of all points in a scene, however. We do not sequentially "scan" through *all* focus settings, and the reason we don't is presumably since estimating depth is just one of many things the visual systems needs to do. That said, some depth information is available from focus, and so we would like to better understand what that depth information is and how the visual system might use it.

One idea is that if the visual system can estimate the depth at which it is currently focusing (by controlling the shape of the lens to bring a desired point into focus), and if it can also estimate the current aperture (which it can, since the eye controls the pupil size), and if it can estimate the blur width at various points – for example, the width of a blurred edge – then the visual sysem can compute the distance in diopters between any blurred point and the focal plane. Recall the relation derived in Exercise 2 Question 4:

$$\text{blur width in radians} = A \mid \frac{1}{Z_{object}} - \frac{1}{Z_{focalplane}} \mid$$

Note the absolute value on the right side of this equation, which is due to the fact that blur occurs for points farther than the focal plane and also points closer than the focal plane. From blur alone, we have a two -fold depth ambiguity.

Interestingly, the eye does not hold the focal distance constant. Rather the eye's focus distance is continuously oscillating. The amount of oscillation is small: the amplitude is roughly 1/3 of a diopter. But this may be enough to resolve the ambiguity. For example, if an object is closer (or further) than the focal plane, then moving the focal plane closer to the eye will decrease (or increase) the blur; the opposite holds when the focal plane is moved further from the eye. In particular, the two-fold ambiguity mentioned in the previous paragraph is easily resolved – at least in principle.

Blur on a Slanted Plane

It is very common to have objects in a scene that are large planes, or at least can be approximated as such over a large region. Examples are the ground we walk on, walls, and ceilings. Let's consider the blur that arises on a slanted plane.

Let the scene depth map be a slanted plane,

$$Z = Z_0 + mY$$

where Z_0 is the depth of the plane at the point that intersects the Z axis. Assume that we are focussed on that depth. Note that this scene has a floor or ceiling slope only. A more general plane would have a slope component in the X direction also.

Recalling $\frac{y}{f} = \frac{Y}{Z}$, we divide by Z to get

$$1 = \frac{Z_0}{Z} + m\frac{y}{f}$$

or

$$Z_0(\frac{1}{Z_0} - \frac{1}{Z}) = m\frac{y}{f}.$$

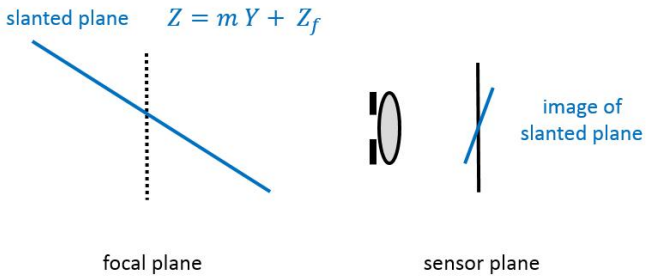
From page 1, the blur width w in radians is

$$w = A|\frac{1}{Z_0} - \frac{1}{Z}|$$

and so

$$w = \frac{mA}{Z_0} \mid \frac{y}{f} \mid.$$

Thus, the blur width on a slanted plane increases linearly with the image coordinate y. This linear dependence scales with A, with the focus distance Z_0 in diopters, and with the slope m of the plane.



Long ago photographers tried to take advantage of this linear dependence. One idea was to build a camera whose sensor plane was slanted slightly in the direction of the depth gradient.[16] When the slant is chosen appropriately, the sensor plane becomes aligned with the focussed image of the points on the slanted plane and one obtains a perfectly focussed image – something that otherwise is not possible to do, especially not with a wide aperture.

Another idea is to tilt the lens in in the opposite direction, so as to *increase* the gradient of blur in the y direction. Examples are shown below. The perceptual effect is that one misinterprets the overall scale of the scenes: the scenes appear to be photos of small toy worlds, rather than photos of large scale environments. While there is some controversy on what is causing his perceptual effect, the general idea is that the large blur gradient needs to be 'explained' by one of the variables in the above equation. Having an extremely large ground plane slant m is not possible, since the perspective cues suggest a particular slant m which is not extreme – I'll discuss perspective cues next lecture. Having a large aperture A is also not possible, since the aperture needed to get such blur gradient in a large scene would be much larger than the aperture of our eyes – we simply don't experience large scale scenes with such a blur gradient. The most likely culprit seems to be the variable Z_0 which is the distance to the point on the optical axis – indeed, making Z_0 small by scaling the whole scene down would explain the large blur gradient, while holding perspective cues constant.

[16] The configuration was a called a *tilt-shift* lens. Details omitted since I just want to give the basic idea.

Binocular Stereopsis (and its relation to blur)

We have discussed the geometry of stereo a few times in the course, for example, in lectures 1 and 6. If the eyes are parallel, then

$$\text{disparity (radians)} = \frac{x_l}{f} - \frac{x_r}{f} = \frac{T_X}{Z}$$

and if the left eye and right eye are rotated by angles θ_l and θ_r relative to the Z axis, then :

$$\text{disparity (radians)} = (\frac{x_l}{f} - \frac{x_r}{f}) - (\theta_l - \theta_r) = T_X(\frac{1}{Z} - \frac{1}{Z_{vergence}})$$

It is easy to show that $\theta_l - \theta_r$ is the *vergence angle*, namely the angle defined by the three points (left eye, scene point where eyes are verging, right eye).

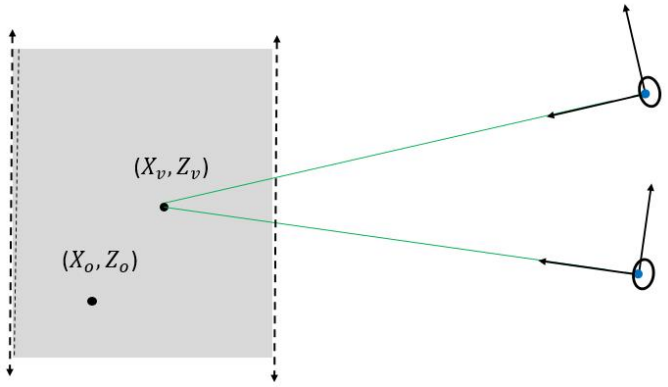
Since the brain controls the vergence, the brain in principle deterimines the depth on which the eyes are verging, and so this depth information is available. This is similar to accommodation as we'll see below, namely the brain controls the shape of the lens and so the brain 'knows' the depth of points that are in focus. Indeed the mechanisms of binocular vergence and accommodation are coupled: when the vergences angle is changed, so does the power of the lenses – at least to the extent possible. (Recall that as you get older, the range of accommodation decreases.) I will discuss blur again a bit later in the lecture.

Crossed and uncrossed disparities, binocular fusion

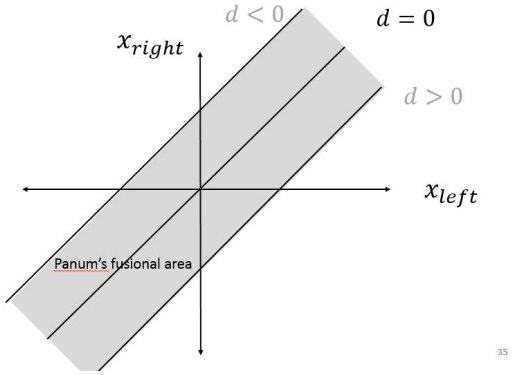
Point that are closer to the eye than the vergence distance have positive disparity, or *crossed* disparity since one needs to cross one's eyes to bring the disparity of such points to 0. Points that are further than the vergence distance have negative disparity, or *uncrossed* disparity since one uncrosses one's eyes to bring the disparity of such points to 0.

If the magnitude of the disparity of a 3D is small enough, then one can perceptually fuse the left and right images of the point, rather than seeing two images of these points – i.e. 'double vision' or *diplopia*. This limited range of fusion disparities defines *Panum's fusional area* which is equivalently a range of depths in front of and beyond the vergence depth – see grey area in figure below below.

That is, for any vergence distance, Panum's fusional "area" is really a 3D volume such that visible points in this volume are fused by the visual system.[17] One often refers to the largest disparity that can be fused as *Dmax*. Experiments have shown that Dmax depends on several scene factors, including the visual angle of the object being fused, the eccentricity, and the pattern on the object.



Panums fusional area can also be illustrated in disparity space, as shown below.



Binocular disparity and blur

Binocular disparity and blur give very similar information about depth.

$$\text{disparity in radians} = T_X \mid \frac{1}{Z_{object}} - \frac{1}{Z_{vergence}} \mid$$

where T_X is often called the 'interocular distance' or IOD.

$$\text{blur width in radians} = A \mid \frac{1}{Z_{object}} - \frac{1}{Z_{focalplane}} \mid$$

[17] In fact the iso-disparity surfaces in the scene are not depth planes, since the retina is not a planar receptor array. But let's not concern ourselves with this detail.

So, if the visual system is verging on the same depth as it is accommodating then

$$\frac{\text{disparity}}{\text{blurwidth}} = \frac{T_X}{A}.$$

With $T_X = 6cm$ and $A = 6mm$, this would give a ratio of 10:1. Indeed one does typically attempt to accommodate at the same depth as one verges – since the scene point we are looking at should be in focus. The above relationship specifies how two cues covary for points that are not on the vergence/accommodation distance. This covariance is presumably important for controlling accommodation and vergence. Indeed the neural control systems that control vergence and accommodate are closely coupled. (Details omitted.)

This close coupling between the accommodation and vergence systems is a problem for 3D cinema. Binocular disparities are used in 3D cinema to make the scenes appear 3D, and yet images are all presented at the display plane – the movie screen or your TV or laptop screen. When you look at an object that is rendered in 3D, you make a vergence eye movement to bring that object to zero disparity. However, normally your accommodation system follows along and adjusts the lens power so that you are accommodating at the same depth that you are verging. But for 3D cinema that is incorrect, since the screen is always at the same depth. If you verge your eyes to a point that is rendered with a non-zero disparity on the screen, then you will verge to a 3D point with depth difference than the screen. In that case, your accommodation system will get conflicting information if it follows the vergence system, namely the image on the screen will become blurred. The system will try to find a different depth to focus on to bring the image into sharp focus. However, this will drive the vergence back to the screen and away from the object that you are trying to verge on. There is no way to resolve this conflict, unless you can decouple the two systems. Most people cannot do this, which is why 3D displays give many people headaches and general viewing discomfort.

The other problem with 3D cinema is that the disparities are designed for a particular viewing distance and position, namely in the middle of the cinema audience. Anyhow who has sat in the front row at a 3d movie or way off to the side is familiar with this problem.

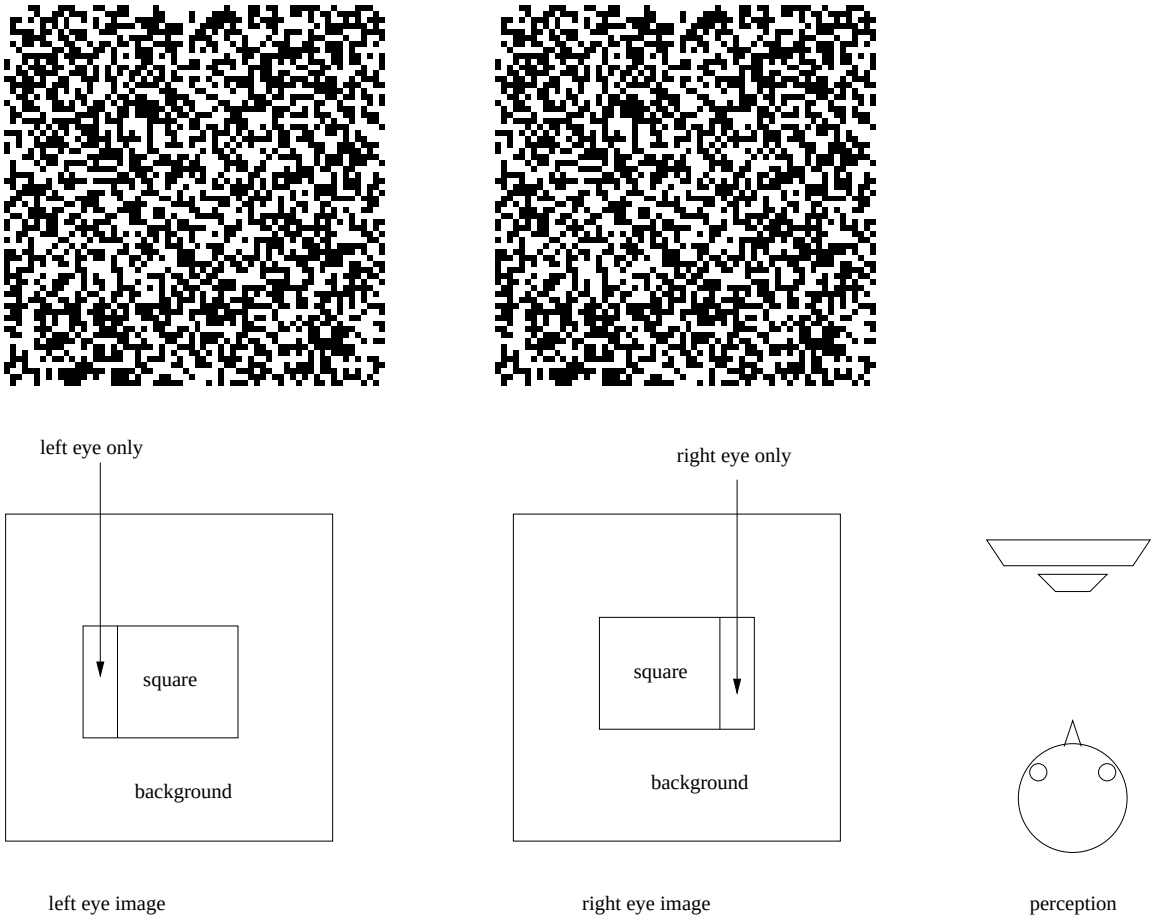
Random dot stereograms

One longstanding question in binocular stereovision is: How does the eye/brain match corresponding points in the left and right images? Up until the middle of the 20th century, it was believed that the brain solved this correspondence problem by finding some familiar pattern such as a line or edge or corner in the left image and matched it to the same familiar pattern in the right image, and vice-versa. This makes sense intuitively, since it was known that the brain follows certain rules for organizing local image regions into small groups of patterns. (We will discuss "perceptual organization and grouping" later in the course.)

In the 1960's, engineers and psychologists became interested in the process of binocular correspondence and fusion, and started using computers to address the problem – they did perception experiments using computer generated images. Computer scientists also began experimenting with writing computer vision programs using digital image pairs. One important type of image that was used was the *random dot stereogram* (RDS). RDS's were invented by Bela Julesz at Bell Labs. An RDS is a pair of images (a "stereo pair"), each of which is a random collection of white and black (and sometimes gray) dots. As such, each image contains no familiar features. Although each image on its own is a set of random dots, there is a relation between the random dots in the two images.

The random dots in the left eye's image are related to the random dots in the right eye's image by shifting a patch of the left eye's image relative to the right eye's image. There is a bit more to it than that though as we'll see below.

Julesz carried out many experiments with RDSs. These are described in detail in his classic book from 1971 and in a paper[18]. His results are very important in understanding how stereo vision works. They strongly suggest the human visual system (HVS) does not *rely* on matching familiar *monocular* features to solve the correspondence problem. Each image of a random dot stereogram is random. There are no familiar patterns in there, except with extremely small probability.



The construction of the random dot stereograms is illustrated in the figure below. First, one image (say the left) is created by setting each pixel ("picture element") randomly to either black or white. Then, a copy of this image is made. Call this copy the right image. The right image is then altered by taking a square patch and shifting that patch horizontally by d pixels to the left, writing over any pixels values. The pixels vacated by shifting the patch are filled in with random values. This procedure yields four types of regions in the two images.

- the shifted pixels (visible in both left and right images)

[18] B. Julesz, "Binocular depth perception without familiarity cues", Science, 145:356-362 (1964)

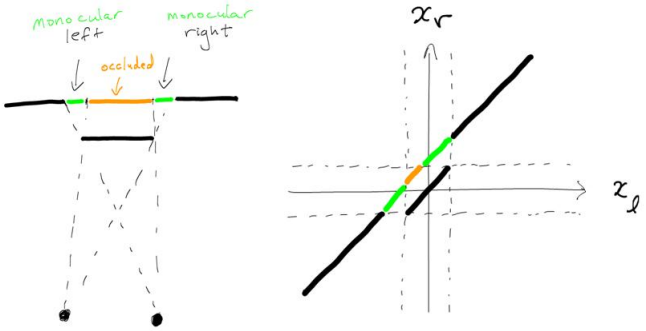
- the pixels in the left image that were erased from the right image, because of the shift and write; (left only)
- the pixels in the right image that were vacated by the shift (right only)
- any other pixels in the two images (both left and right)

To view a stereogram such as shown above, your left eye should look at the left image and your right eye should look at the right image. (This is difficult to do without training.) If you do it correctly, then you will see a square floating in front of a background.

Disparity space

Let's relate the above example to a 3D scene geometry that could give rise to it. The scene contains two depths: the depth of the square and the depth of the background. Suppose the eyes are verging on the square. We approximate the disparity as 0 on the whole square, and then the background has negative disparity.

Let's consider a 'disparity space' representation of the scene. Specifically consider a single horizontal line $y = y_0$ in the image which cuts across the displaced square. We wish to understand the disparities along this line. The figure below represents this line in the two images using the disparity space coordinate system (x_l, x_r). For each 3D scene point that projects to this line $y = y0$, there is a unique x_l and x_r coordinate, regardless of whether the point is visible in the image. (It may be hidden behind another surface.) Moreover, each depth value Z corresponds to a unique disparity value, since $d = x_l - x_r = T_x/Z$.



Notice that the set of lines that arrive at the left eye are vertical lines in the figure on the right, and the set of lines that arrive at the right eye are horizontal lines in the figure on the right. Similarly, each horizontal line in the figure on the left represents a line of constant depth (constant disparity). Each diagonal line in the figure on the right represents a line of constant disparity (constant depth).

In the sketch, we have assumed that the eyes are verging at a point on the foreground square. The background square has $x_l < x_r$ and so disparity d is negative.

Because of the geometry of the projection, certain points on the background surface are visible to one eye only; others are visible to both eyes; still others are visible to neither eye. Points that are visible to one eye only are called *monocular* points. In the exercises and assignment, you will explore this a bit further.

 ©Michael Langer

Last lecture when I discussed defocus blur and disparities, I said very little about neural computation. Instead I discussed how blur and disparity are related to each other and to depth – in particular, how blur and disparity vary with accommodation and vergence.

Today I will discuss other sources of image information – called *cues* – namely perspective, texture, and shading. I will briefly describe the information available to the visual system and what problems the visual system is solving when estimating scene depth using these cues. One general idea is that we've consider so far only *depth* of isolated points or small patches around some (X, Y, Z). But our perception of the visual world doesn't represent the world as a set of points or small patches. Rather we group patches together into large surfaces. We don't just perceive depths of points and patches, but rather we perceive the *layout* of scenes – the slants of large surfaces such as a ground plane or wall. We also perceive the 3D *shap*e of objects and whether certain parts of objects are concave or convex and how these local parts of an object fit together.

Perspective and vanishing points

You are all familar with the fact that parallel lines in the world might not appear parallel; when you look at them. The two rails of a train track or the two sides of a road will meet at the horizon, for example. Note that the lines don't actually meet; they only meet "in the limit". The image point where such parallel lines meet is called a vanishing point.

To say that parallel lines meet at infinity just means that if we take two parallel lines and consider where they strike some constant $Z = Z_0$ plane, then the XY distance between the points where the lines intersect a $Z = Z_0$ plane will *not* depend on depth (since the lines are parallel). However, when you project the two points at depth Z_0 into the image, the xy image distance between them will fall as $\frac{1}{Z_0}$.

Although we are most familiar with vanishing points that are defined by lines that lie in a plane such as the above examples, a vanishing points in fact is defined by any set of parallel lines. Consider a hallway, for example. The floor and ceiling and tops of the door frames will all be parallel lines but these lines lie in multiple depth planes. Similarly, the vertical lines in the door frames on the two sides of the hallway lie in different planes.

Images of manmade environments typically have three vanishing points, which correpond to vertical (gravity) and the natural axes of the two "floor plan" dimensions of the buildings or rooms in the environment. If one or two of these axes is perpendicular to the camera/eye axis, then the lines in these directions will be parallel in the image and will meet only at infinity in the image. In the slides I give an example of the McConnell Engineering Build at McGill and I indicate three sets of parallel lines in three consecutive slides. The third set of lines is parallel to the scene gravity axis which is roughly perpendicular to the camera Z axis and so its vanishing point is well outside the image frame.

Vanishing points provide strong cues about 3D. But what information do they convey? Vanishing points allow us to assign a Z component (or extent) to image lines and edges. Detecting a line or edge in an image only identifies (x, y) coordinates in visual direction, but it doesn't tell us about the Z component. By associating an line or edge with a vanishing point, we can identify a slope of that line or edge in depth. I'll have more to say about depth slope soon.

What is the computational problem that the visual system needs to solve when identifying vanishing points? Vanishing points aren't given, but rather they must be found. If a scene contains multiple sets of parallel 3D lines, then the visual system must form groups of these lines and

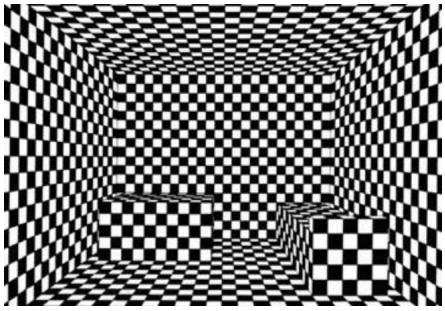
edges, corresponding to different vanishing points. There is a chicken-and-egg problem here. The visual system cannot decide whether a line or edge corresponds to a vanishing point without first knowing where the vanishing point is. And it cannot decide where a vanishing point is, unless it has identified a set of lines or edges that correspond to a common (unknown) vanishing point. Many computational models have been proposed for solving this chicken and egg problem – most of these in computer vision.

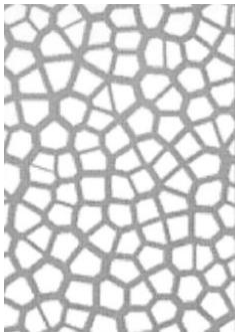
Shape from texture

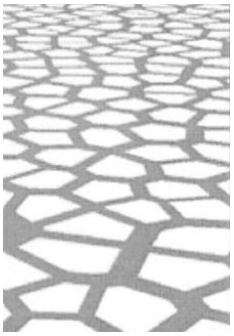
Many surfaces in the world are covered in small surface facets that have a range of sizes. Examples are grass or leaves on the ground, stones, bricks. Sometimes these texture elements are arranged in a regular pattern, for example, floor tiles.

An extreme example is the (computer graphics generated) checkerboard room shown below on the left. Such regular texture patterns on surfaces can convey rich 3D information. These scenes contain parallel lines and so they have vanishing points, but there is more information than that since the lines are *regularly spaced* in 3D and so the distances between the lines in the image varies gradually and systematically with depth. The visual system can potentially relate such gradients in image structure to 3D depth gradients.

Such gradients are defined for random textures as well, such as the examples on the middle and right. The randomness of the sizes and positions of the texture elements and the lack of vanishing points reduces the amount of information about 3D geometry that is available. Yet for those two examples, we get an impression of how depth varies across the image. In the middle panel, the surface appears to be frontoparallel (constant depth) whereas in the right panel the surface appears to slope backwards like a ground plane.







Below see a few photographs of real textures e.g. coins randomly distributed on a plane, and leaves on the ground. In each case, you have a sense of the slope of the plane. In the case of the coins which are all disks, you can use the compression of the coins to tell you something about the slope of the plane. The case with the leaves is more complicated since the leaves have a wide range of sizes.

 ©Michael Langer



Slant and tilt

We are considering the problem of perceiving the slope of a ground plane. The slope downward (a ceiling) or upward (a floor) or it may be to the left or right or some intermediate. Consider a general *oblique* plane which depth map

$$Z = Z_0 + AX + BY$$

where XYZ is the viewer's coordinate system which we have been using throughout the course. Note that such a plane always intersects the Z axis, namely at Z_0. (The ground plane $Y = -h$ does not satisfy this property if we are looking at the horizon.)

The *depth gradient* is the direction in 3D in which the depth of the plane is increasing fastest:

$$\nabla Z \equiv (\frac{\partial Z}{\partial X}, \frac{\partial Z}{\partial Y}) = (A, B).$$

The magnitude of the gradient

$$|\nabla Z| = \sqrt{(\frac{\partial Z}{\partial X})^2 + (\frac{\partial Z}{\partial Y})^2}$$

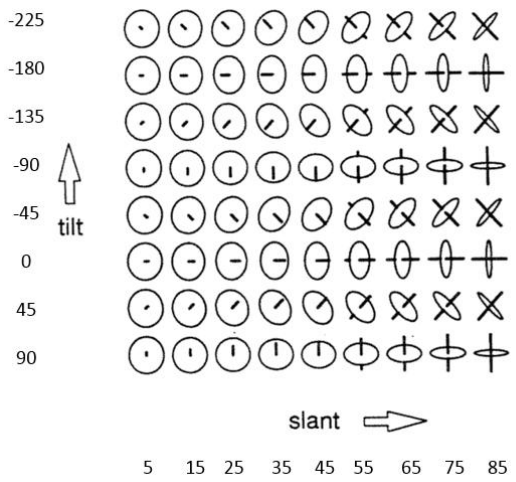
is the tangent, i.e. tan(), of the angle σ between the plane $Z = Z_0 + AX + BY$ and the constant depth $Z = Z_o$. This angle σ is called the *slant*, i.e.

$$|\nabla Z| \ = \tan \sigma.$$

We also define the *direction* of the depth gradient, which is the angle τ such that

$$\nabla Z \ = \ |\nabla Z| \ (\cos \tau, \sin \tau).$$

The angle τ is called the *tilt*. It is the angle from the viewer's X axis to the depth gradient vector $(\frac{\partial Z}{\partial X}, \frac{\partial Z}{\partial Y})$. Tilt τ is only defined when $|\nabla Z| > 0$ since when the plane is frontoparallel (constant Z) we cannot say in which direction it is sloped as it isn't sloped in any direction! The figure below (from Koenderink 1992) shows examples of slant and tilt.

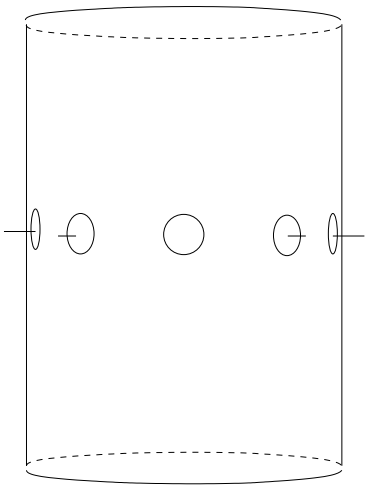


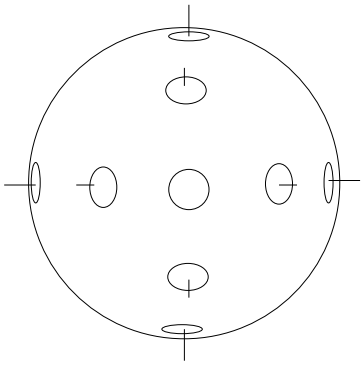
Curved surfaces

Slant and tilt are very commonly used in surface shape perception, and they seem to capture different qualitative aspects of surface orienation in space – i.e. how sloped versus which direction is the slope ? Their usage goes beyond the case of a single slanted plane though. To give you some intuition, here is an illustration of slant and tilt of a cylinder and sphere. We can also talk about the slant and tilt of points on a curved surface. Slant is the angle by which the *local tangent plane* is rotated away from a front facing disk. Tilt is the direction of this rotation.

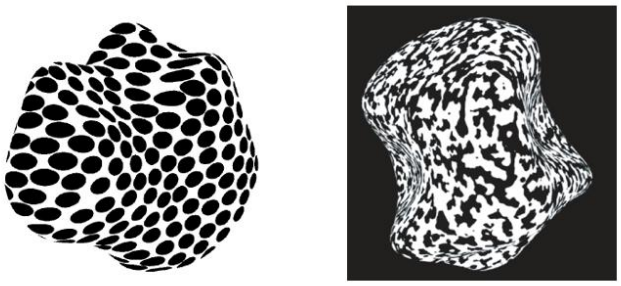
In the cylinder example below, the tilt is 0 deg for the two on the right and 180 deg for the two on the left. The tilt is not defined at the center one since the surface has zero slant there. The slants are about 80, 45, 0, 45, 80 going from left to right.

For the sphere example, the slants are close to 90 deg for four examples near the boundary. (Slants are 90 degrees always at a depth boundary of a smooth surface!) The tilts go from 0, 90, 180, 270 counter-clockwise.





Consider two surfaces rendered below[19] which are smooth random blobby shapes. The texture on the surface tells us something about the surface 3D shape. On the left, the texture elements are elongated disks. On the right, the texture elements are random blotches of various sizes. The surface orientation causes these texture elements to be compressed *in the tilt direction, and by an amount that depends on the slant*. The visual system uses this compression information to perceive shape, and this has been shown in various studies. But we don't know how this processing is done. One limitation is that the visual system does not know what the texture *on the surface* would look like if it were somehow placed on a frontoparallel plane so that there were no view-based distortion. So, the visual system cannot be sure how much of the compression that is observed *in the image* is due to the image projection, and how much of the compression is due to compression that might be present in the original texture itself. For example, on the left, the disk-like texture elements on the surface are not perfectly round disks, but rather they are already elongated i.e. *prior to projection*. This seems to be the case on the left side of the left figure, for example, as the disk texture elements there appear to be horizontally elongated on the surface (as well as in the image).



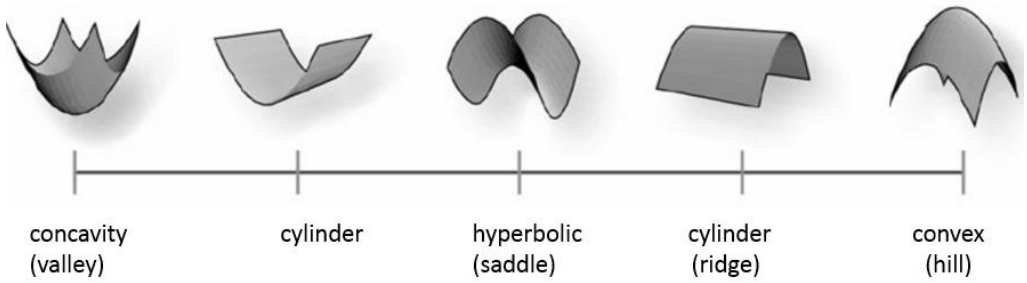
We have been discussing surface orientation (depth gradient, slant, tilt), but these properties seem inadequate for capturing our intuitions about surface shape, in particular, curvature. One can define surface curvature formally in terms of the second derivatives of the surface, and this is part of the topic of *differential geometry* which some of you who have spent more time in the math department may be familiar with. However, differential geometry isn't intended to capture what we perceive about shape and so one can ask if there is a way of (mathematically) defining local shape properties which *does* correspond to our intuitions.

One nice example of how this can be done[20] is illustrated in the figure below. The examples are local surface regions that are (from left to right) concave, an extended valley, a saddle, an extended ridge, or a convex region. We can define a continuum of shapes between the identified ones. This continuum of shapes can be defined mathematically by varying one parameter (which Koenderink and van Doorn (1992) call the *shape index*).

One parameter doesn't define the surface uniquely though, since a surface can curve in two directions at each point. The second parameter has to do with the amount of curvature; think of the scale of the surface in 3D. For the concavity case on the right, compare a golf ball to the planet earth. Both are spheres and have the same shape, but the curvature amounts are quite different.

[19] papers by Jim Todd, Roland Fleming, and colleagues

[20] Koenderink and van Doorn 1992



The slides show an example of a 3d model of a face which uses a color map to indicate the type of local shape (shape index) and the amount of curvature ("curvedness"). For example, there are just two convex spherical regions on the surface: the top of the head and the tip of the nose. Note that the curvedness is quite different in these two cases.

Surface tangent plane and unit surface normal

To define slant and tilt at different points on a general curved surface, we consider the tangent plane at each visible point. As we saw above, the tangent plane will change from point to point along the surface. For any visible point (X_p, Y_p, Z_p) on the surface, the surface in the neighborhood of that point can be approximated as a planar depth map (the *tangent plane*):

$$Z(X_p + \Delta X, Y_p + \Delta Y) = Z_p + \frac{\partial Z}{\partial X} \cdot \Delta X + \frac{\partial Z}{\partial Y} \cdot \Delta Y$$

where $(X, Y) = (X_p + \Delta X, Y_p + \Delta Y)$. Note that this is different from the equation of a plane $Z = Z_0 + AX + BY$ mentioned earlier, since (X_p, Y_p) is not necessarily $(0, 0)$. [ASIDE: I did not make this distinction originally in the slides, but I have now changed them.]

It is often useful to talk about the unit vector that is perpendicular to the local tangent plane. This is called the *local surface normal*. Let's derive what this vector is. Not surprisingly, it depends on the depth gradient of the tangent plane.

Taking Z_p from the right side to the left side of the above equation, we get

$$\Delta Z = \frac{\partial Z}{\partial X}\Delta X + \frac{\partial Z}{\partial Y}\Delta Y$$

or

$$(\Delta X, \Delta Y, \Delta Z) \cdot (\frac{\partial Z}{\partial X}, \frac{\partial Z}{\partial Y}, -1) = 0 \ .$$

Since this inner product relationship holds for any step $(\Delta X, \Delta Y, \Delta Z)$ in the tangent plane of the surface, it follows that the vector $(\frac{\partial Z}{\partial X}, \frac{\partial Z}{\partial Y}, -1)$ is perpendicular to the surface and hence this vector is in the direction of the *surface normal*. We rescale this vector to unit length, and call it as the *unit normal vector*

$$\mathbf{N} \equiv \frac{1}{\sqrt{(\frac{\partial Z}{\partial X})^2 + (\frac{\partial Z}{\partial Y})^2 + 1}} (\frac{\partial Z}{\partial X}, \frac{\partial Z}{\partial Y}, -1) \ .$$

Notice that we are not considering steps $(\Delta x, \Delta y)$ in the image, but rather we are considering steps on the surface, or more specifically, on the surface tangent plane.

Shape from shading

Suppose we have a surface defined by a depth map $Z(X, Y)$. We have in mind here a curved surface. If we illuminate this surface by a parallel source such as sunlight from some 3D direction $\mathbf{L}$, then the amount of light reaching a surface patch depends on the orientation of the patch. If the patch faces the light source directly, then the maximum amount of light reaches that patch. As the patch orienation is rotated away from the light source direction, the amount of light that reaches the patch falls off. Lambert's law says that the amount of light reaching the patch depends on the cosine of the angle between the normal of the patch and the direction to the light source:

$$I(X, Y) = \mathbf{N}(X, Y) \cdot \mathbf{L}.$$

Here I have normalized the intensities so that the point that faces directly to the source has value 1. Also note that there is no dependence in this model on the distance from the light source to the surface. Essentially we are assuming that the light source is very far away, like sunlight.

The computational problem of *shape from shading* then goes as follows: given an intensity image, estimate a surface $Z(X, Y)$ such that the surface normal satisfies the above model. There are a number of reasons why this problem is difficult. First, the vision system needs to estimate the light source direction. People have come up with methods for doing so, which I won't go into. Let's just assume for now some hypothetical direction $\mathbf{L}$. The second reason the problem is still difficult is that the surface normal has two degrees of freedom at each point, namely it is some direction on a unit hemisphere facing the light source. The intensity $I(X, Y)$ only specifies the angle between the normal $\mathbf{N}$ and the light source, but there are many possible $\mathbf{N}$ that have this angle.

There are other versions of the shape from shading problem, and I'll return to them next lecture. For now, let's just think about what problem we are solving here. I said that the problem was to estimate the surface normal at each point. You can imagine that, with the estimates of surface normals, you can estimate the surface depths $Z(X, Y)$ by fitting oriented patches together. You could also estimate the surface shape (see shape index on previous page) by piecing patches together. But notice that this discussion is pure handwaving. It does not say *how* to estimate the surface normals. Frankly there are far more unknown aspects of this problem than known aspects – and this is despite decades of experiments and theorizing. (This is in remarkably sharp contrast to the situation of binocular stereo, which relatively well understood.)

 ©Michael Langer

Shading and shape (continued from lecture 11)

Last lecture we examined the problems of perceiving surface shape from texture and shading. The discussion was not at the level of neural coding, but rather it was at the level of what problem was to be solved. What are the 3D scene properties that we mean when we say "shape' (e.g. depth, depth gradient – slant and tilt – and curvature). How are these properties related to image intensities?

We begin today by considering a few variations on the shape from shading problem. We assume that the physical intensity of light reflected from an image depends on the angle between the surface normal and some light source direction $\mathbf{L}$ which we assume to be constant i.e. the source is far away like the sun:

$$I(X,Y) = \mathbf{N}(X,Y) \cdot \mathbf{L}$$

where

$$\mathbf{N} \equiv \frac{1}{\sqrt{(\frac{\partial Z}{\partial X})^2 + (\frac{\partial Z}{\partial Y})^2 + 1}} (\frac{\partial Z}{\partial X}, \frac{\partial Z}{\partial Y}, -1) \ .$$

Note that intensity $I(X,Y)$ is being defined in XY variables of 3D scene coordinates XYZ – where $Z(X,Y)$ is a depth map – rather than as $I(x,y)$. Similarly the depth map is defined on (X,Y) rather than on (x,y). We do so because it is simpler to define scene planes (tangent planes and surface normals) on (X,Y).

The above model holds only when $\mathbf{N} \cdot \mathbf{L} \ \geq 0$, since it is meaningless to have negative intensities. It can happen that the inner product of $\mathbf{N}$ and $\mathbf{L}$ is less than zero, and in this case the surface is facing away from the light source and would not be illuminated by the source. The illuminance from the source would be zero (not negative). We could consider this case in the model by writing $I(X,Y) = max(\mathbf{N}(X,Y) \cdot \mathbf{L}, 0)$.

We refer to the situation above in which $\mathbf{N}(X,Y) \cdot \mathbf{L} < 0$ as an *attached shadow*. In this situation, the surface received no direct illumination from the source. This is distinguished from a *cast shadow* where $\mathbf{N}(X,Y) \cdot \mathbf{L} > 0$ but the light source is not visible because it is occluded by some other object.

Linear shape–from–shading

One variation of the above shading model occurs when the surface is nearly planar (Z is approximately constant) and has low relief bumps and dents on it, and is illuminated from a direction that is oblique to the surface normal. By "low relief", specifically we mean that the partial derivatives of Z with respect to X and Y are small i.e. $\frac{\partial Z}{\partial X} \approx 0$ and $\frac{\partial Z}{\partial Y} \approx 0$ and so

$$\frac{1}{\sqrt{(\frac{\partial Z}{\partial X})^2 + (\frac{\partial Z}{\partial Y})^2 + 1}} \approx 1$$

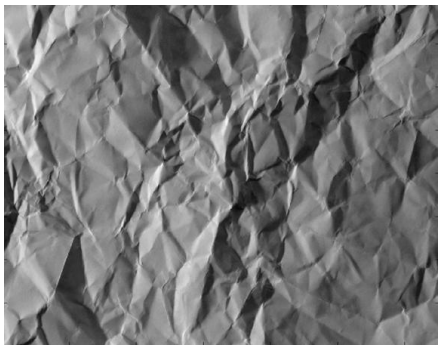
In this case, we obtain an approximation:

$$I(X,Y) \ \approx \ (\frac{\partial Z}{\partial X}, \frac{\partial Z}{\partial Y}, -1) \cdot (L_X, L_Y, L_Z).$$

An example is shown below of uncrumpled paper illuminated by a light source that is off to the side. The surface has a slant near 0, so that the Z variable in the linear equation above corresponds to the Z axis of the viewer.

 ©Michael Langer

The image looks like a real surface, not just some random patterns of grey level intensities. Moreover, you perceive it to be a relatively flat surface. You also may perceive the light direction to come from the left. How sure are you about that? And why do you not think the illumination is coming from the right? We will return to these questions a few lectures from now.



To understand the shading effects better for this model, let's consider a simple example of a surface depth map:

$$Z(X,Y) = Z_0 + a\sin(k_0 X).$$

An example would be hanging drapery (curtains). The frequency is some constant k_0 which is the number of radians per length in X variable. (We could put in a factor of 2π to make the units of k_0 number of cycles per unit distance.)

What is the linear shading model for this example? Computing partial derivatives

$$\frac{\partial Z(X,Y)}{\partial X} = ak_0\cos(k_0 X), \qquad \frac{\partial Z(X,Y)}{\partial Y} = 0$$

and plugging into the shading model gives:

$$I(X,Y) = a\ k_0\ L_X\cos(k_0 X) - L_Z.$$

Notice that the intensity is 90 degrees out of phase with the depth map, i.e. sine versus cosine. So the maximum and minima of intensity don't occur on top of the depth hills and valleys. Rather, they occur on the sides of the slopes. Also notice that $L_Z < 0$ for this model to make sense, since we need to have positive intensities and the cosine oscillates between positive and negative values. Also notice that $a\ k_0\ L_X$ cannot be too large, otherwise the intensity will become negative when the cosine is negative. This creates an attached shadow effect. For this particular surface, whenever there is an attached shadow there is also a cast shadow, as was observed in class.

Shape from shading on a cloudy day

Another shading model[21] addresses quite a different lighting condition, namely a high relief surface under diffuse lighting such as on a cloudy day. The sunny day model cannot capture this shading because there is not a single light source direction $\mathbf{L}$. Rather on a cloudy day there are many light source directions. Indeeed there is a hemisphere (sky) of directions.

[21] introduced by yours truly in my Ph.D. thesis

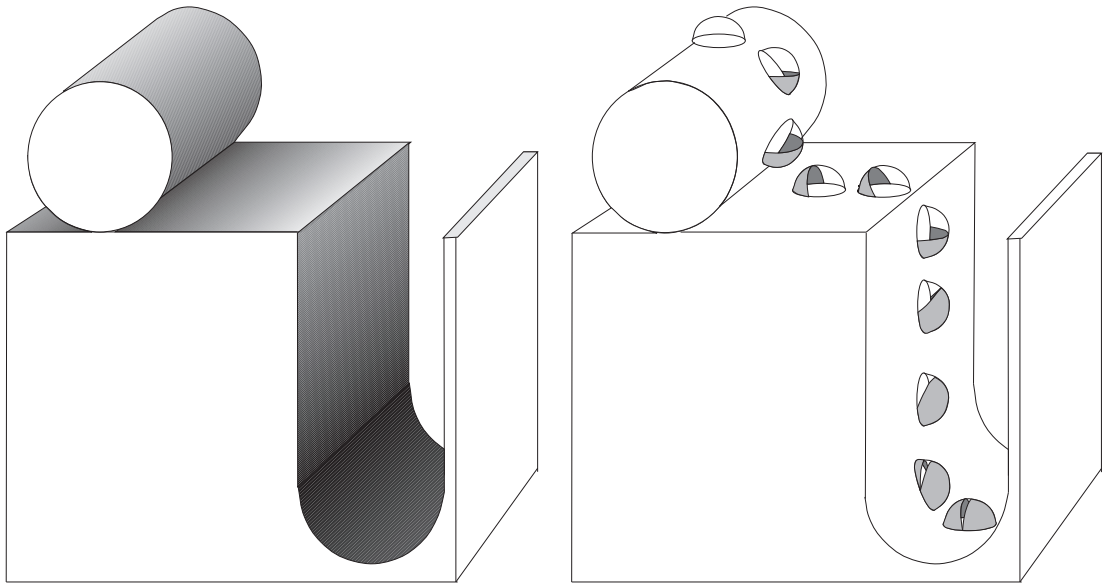
 ©Michael Langer

Although light comes from all directions on a cloudy day, the surface is not uniformly illuminated. The reason is that not all of the sky is visible from every point on the surface, and this varying sky visibility is a shadowing effect. We can integrate the previous model over directions $\mathcal{V}(\mathcal{X}, \mathcal{Y}, \mathcal{Z})$ in which the sky is visible:

$$I(X,Y) \;=\; \int_{\mathcal{V}(\mathcal{X},\mathcal{Y},\mathcal{Z})} \mathbf{N} \cdot \mathbf{L} \; d\mathbf{L}$$

This implicitly assumes the sky is equal intensity in all directions (which isn't true, but we'll assume it for simplicity).

The above model is much more complicated than the sunny day model because now both the surface normal and the region of the visible sky vary along the surface. The graphic on the right shows the amount of the hemispheric sky that is visible for different points on the surface. At the top of the cylinder, the entire sky is visible and this is the brightest point in the scene.[22] As we go around the cylinder towards the bottom, the amount of sky that is visible decreases. Similarly as we go down into the valley the amount of visible sky decreases. Note that the amount of visible sky can change along the surface because of cast shadow effects.



One can simplify the model by ignoring the $\mathbf{N} \cdot \mathbf{L}$ term and just considering:

$$I(X,Y) \;=\; \int_{\mathcal{V}(\mathcal{X},\mathcal{Y},\mathcal{Z})} d\mathbf{L}$$

which says that the surface intensity at each point is proportional to the fraction of the sky that is visible. FYI, I was able to come up with a computer vision algorithm for computing a depth map $Z(X,Y)$ that is consistent with this model, given an image $I(X,Y)$.

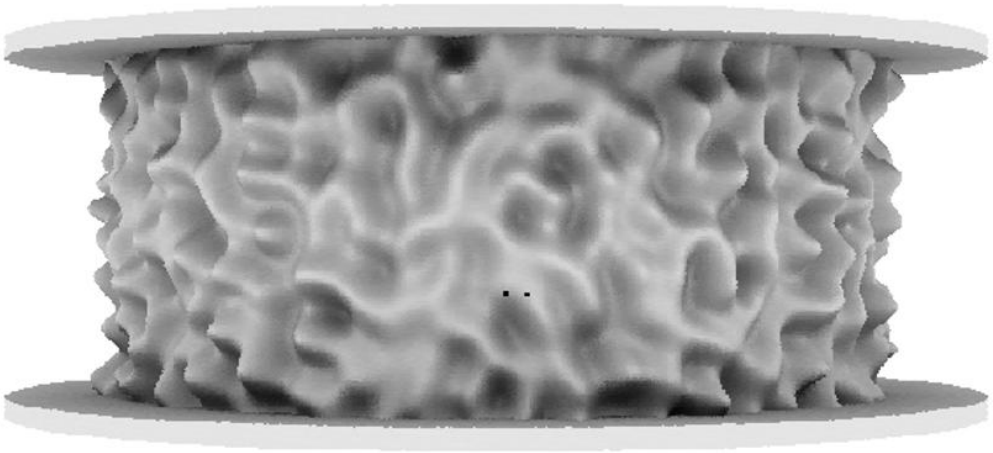
A few years after my PhD, I carried out perception experiments that studied how people perceive shape from shading under diffuse lighting. These experiments used rendered images such as the one

[22] The shading doesn't indicate this property. Evidently I wasn't careful enough when I made this figure many years ago.

 ©Michael Langer

shown below. Notice the little local intensity maxima in the valleys of this rendered surface. These little maxima are due to the surface normal in the valley turning to face directly towards the part of the sky that is vaible (rather than facing a side wall of the valley). For these visible parts of the sky and for points at the bottom of valleys, $\mathbf{N} \cdot \mathbf{L}$ tends to be close to 1. This leads to a local peak in intensities in the bottom of the valleys, which you can easily see in the rendered images (which used fancy computer graphics that took account of sky visiblity and surface normal effects).

In my experiments, I wanted to know if people would be fooled by these local intensity maxima. Their task was to judge the relative depths of pairs of points (such as the little black squares in the image). Sometimes the darker point was deeper, but sometimes the darker point was shallower such as in the case of a point on the side wall versus a point at the bottom of the valley. I found that in many cases people correctly identified that the brighter point was deeper. It was as if people correctly attributed the local intensity maxima to the surface normal effect rather than attributing it to a small hill within a bigger valley.



All the above shading models assume that the surface has a constant reflectance, and that all intensity variations are due to changes in the amount of illumination. But surfaces can have reflectance variations too. The rest of this lecture will examine situations in which the reflectance (and illumination) change.

Lightness versus Brightness

[See the slides to accompany illustrations for the text below.]

I began with an example photograph showing two pieces of white paper laying on a carpet. One paper was in shadow and one paper was not. The paper in shadow naturally receives less illumination from the light source and so its images has lower intensities. Although it is physically darker and appears less bright too (this distinction will be discussed below), it still seems to be a white piece of paper, as if the visual system takes account of the shadowing effect.

A second photograph replaces the intensities on the illuminated paper with intensities that are equal those of the shadowed paper. Remarkably, now the illuminated paper appears to be a darker

 ©Michael Langer

color paper than the shadowed one. Again, the visual system is taking account of the lighting effect (or discounting the effect of the illumination and shadow). The paper on the right *appears* darker now as if this is the best way for the visual system to explain how the two papers have the same physical intensity. To see this, consider:

$$I(x,y) \;\equiv\; illuminance(x,y) \;\times\; reflectance(x,y)$$

and note that if $I(x_1,y_1) = I(x_2,y_2)$ and if the shadows suggest that

$$illuminance(x_1,y_1) < illuminance(x_2,y_2)$$

then it follows that

$$reflectance(x_1,y_1) > reflectance(x_2,y_2).$$

Is this the correct way of the thinking about what the visual system is doing? Indeed some vision scientists shun these sorts of explainations, and prefer to explain everything in terms of neural coding. But notice that this example is just another version of the simultaneous contrast effect which you saw back in Assignment 1. Perhaps you can explain this effect in terms of neural coding (and lateral inhibition). However, as you saw in Assignment 1 with White's effect, sometimes the simple models also predict the wrong thing.

For today, let's not wring our hands over this issue. Instead let's just try to understand the computational problem that is being solved. The problem is to discount (or at least partially discount) the effects of illumination. The idea is that it isn't as useful for the visual system to estimate the exact magnitude of the intensity at each point in an image. Rather it is more useful to know the reflectance of the surfaces.

As the relation above says, we can think of an image $I(x,y)$ as consisting of the product of two *intrinsic* images: the $illuminance(x,y)$ which captures the shading and shadows, and the $reflectance(x,y)$ which is the fraction of light arriving at the surface that gets reflected. It is straightforward to model the physics of light reflecting off a surface, such as the models above. But how to model the perception of such situations?

First, we need to distinguish between physical and perceptual quantities. The term *luminance* refers to the physical intensity of the light reflected from a surface, whereas the term *brightness* refers to the *perceived* intensity. The two are (hopefully obviously) not the same thing – not just because physical quantities are differnet from perceptual quantities, but also because the light in one image patch might be physically more intense than the light in another patch, and yet the first might be perceived as less intense (less bright).

Second, people also sometimes are capable of judging the reflectance of surfaces. I emphasize that reflectance refers to the fraction of light that gets reflected from a surface, and it is physical quantity. One uses the term *lightness* to describe the *perceived reflectance*. When you look at a surface and judge its colour (grey vs. black vs white, etc), you are making a lightness judgment – not a brightness judgment.

Distinguishing lightness judgments from brightness judgments is difficult. If you run an experiment and you ask people off the street to make different judgments, they typically don't know what you are asking. Even people who are worked in the field sometimes get confused. This is especially a problem when we are looking at pictures, rather than physical objects in a real 3D scene. The following example is susceptible to this problem, but I'll describe it anyhow because it is so nice in other ways.

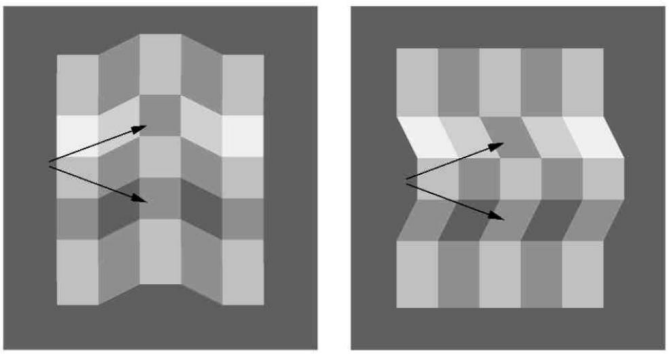
 ©Michael Langer

Adelson's corrugated plain illusion below shows a random checkerboard pattern, which has been folded. The folding is either along vertical lines or horizontal lines, and the two images each naturally contain five groups of five tiles. (Both images are consistent with either a concave or convex folding, but let's not deal with that now.)

Consider the four square tiles that the arrows point to. In fact, they all have the same shade of grey – same physical intensity. In the example on the left, the two tiles that are pointed to appear in the same vertical group of five tiles which all lie in a common plane. The two tiles appear to be the same, whether we are judging brightness or lightness. Indeed it is difficult to say whether our percepts are brightness or lightness since we do not have strong cues about illumination.

In the example on the right, the two tiles that are pointed two now appear to have different shade of grey: the tile on the top appears darker. The most basic explanation for this is that the tile on the top is grouped with the four other tiles in the same row (respecting the 3D interpretation of the folding). The upper tile is the darkest tile in its row. The lower tile belongs to a row that has only two intensities, and the lower tile is in the more intense group.

To explain why the upper tile looks darker (in this example, we don't distinguish brightness from lightness), we suppose that the visual system only compares a tile with others in the same 5-tuple which appear to lie on a common plane and hence have roughly the same illumination (since there is no evidence of illumination changing, like a shadow). If a tile is the brightest in its group, it is perceived as closer to white, whereas if it is the darkest then it is perceived closer to black. That's it, and that idea of comparing within groups takes you a long way – as many other examples show.



Color constancy

When we discussed the basics of color earlier in the course, we used the terms hue and saturation. These properties of color (along with lightness) are often used to recognize objects. In particular, we are very sensitive to the color of skin, when judging the emotional states of others (embarassment, anger, or whether someone has had a bad night). We also judge the ripeness and edibility of food: think meat, bananas, oranges, etc. But our discussion of color earlier in the course did not distinguish between the illumination and the surface reflectance. So let's address this now.

Recall the relationship from lecture 3:

$$I_{LMS} = \int C_{LMS}(\lambda)\ E(\lambda) d\lambda$$

which describes the linear response of a photoreceptor as the sum over all wavelengths of the product of the absorption and the spectrum of the light that arrives at that point on the retina. For a color image, we need to add a pixel position dependence:

$$I_{LMS}(x,y) = \int C_{LMS}(\lambda)\ E(x,y\lambda)d\lambda$$

Notice that the cone absorption C doesn't depend on position. We are assuming that LMS cones have the same properties at all positions.

We also need to say more about spectrum $E(x,y,\lambda)$ which arrives at the retina. In particular we are interested in cases the light is reflected from a surface. Just as in the grey level case of brightness and lightness, we would like to know how a vision system can discount the illuminant color.

The spectrum of light $I(x,y,\lambda)$ arriving at a point in the image depends on the spectrum of the light source (as a function of wavelength) and the percentage of light that is reflected by a surface at each wavelength. Here are few details (not mentioned in class):

- *illuminance*: Each light source has a characteristic spectrum. Sources that emit light because they are hot have a spectrum that depends on their temperature. A fire, tungsten light bulb, the sun all have quite different spectrum. The sun's spectrum is relatively flat over the range of visible light, whereas a tungsten light bulb has much more energy at long wavelengths than short wavelengths. The spectra of sources such as fluorescent light and CRT phosphors are must more spiky as a function of λ than are natural source spectra (which are relatively smooth)

- *surfaces reflectance*: when light reflects off pigmented surfaces – paints and dyes – certain wavelengths are reflected more than others. Foliage is green because it contains a pigment chlorophyll. (In the Fall, because of changes in temperature and other factors, the chlorophyll pigment breaks down. This why leaves change color and lose their green.)

Suppose light is emitted by a source and has a certain amount of energy per wavelength. Call this spectra the $illuminance(\lambda)$. Suppose this source light is then reflected from a surface. For each wavelength, a proportion of the incident light is reflected and this proportion is $reflectance(\lambda)$. For example, objects that appear red reflect long wavelength light (> 600 nm) better than short wavelength light (< 500 nm). The spectrum of reflected light is the wavelength by wavelength product,

$$I(\lambda)\ \equiv\ illuminance(\lambda)\ \times reflectance(\lambda)$$

Since these values can vary along the surfaces and across the image, they depend on image position (x,y), so we write

$$I(x,y,\lambda)\ \equiv\ illuminance(x,y,\lambda)\ \times\ reflectance(x,y,\lambda).$$

This is similar to what we saw above in the black and white domain, but now we have put wavelength into the equation. (Note that the illuminance implicitly includes shading effects too.)

The perceptual problem now is to take the photoreceptor intensities $I_L(x,y), I_M(x,y), I_S(x,y)$ and to infer as much as possible about the reflectance spectra of surfaces in the scene and the illuminant spectrum.

 ©Michael Langer

You may think this problem is hopelessly impossible. However, this pessimism ignores the facts of our everyday experience. We are able to judge the colors *of object surfaces* quite well, and discount the illuminant to some extent. This observation holds both informally (our day-to-day experience) and when you go into the lab and do careful experiments with people, asking them to judge color of surfaces as you vary the illumination. *People do make mistakes* (some of them systematic), but the mistakes are surprising small. The ability to see colors such that they appear roughly the same under different illumination is called *color constancy.*

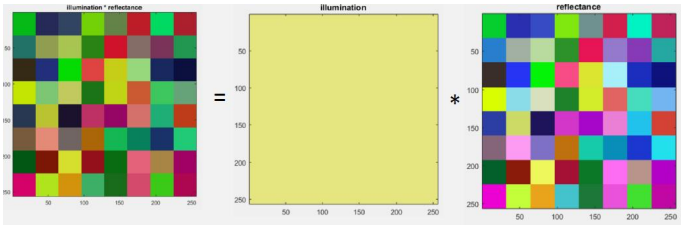
Let's sketch out a few basic ideas for how this can be done. First, if we suppose that the cone response curves don't overlap, then we can think of three ranges of wavelengths. This let's us treat the three channels (LMS or RGB) as independent. (I'll write RGB instead of LMS to be consistent with the slides.) At each point (x, y), we can think of having three intensity values $I_{RGB}(x, y)$ and three illuminance values $illuminance_{RGB}(x, y)$ and three reflectance values $reflectance_{RGB}(x, y)$, so we can write:

$$I_{RGB}(x, y) \equiv illuminance_{RGB}(x, y) \times reflectance_{RGB}(x, y).$$

We are essentially ignoring the details *within* each of the three frequency bands. This is an approximation which let's us cut to heart of the problem, as follows.

Case 1: uniform illuminance (grey world, von Kries)

It often happens that a surface has roughly constant illuminance. As an example, consider the first row below. The checkerboard on the left has colors that are more yellowish than the surface reflectance (on the right) because the illumination is yellow.



The image on the left is literally just the one on the right, multiplied by the one in the center – point by point and channel by channel.

In the real world, the vision system's task is to take the image on the left and to discount the (yellow) illuminant. Obviously we don't do this completely when looking at the little images here; the images on left and right appear different. But this is because we are also comparing these little images to the white page that surrounds them ! In the real world, all the objects that are visible will be colored by the illuminant.

One key idea for discounting the illuminant for the vision system to assume that the surface reflectances in the scene are grey on average.[23]. Then the vision system could take the average RGB value of the scene, and if it not neutral colored (grey) then the vision system could normalize the

[23] This was suggested by a student in the class, and indeed the idea has been tested and holds some water. Its called the *grey world assumption.*

image channel-by-channel by dividing by the average intensity for each color.

$$(\frac{I_R(x,y)}{mean_{x,y}I_R(x,y)}, \frac{I_G(x,y)}{mean_{x,y}I_G(x,y)}, \frac{I_B(x,y)}{mean_{x,y}I_B(x,y)})$$

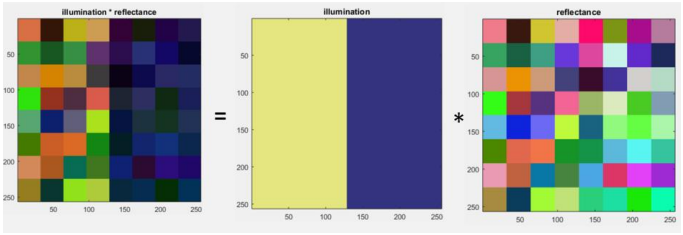
The result is that the new average over the image would be (1,1,1). Of course, you would need to scale those values down if you wanted them to represent reflectances, since the maximum reflectance is 1. For example, you could further divide all channels by a scalar which is the maximum value over all changes, which would ensure that all values are at most 1.

A second approach, which I listed on a slide but ran out of time to discuss is to normalize each image channel by the maximum value that occurs in that channel. That is, compute

$$(\frac{I_R(x,y)}{max_{x,y}I_R(x,y)}, \frac{I_G(x,y)}{max_{x,y}I_G(x,y)}, \frac{I_B(x,y)}{max_{x,y}I_B(x,y)})$$

This is not the same solution as the grey world one above, but the idea is similar: scale down brighter channels to try to discount the illumiant.

Case 2: the shadow revisited



A more challenging problem is the case of a shadow. In natural scenes that are illuminated by sunlight and blue sky, parts of the scene that are not in shadow have yellowish illumination (plus a much weaker blueish illumination from the sky) whereas shadowed regions have just blueish illumination from the sky. This situation is illustrated abstractly in the example here. How might a vision system discount the illuminant in this case?

The take home message from today is that the intensities and colors that we measure with our eyes are the product of a few different factors (literally) and that our vision systems often seem to disentangle these factors, allowing us to perceive the illuminance separately from the surface reflectance. How this is done is only partly understood.

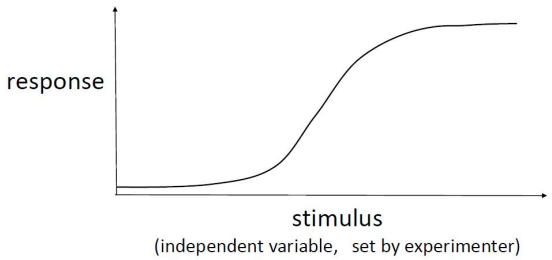
One final side note is worth mentioning: shading and shadowing primarily affect the intensity (and less so the hue and saturation). We often perceive changes in intensity as being due to shading (shape) and shadows (either attached or cast) and we often rely on geometric cues to help us figure out the 3D situation, such as the boundaries of the surfaces in the Adelson example. Once these intensity effects from shading and shadow effects have been accounted for, the visual system can more easily rely on simple normalization processes to discount the color of the illuminant.

Most of our discussion in this course up to now has been about early vision problems and the early processing in the brain to solve these problems. In the next few lectures, we will turn to how well humans solve these computational problems. These problems may include detecting a change in image intensity or color in a region, or detecting motion a depth increment from disparity, or discriminating the slope of a surface.

The term *psychophysics* refers to experimental methods that measure the mapping from some physical stimulus to a response. A person is shown some images – usually presented on a display screen – and answers questions about the images by pressing on some buttons. Psychophysics is the field of science that characterizes how responses depend on the parameters of the images. One is more interested in the underlying perceptions, and less interested in the responses themselves. But typically we can only find out about the perceptions by asking people to press buttons. (If one is doing psychophysics on monkeys, one can ask them to press buttons and one can also record from cells in their brains. Both kinds of experiments count as psychophysics.)

Psychometric function

A *psychometric function* is a mathematical function from a stimuli level (a parameterized variable) to a response level. The response can be a parameter level that is set by an observer, or it can be a statistic such as percent correct in some task. Most of our examples will consider the latter. We will typically consider S-shape (called sigmoid shaped) psychometric functions.



An example task a background patch of intensity I_0 and a central square with a different value of intensity $I + \Delta I$, where ΔI is negative or positive. The task could be to judge if there is an increment or decrement. In order to get a psychometric curve that is 'S shaped' and increasing, one would plot the percentage of times that the subject responded that the center was an increment. The response would go from 0 percent (for large decrements) to 100 percent (for large increments).

In the slides, I discussed a few other ways to set up the problem. One could have a square that is an increment only, and the task would be to say if it is in the left or right half of the display. If the ΔI is very small, then the subject will be at 50 percent correct. But as the ΔI increases, performance will rise from 50 percent to 100 percent.

Psychometric curves are typically not step functions. The reason that there is a gradual change in performance is that there are various sources of uncertainty that subjects face when doing the tasks:

- Noise in the display or stimulus (because it is a physical device)
- Random number generators in the computer program that creates the display image

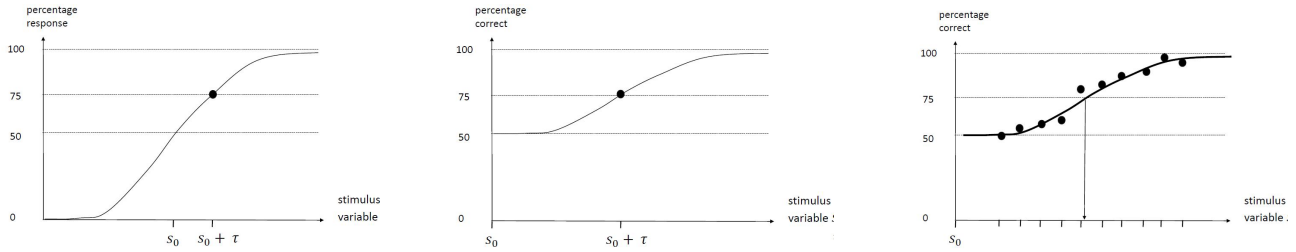
- Noise in the sensors/brain
- Limited resolution of the display or vision system e.g. finite samples in the photoreceptor grid
- Subjects press the wrong button (stop paying attention)

Different sources of uncertainty play more or less of a role in different experiments. In general, as the noise or uncertainty increases, it takes more stimulus to reach high performance. Sometimes the psychometric curves stretches out as uncertainty (noise) increases and sometimes it shifts the right, and sometimes it both stretches and shifts.

It is important to note that some of the factors that limit performance are within the observer (noise in the brain, failing to pay attention) but that some factors are inherent in the stimulus. Even a vision system that had no 'brain noise' and always paid 100 percent attention would still make mistakes since the stimulus itself could have randomness – so even an *ideal observer* would need to guess sometimes.

Psychophysical thresholds

A psychometric function has a lot of information, and often we just want to summarize it with one number. We arbitrarily take a particular performance level (for example, 75 percent correct) and consider the stimulus level that produces this performance level. This stimulus level is called a *threshold*. Such a threshold can be defined whether the responses go from 0 to 100 percent (left below) as in the case of deciding if an center square has an intensity increment or decrement with respect to the background, or in the case of a psychometric function going from 50 to 100 percent (middle below) as in the case of detecting if an increment is present (e.g. s_0 is the background intensity of square).

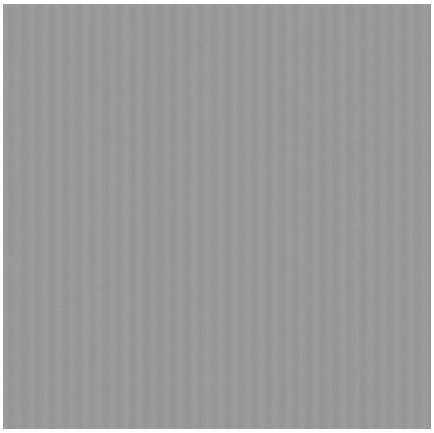


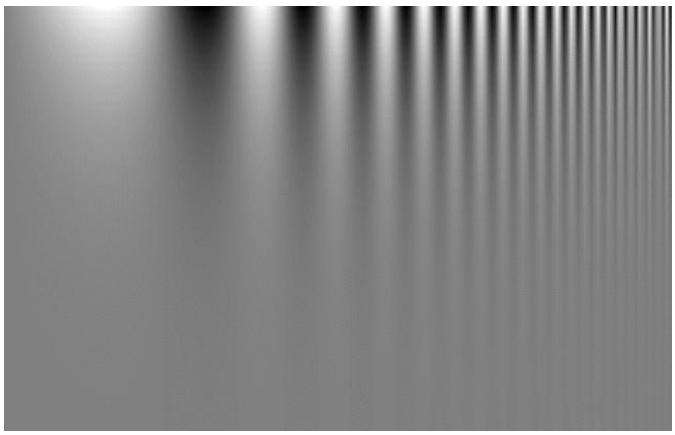
In a real experiment, one fits the parameters of some model curve to some noisy data. One then takes the 75 % threshold point *from the fitted curve* rather than from the data. Note that the fit is never perfect (above right), and often one makes strong assumptions about the shape of the curve. This is ok, as the exact threshold values are not the main point. Rather, as you will see with some examples, what is more interesting is how the values vary in different compare threshold values for different versions of the experiment – that is, across different psychometric curves. This will be more clear once you see a few examples.

Finally, one often thinks of thresholds as values above which a person can do the task and below which the person cannot do the task. (Recall for example Panum's fusion area for binocular stereo vision.) But of course that's oversimplified, since one's ability to do a task varies continuously with the amount of stimulus relative to the noise.

Michelson contrast

For several of the examples that we discuss, the stimulus is a 2D sinusoid variation. An example is a 2D intensity pattern such as below on the left. The task might be to decide if the 2D sinusoid is vertically or horizontally oriented. We would like to know how well people can perform this task as a function of the range of intensities in the pattern, and also whether performance depends on the frequency.





Define the Michelson contrast:

$$Michelson\ contrast \equiv \frac{I_{max} - I_{min}}{I_{max} + I_{min}}.$$

To understand this definition, write it slightly differently as

$$\frac{(I_{max} - I_{min})/2}{(I_{max} + I_{min})/2}$$

For the case of a 2D sinusoid function, $I(x) = I_0 + a\sin(2\pi k_x x)$, the numerator is the amplitude a of the sinusoid and the denominator is the mean I_0 of the sinusoid, so the contrast would be a/I_0.

Note that this quantity ranges from 0 to 1, where 0 means constant intensity (no contrast) and 1 means maximum contrast. In the example image above on the left, the Michelson contrast is 0.02.

Contrast detection thresholds

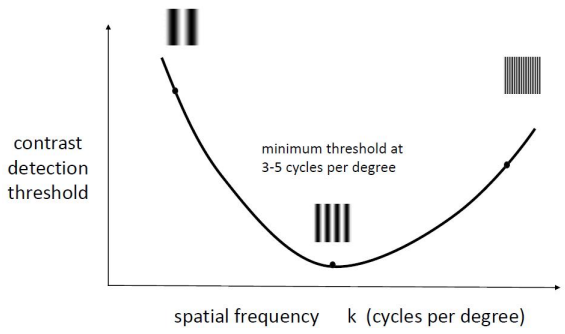
2D sinusoidal stimuli are often used in psychophysics to examine sensitivity to oriented structure and structure at different scales. Consider the example above right, which shows an image whose spatial frequency varies continuously from left to right and whose contrast increases from bottom to top. Note that the perceived boundary between grey (contrast below threshold) at the bottom and white/black alternation at the top is not a horizontal line, but rather the threshold seems to dip down and up. *The contrast threshold is lowest at the middle frequencies.*

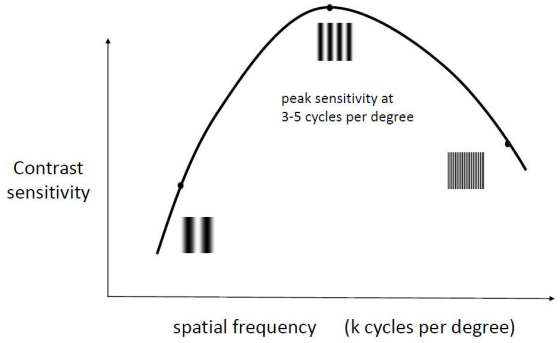
The figure just mentioned is a demo, not an experiment. A formal experiment to measure contrast detection thresholds at various spatial frequencies would measure thresholds from images

that each contain just one spatial frequency, such as above on the left. See the Exercises for some examples.

One often plots *contrast sensitivity* which is defined as the inverse of contrast. For example, a contrast detection threshold of 0.02 is equivalent to a contrast sensitivity of 50. Just as contrast threshold curves typically have a minimum at some middle frequency, contrast sensitivity curves have a maximum at that same middle frequency. See exercises for some examples.

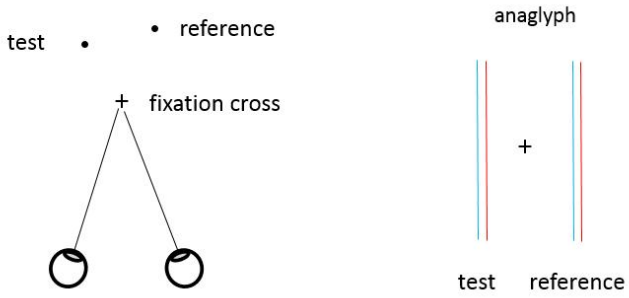
The shape of the contrast sensitivity curves is presumably due to constraints on how many cells in the visual system have their peak sensitivity to different spatial frequencies. We are more sensitive to those spatial frequencies to which more of our cells are tuned. (Recall that DOG cells in the retina and LGN each have a particular range of sizes and this range varies from the fovea to the periphery. As you will see in one of the exercises, contrast sensitivity varies with eccentricity as well.)





Binocular disparity discrimination

Below I illustrate a common depth discrimination task based on binocular disparity cues. The subject fixates (verges) on a cross, so the cross has disparity 0. A test and a reference stimulus is also shown, which are vertical lines presented at different depths. In practice these are displayed on a monitor, so that they produce different binocular disparities which give rise to different depth perceptions.

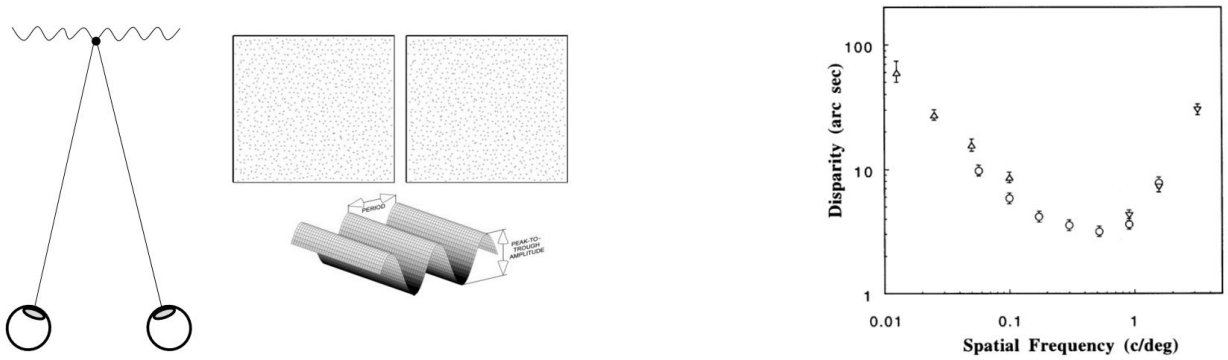


Suppose we hold the disparity of the reference constant, and we vary the disparity of the test, and the task is to say if the test is closer or further than the reference. Responses ("test further") will

go from 0 percent (when test is much closer) to 100 percent (when test is much further). As usual one takes some arbitrary level (say 75 percent) as the threshold.

Note that one obtains a different psychometric function for each reference depth, and hence one obtains a different threshold for each reference depth. One can plot the thresholds as a function of reference depth (not shown here).

There are various versions of such an experiment. In the slides, I showed a square on background configuration. The idea is similar. There is reference depth which might be the square, and a test depth which might be the background (or vice-versa). One can also measure binocular disparity thresholds with 2D sinusoids. One would define a random dot disparity image and the disparity itself would vary as a 2D sinusoid! Since figure below which is from a paper by Banks (2004). The data plot on the right shows the threshold amplitude of the disparity sinusoid as a function of spatial frequency[24]. Note the threshold levels of disparity are remarkably low. A threshold of 5 arc seconds of disparity corresponds to quite a small depth amplitude. (See Exercises.)



Also note that the lowest thresholds occur at spatial frequencies below 1 cycle per degree of visual angle, which is close to a factor of 10 less than the spatial frequencies where the peak sensitivity for luminance contrast sensitivity occurs (which I mentioned earlier is about 3-5 cycles per degree).

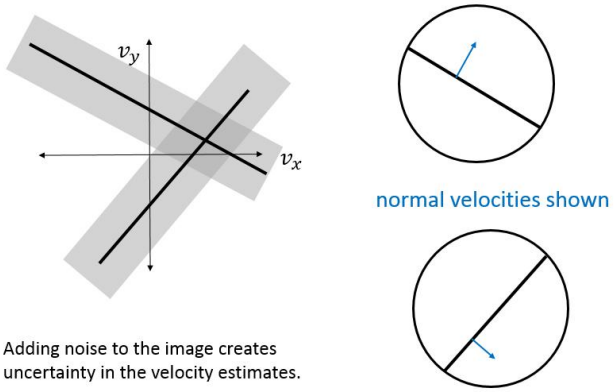
Is this surprising? Not really. For your visual system to measure small spatial *variations* or gradients in disparity from random dot stereograms, it needs to precisely represent the disparities in those regions (since otherwise, how would it know that the disparity is changing?) But to precisely estimate the disparity of a small local region – i.e. to match small local regions of the left and right image – the visual system requires that the image has intensity gradients that are high e.g. sharp edges or lines *and* that the visual system needs to be sensitive to these sharp image structures. As we will see a few lectures from now, sharp edges and lines in the intensity mean that there are high spatial frequencies in intensity present. (Wait for after study break for this.)

Motion Discrimination

Now that we know what noise is for, let's consider how we can add noise to other cues. For motion, if we add pixel noise to a video $I(x, y, t)$, then the partial derivatives of I with respect to x, y, t

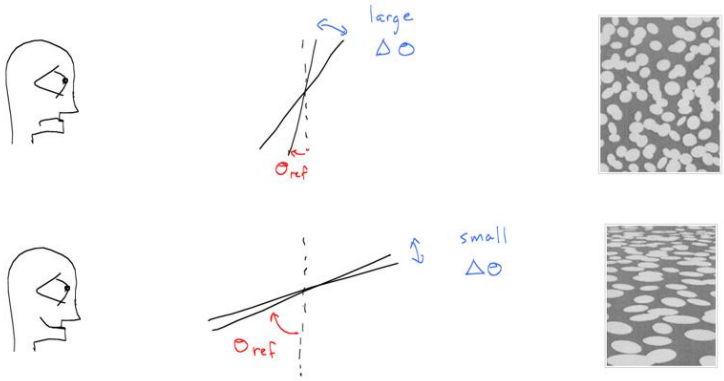
[24] from a paper by Bradshaw and Rogers

essentially become noisy.[25] The result is that the motion constraint line would become uncertain. There would be a *distribution* of motion constraint lines. The intersection of constraints would then not give a unique solution but rather would give a *region* of uncertainty for the solutions. Experiments have shown that this is indeed what happens perceptually. Intuitively, it is not surprising. Adding pixel noise makes it difficult to judge exactly what the image velocity is.



Slant from texture

The last example we discuss today is slant from texture. Here the noise is often not pixel noise, but rather it is randomness in the texture pattern itself, namely the shape and size of texture elements. Even if we assume the visual system knows the mathematics of perspective mappings from 3D, it cannot know the size and shape of the texture elements in 3D if these are random and there will be some undercertainty in the surface slant and tilt. We will discuss this more next lecture, but let me just mention one important idea here, illustrated below.



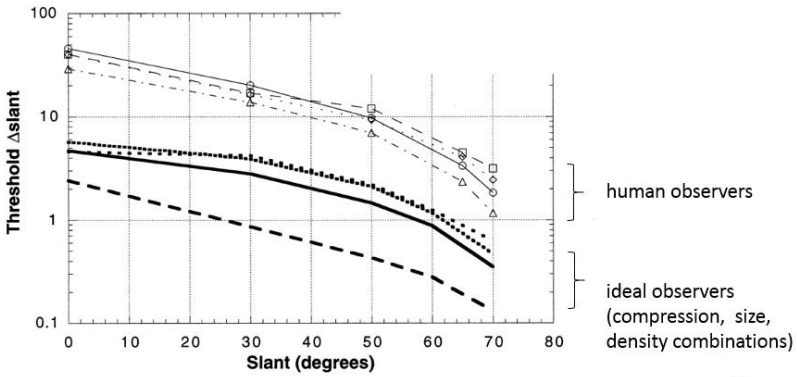
The claim is thatit is inherently more difficult to discriminate the slant of a textured surface when that surface is close to frontoparallel than when it is highly slanted away from the line of

[25] We can't take derivatives of a discrete image, but we can take local differences at neighbhoring pixels, and they would be noisy.

sight. Here the task specifically is: given two images (reference and test) of a textured surface, decide which has greater slant. The claim is that thresholds decrease as slant increases.

To understand why slant is more difficult to discriminate when a surface is close to frontoparallel, notice that slanting a (circular) disk slightly away from frontoparallel doesn't change the projected shape by much; the aspect ratio (width:height) of the disk in the image is $\cos\sigma$ which remains close to 1 when $\sigma \approx 0$, since the cosine curve is flat at 0. When σ is large, however, the cosine function changes more quickly with σ and so a small change in slant σ leads to a larger change in the aspect ratio and the larger change would allow for greater discrimability of the slant of the disk.

The experiments illustrated in the figure above don't use disks. Instead they use ellipses. But the same idea holds, namely that there is relatively less information about the foreshortening of the ellipses when the surface is frontoparallel than when it is slanted. To estimate slant, the visual system needs to use probabilities of various ellipses. The calculation is non-trivial for an ideal observer, and not surprisingly, the visual system does not perform as well as various ideal observers. (The figure below is from Knill 1998 and shows a few different ideal observers that were used to estimate the slant. These different ideal observers used combinations of the texture cues to slant, namely foreshortening (also called "compression"), size, and density. Never mind the details for now. The main point, which you can see in the figure, is that the threshold on slant decreases as the slant increases: we are better at discriminating the slant (angle) of highly slanted surfaces than frontoparallel surfaces. This is true both for ideal observers and for human observers. So, humans are just using the information that is there.



To summarize, plots of thresholds as a function of scene parameters can reveal two different aspects of how the visual system is solving a problem. First, the plots can reveal underlying mechanisms. There may be cells that encode *limited* ranges of size, orientation, disparity, motion, etc. The contrast sensitivity plots from today are a good example. Note that for these types of examples, the performance of human observers might exhibit quite different patterns than the performance of ideal observers; people might not be using information that *is* available, for whatever reason. Second, the plots can reveal how the inherent difficulty of the computational problem varies over different ranges of parameters. Slant from texture is a good example of this. For such examples, human and ideal observer performance tends to exhibit similar patterns, with the human typical performing consistently worse than the ideal (since humans are not ideal).

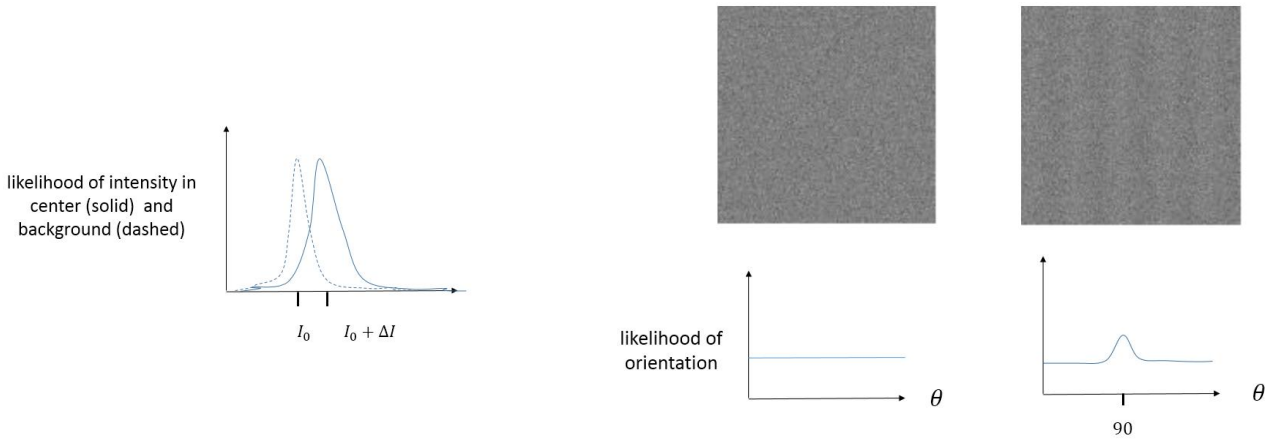
Today I will revisit some of the ideas that I introduced last lecture, and I will try to pose these ideas in terms of probabilities. I'll then use probabilities next lecture when I describe a theory of how the visual system combines different cues.

Motivation

Recall the task of detecting an increment in intensity ΔI in the center of a uniform intensity field I. How can we think about this in terms of probabilities? Suppose that the observer has some uncertainty in the intensity of both the center intensity $I_0 + \Delta I$ and the background I_0. This uncertainty can be due to noise in the monitor or to the noise in the visual system. (Or pixel noise could also be added to the image intensities themselves.) We would like to model the observer's uncertainty that comes from this noise.

We will use the term 'likelihood' as well a probability this lecture. Likelihood has a formal definition which I will review a bit later. For now, let us take the likelihood intuitively to be proportional to the probability that a given image was the result of a background that has intensity value I_0 plus noise, and that the center square has intensity value $I_0 + \Delta I$ plus noise. So the likelihoods are a function of the uniform (pre-noise) intensity of the two image regions. These likelihood functions are sketched below on the left.

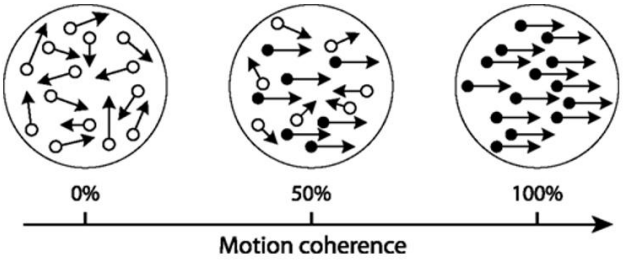
As another example, consider the orientation of a 2D sinusoid image with some additive noise. If the contrast of the sinusoid is low relative to the amount of noise, it will be difficult to discern the orientation of the sinuoid. The image will have been equally likely to occur for any orientation. As the contrast of the sinusoid is increased, the sinusoid will gradually become visible in the image and the likelihood will be elevated at the correct orientation. That is, the noisy image will be more likely to have occured for the correct orientation than for some other orientation.



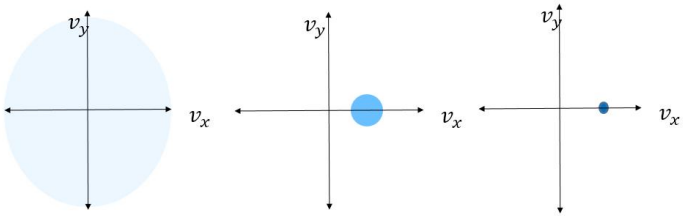
Another example is image motion. Below is a image motion stimulus which consists of dots moving in 2D. Each dot either moves with some fixed image velocity (v_x, v_y) or else it moves with some random velocity with mean 0 drawn from a distribution. The observer's task might be to judge the velocity (v_x, v_y). We assume that the stimulis is filtered through orientation/motion sensitive cells and the percepts must be made based on the responses of these cells. To vary the difficulty of the task – that is, the level of uncertainty in (v_x, v_y) – the experimentor typically varies the fraction

 ©Michael Langer

of dots that move with (v_x, v_y) versus the fraction that move with the random velocity. The fraction that moves with (v_x, v_y) is sometimes called the *motion coherence*.



Below I sketch out informally the likelihood function for (v_x, v_y) for this case as the coherence increases. For 100 % coherence (right), the likelihood will be concentrated around the true velocity. There still will be some spread, however, because the stimulus consists of random dots and there are multiply possible pairings in principle from frame to frame. For 50 % coherence (middle), there will be less of a spread in the likelihood because only half the velocities are in a random direction. For 0 percent coherence (left), there will very high uncertainty in what (v_x, v_y) is. (Indeed it is not even well defined, if no dots move with that velocity.)



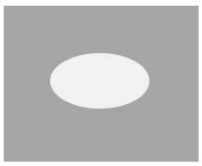
The next example is for depth from binocular disparities. (Figure omitted.) If the stimulus is a random dot stereogram with a center square protruding from a background then the likelihood function will be similar to the one on the previous page for I and $I + \Delta I$ except that now it will be disparity d of the background and the disparity $d + \Delta d$ of the center.

The final example for now is shape from texture. Take a class of stimuli in which the texture is generated from random shaped ellipses placed on a slanted 3D plane. The shapes and sizes and positions of the ellipses are chosen randomly from some distribution. As illustrated in the image on the left below, for a single image ellipse, it is uncertain whether it comes from an ellipse of the same shape and on a frontoparallel plane or whether it comes from a disk on a suitably slanted plane, or from other shaped ellipse on a differently slanted plane.

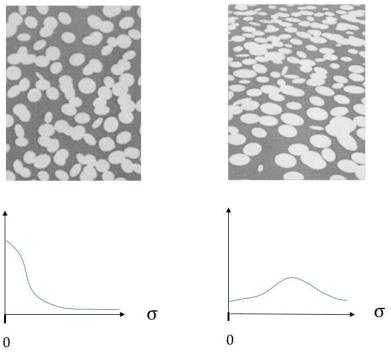
The texture consists of many ellipses on the surface. If the observer knows or assumes the distribution of ellipse shapes on the surface then, for any given image containing many ellipses, some surface slants will be more likely than others. See the two examples below. For the example on the left below, there is a higher likelihood that the plane is close to frontoparallel (slant near 0). For the example on the right, there is a higher likelihood that the plane is slanted backwards at some angle. This is because, for the case on the right, the ellipsoids near the top of the image

 ©Michael Langer

are more foreshortened and more dense and smaller, which is consistent with the deformation that occurs when the surface is indeed slanted back. It would be very unlikely, for example, to have a frontoparallel plane that produced that gradient in the sizes and foreshortening of ellipses observed in the right image.



Is this an ellipse on a frontoparallel plane, or a disk on a slanted plane?



Probability review

Let's now be more formal about what we mean by "likelihood." We will be talking about image random variables I and scene random variables S. Let $I = i$ refer to some 'image'. In practice this could refer to the 2D matrix of image intensities themselves, or it could refer to the responses of a set of cells e.g. photoreceptors, retinal ganglion cells, or simple and complex cells in V1. Or, in the case of shape from texture, it could refer to the image positions and aspect ratios and orientations of the ellipses that define the image and are assumed to be accurately measured.

Let S be a random variable that corresponds to some scene property that is manipulated in the experiment e.g. luminance, depth, orientation, binocular disparity, slant or tilt, etc. The key difference between the I and S is that the I are measurable image quantities whereas the S are scene quantities that are inferred.

We assume that these are discrete random variables. Sometimes this is already the case e.g. 8 bit images, but if not then we can partition the (continuous) sample spaces of I and/or S into bins and consider $I = i$ to be some bin and/or $S = s$ to be some bin. This is essentially what is already done with image intensities, namely we've broken the infinitely many possible intensities into a finite set of possibilities, say 0 to 255.

Here is the notation that we'll be using for basic probability definitions. You should be familiar with this. If not, you'll need to brush up.

- *joint probability* $p(I = i, S = s)$
- *marginal probability*

$$p(I = i) = \sum_{s \in S} p(I = i, S = s)$$

$$p(S = s) = \sum_{i \in i} p(I = i, S = s)$$

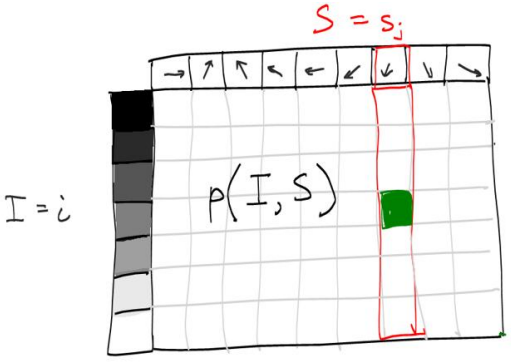
- *conditional probability*

$$p(I = i|S = s) = \frac{p(I = i, S = s)}{p(S = s)}$$

$$p(S = s|I = i) = \frac{p(I = i, S = s)}{p(I = i)}$$

You can think of the joint probability function $p(I,\ S)$ definitions as follows. (See slides for illustrations of the marginal or conditional probabilities.) Consider a 2D matrix where the rows are different values of $I = i$. Each row is an image or the set of DOG or Gabor (or complex cell) responses to an image. The different rows are not single pixels or single responses from one DOG or Gabor cell, etc, but rather they are entire images or sets of responses.

Each column $S = s$ could represent different values of a scene parameter. I have indicated vectors in the illustration: these could be 2D image motion vectors, or image orientations (a 2D sinusoid) or surface slants or tilts.



Likelihood function

The *likelihood* function is the conditional probability $p(I = i|S = s)$. It depends on both I and S. In the problems that I discussed earlier, I mentioned 'likelihood' intuitively and it was tempting to think it is as a probability of a scene. But that is not quite the idea. Rather the likelihood is the probability that an image $I = i$ occurs, in the case that the scene was $S = s$. If we fix s and vary i, then $p(I = i|S = s)$ is a probability density function over i and it integrates to 1. However, that's not what we are interested in here. Rather, the image i is given and we are comparing different scenes s as possible ways of explaining i. If we integrate over s (for fixed i), we do not get 1. That is why we do not call $p(I = i|S = s)$ a probability of s, and we instead some other word was invented, namely 'likelihood'.

The vision system doesn't know what this scene s is. It is only given the image $I = i$. The *maximum likelihood* method is to choose the scene $S = s$ that maximizes $p(I = i|S = s)$, that is, the image $I = i$ arises with a greater likelihood for that scene $S = s$ than for any other scene. By 'likelihood' here, we are referring to something random, namely the randomness of image formation. We are not referring to the randomness in S, since $S = s$ has already occured. A few examples should help with this rather subtle distinction.

Maximum likelihood for an intensity increment

Take again the example of the intensity increment. Suppose we have a background region of some intensity I_0 and we have a center square region of some intensity $I_0 + \Delta I$. Assume that the visual system knows the location and size of the square. Also assume that noise is added to each pixel in center and surround. For example, one often assumes that the noise values $n(x, y)$ have a Gaussian probability density with mean 0 and variance σ_n^2. As mentioned earlier, image intensities are quantized into discrete values, so we would need a discrete approximation but let's not concern ourselves with that detail.

Suppose the task is to decide if there an intensity increment as opposed to decrement. Say one solves this by estimating I_0 in the surround and $I_0 + \Delta I$ in the center. Let's just take the case of estimating I_0 in the surround. We can write the probability of one noisy pixel in the surround, given I_0 is as follows.

$$p(n(x, y) = n_i) = \frac{c}{\sqrt{2\pi}\sigma_n} e^{-\frac{n_i^2}{2\sigma_n^2}}$$

The constant c is there because the Gaussian is continuous density and I want to write the probability as discrete, i.e. we discretize the range of noise values into bins.

The probability of a particular set of noise values $p(n)$ in the surround patch depends on the values of I_0 and on the (noisy) image $i = I_0 + n$, and can be written as a likelihood:

$$p(n) = p(I = i \mid I_0)$$

We want to find the value of I_0 that has the highest likelihood, that is, the I_0 value such that the noise n required to produce the given image i would have had the highest probability of occuring.

It is standard to assume the noise at different pixels is independent.[26] Let N be the number of pixels in the surround region. Then the joint probability of the N noise values $n(x, y)$ for the different (x, y) is the product of probabilities of noise for the individual pixels:

$$\begin{aligned} p(n) &= \Pi_{x,y} (\frac{c}{\sqrt{2\pi}\sigma_n})^N e^{-\frac{n(x,y)}{2\sigma_n^2}} \\ &= constant * e^{-\sum_{x,y} \frac{n(x,y)^2}{2\sigma_n^2}} \\ &= constant * \ e^{-\sum_{x,y} \frac{(I(x,y)-I_0)^2}{2\sigma_n^2}} \end{aligned}$$

If you want to see a particular example, then here is some matlab code.

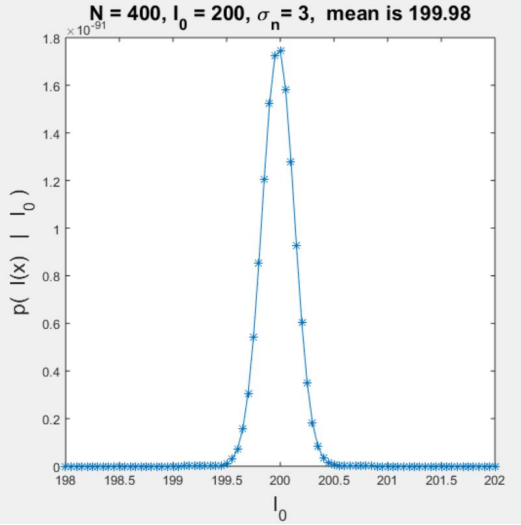
`http://www.cim.mcgill.ca/~langer/546/MATLAB/likelihood.m`

The plot below was generated from that code. It plots the likelihoods of ΔI when the true ΔI is 200 and there are 400 noise values. (Here we just set $I = 0$ for simplicity.) If you run the code a few times, you'll see that the likelihood function changes quite a lot between runs. Even with 400 samples, there is a lot of variability.

Note that the title shows the 'mean' likelihood, which I am taking as a proxy for the 'maximum' since the distribution is roughly symmetric here. (The max is harder to compute since one would need to interpolate between the discrete samples.) As you can see if you run it a few times, the mean likelihood typically has a value close to the true value of $I_0 = 200$.

[26] Two random variables X_1 and X_2 are independent if for any $X_1 = x_1$ and $X_2 = x_2$, their joint probability $p(X_1 = x_1, X_2 = x_2)$ function is equal to the product of their marginal probabilities $p(X_1 = x_1)p(X_2 = x_2)$.

 ©Michael Langer



How can one define likelihood functions for other problems such as deciding on the orientation(s) present in the neighborhood of a pixel, or the binocular disparity or image motion or the slant and tilt of a surface ? I will not go into details on the mathematics here because it is more an exercise in probability than in vision. But let's at least sketch the idea of how you could construct such a model.

Take the case of binocular disparity. Recall Assignment 2 where you computed the responses of V1 complex cells tuned to particular disparities, or you considered MT cells tuned to particular image velocities (v_x, v_y). In each case, it is possible in principle to write down a mathematical model. For example, take a region of constant disparity d, the likelihood of the responses $r_{d_1}, r_{d_2}, \cdots, r_{d_k}$ of cells tuned to different disparities, if the actual disparity was d. One could try to write down a model:

$$p((r_{d_1}, r_{d_2}, \cdots, r_{d_k}) \mid disparity = d).$$

This is not easy to do, but it can be done. You could something similar for the 2D motion estimation problem. You have shift detector cells tuned to different orientations and motions. You could come up with a likelihood function for the responses of these cells:

$$p((r_{\theta_1, speed_1}, \ \ldots \ , r_{\theta_i, speed_j}, \ldots) \mid \text{actual velocity is } (v_x, v_y).$$

where $r_{\theta_i, speed_j}$ is the response of cells with peak tuning to spatial orientation θ and $speed_j$ in that orientation. Again, not easy, but it can be done.

For shape from texture, it is also possible to come up with likelihood functions. I mentioned the work of David Knill whose model assumed that textures were ellipses distributed over a slanted plane. The ellipses on the surface had a random distribution, in size, orientation, and elongation (aspect ratio) and Knill wrote out precise mathematical model for this. He also considered the projection of the ellipses into the image, which gave rise to image distributions of size, orientation, and elongation. He was able to write down likelihood functions of the form:

$$p(\text{image ellipses} \mid \text{surface slant})$$

This allowed him to estimate the maximum likelihood of a surface slant for a given image. Note that there is no intensity pixel noise here. Rather, the randomness is in the distribution of ellipses themselves and it is assumed that the ellipses can be measured in the image.

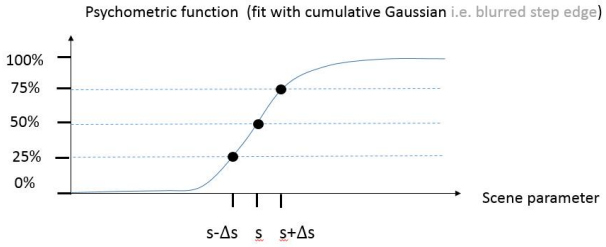
Likelihood functions and psychometric functions

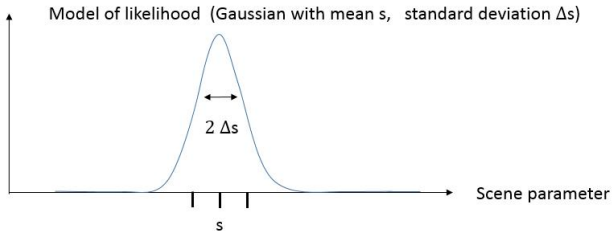
The mathematical models that I mentioned above don't necessarily predict the behavior of real observers. How can we relate real observer behavior to such models?

One common approach is illustrated below. Given a psychometric function in some experiment, one fits this function using a cumulative Gaussian function which by definition is the integral of a Gaussian from negative infinity up to some value x:

$$cdf(x) \equiv \int_{-\infty}^{x} G(x', \mu, \sigma) dx'$$

where *cdf* stands for "cumulative density function". Here μ is the mean and σ is the standard deviation for Gaussian.





The experiment might be to say whether some test stimulus has a greater or smaller scene parameter s than some reference stimulus. We've seen several examples:

- a background intensity I and a central square with intensity $I + \Delta I$
- a background random dot pattern with disparity d and a central square with disparity $d + \Delta d$

The subject's performance in the task is described by the psychometric function. If one fits the performance using a cumulative Gaussian model with mean at some scene parameter value s (e.g. the I or d in the above examples), then the subject's uncertainty in doing the task can be associated with the standard deviation of the Gaussian whose cdf is the best fit to the psychometric function.

It is common to treat this fitted Gaussian as if it were the person's likelihood function[27] which they use to estimate the scene parameter s. As we will see next lecture, this is useful for considering how people combine different visual cues.

[27] We don't really believe that people have a likelihood function in their brains — any more than we believe that your brain solves differential equations when you walk and throw a ball. But let's not get into that philosophical issue here.

Up to now, we have been considering one type of image information at a time e.g. intensity, stereo, motion, texture, shading. But in real situations, an observer has multiple sources of information available and would like to combine this information. We will use the term 'cue combination' for this. By 'cue', we mean both a particular type of image and scene information, along with a mapping which allows an observer to perceive the scene property given the image property.[28] We have discussed in particular how texture is a cue for slant and tilt; binocular disparity, blur, and motion parallax are cues for depth; shading is a cue for local surface orientation and surface curvature.

Cue combinations

What do we do when we have multiple cues available? Suppose we have two sources of image information which I will simply call I_1 and I_2. These variables represent some image measurement, such as binocular disparity, motion, or a description of a texture. We wish to use I_1 and I_2 to estimate a scene variable S.

Let $p(I_1|S)$ and $p(I_2|S)$ be the likelihood functions for each cue on its own and let $p(I_1, I_2|S)$ be the likelihood function of the two cues together. It is common to *assume* I_1 and I_2 are "conditionally independent":

$$p(I_1,\ I_2 \mid S) \ = \ p(I_1 \mid S)\ p(I_2 \mid S).$$

Intuitively, for a fixed scene, conditional independence says that the value of one image variable I_1 tells us nothing about the value of the other image variable I_2. For example, I_1 might be the sizes of texture elements and I_2 might be the foreshortening of texture elements. Or I_1 might be all the texture cues and I_2 might be the binocular disparities of the texture elements. Note: conditional independence is just a model. In reality, there might be a weak dependence, but we ignore this dependence to keep the model simple.

Suppose the likelihood functions $p(I_1 \mid S = s)$ and $p(I_2 \mid S = s)$ both have a Gaussian shape[29] and with means s_1, s_2 and variances σ_1^2, σ_2^2, respectively.

$$p(I_1 = i_1|S = s) \ = a_1 e^{-\frac{(s-s_1)^2}{2\sigma_1^2}}$$

$$p(I_2 = i_2|S = s) = a_2 e^{-\frac{(s-s_2)^2}{2\sigma_2^2}}.$$

where a_1 and a_2 are constants. If we assume conditional independence, then the likelihood function $p(I_1 = i_1,\ I_2 = i_2 \mid S = s)$ is just the product of these two likelihood functions.

What is the s that maximizes the likelihood $p(I_1 = i_1,\ I_2 = i_2 \mid S = s)$? We next show that the maximum likelihood estimate is a linear combination of the maximum likelihood estimates of the two cues when they are on their own. We want to find the s that maximizes

$$p(I_1|S)\ p(I_2|S) = a_1 a_2 e^{-\frac{(s-s_1)^2}{2\sigma_1^2}}\ e^{-\frac{(s-s_2)^2}{2\sigma_2^2}}$$

[28] Usually the mapping is from scene to image, whereas the vision system wants to map from image to scene, which is why vision is a more difficult problem than graphics!

[29] Recall that this doesn't mean that they are Gaussian probability functions, in the sense that they have unit area. Likelihood functions in general do not integrate to 1 when you integrate over the scene variable $S = s$.

 ©Michael Langer

and so we want to minimize

$$\frac{(s-s_1)^2}{2\sigma_1^2} + \frac{(s-s_2)^2}{2\sigma_2^2}.$$

Take the derivative with respect to s and set it to 0. This gives

$$\frac{s-s_1}{\sigma_1^2} + \frac{s-s_2}{\sigma_2^2} = 0$$

and so

$$s = (\frac{s_1}{\sigma_1^2} + \frac{s_2}{\sigma_2^2}) \ / \ (\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2})$$

Note that this is of the form

$$s = w_1 s_1 + w_2 s_2$$

where $0 < w_i < 1$ and $w_1 + w_2 = 1$. In particular,

$$w_1 = \frac{\sigma_1^{-2}}{\sigma_1^{-2} + \sigma_2^{-2}} \qquad w_2 = \frac{\sigma_2^{-2}}{\sigma_1^{-2} + \sigma_2^{-2}}$$

which can be rewritten:

$$w_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \qquad w_2 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}$$

For example, if $\sigma_1 \ll \sigma_2$, then $w_1 \approx 1$ and $w_2 \approx 0$.

This *linear cue combination* method says that if one cue is more reliable than the other, then the more reliable cue should have a heavier weight. The linearity might not be intuitive, however. You might think that a "winner take all" approach would be better, namely that one should put *all* the weight on the more reliable cue. To understand why "winner take all" is wrong, note that we are assuming conditional independence of the cues, which intuitively means that $I_1 = i_1$ and $I_2 = i_2$ give you different information about s. Even though one cue may be more reliable than the other, the less reliable one still gives information and so it should not be entirely ignored.

Many experiments have shown that this cue combination theory often does describe human performance. For example, one can vary the noise in one of the cues, and show that the psychophysical thresholds change, as if the vision system were giving less weight to that cue. (Recall that this requires some assumptions, namely that the psychometric curves can be interpreted in terms of likelihood functions.)

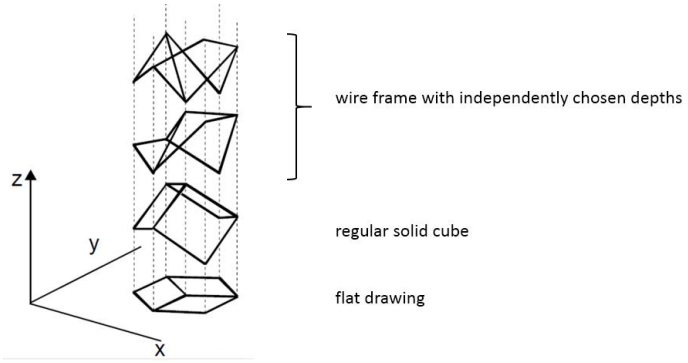
Bayes Rule and MAP (Maximum *a Posteriori*) estimation

We have thus far concentrated on the likelihood function $p(I = i \mid S = s)$. One estimates the scene s that produces the image i, such that the probability of the noise that is required to explain the s to i mapping is maximal. One limitation of this maximum likelihood approach is that it ignores the fact that some scenes s have a higher probability of occurring than other scenes.

As an example, consider that an image of a cube can be arise from several different wire frame figures in 3D. One of these figures is a 3D cube, but there are infinitely many others since one can move the depths Z of the points of the cube without changing the image. (Here we are assuming for simplicity that the image formation occurs by a projection that is parallel to the Z axis, but the same projection ambiguity holds if we use perspective projection.)

 ©Michael Langer

We can think of the likelihood function as being uniform over s, since any of the scenes s that project to the image shown is as good as any other one in accounting for the image. (We aren't formally writing down a model of the 'noise' here, but we could do so in terms of a slight jittering of each vertex of the wireframe in 3D and its corresponding jittering in 2D.)



Why does the visual system prefer the cube interpretation over any particular other wire frame interpretation? One idea is that cubes have higher probability of occuring in our world than individual complex shapes that happen to look like cubes just because we are viewing them from a particular direction and some accidental alignment. If the visual system takes account of the higher probability of cubes occuring (than random wire shapes that happen to look like cubes), then it would infer a 3D cube when it sees an image that is consistent with a 3D cube.

A more elaborate example is the *Ames Room* illusion. See the videos:
`http://www.youtube.com/watch?v=Ttd0YjXF0no`
`https://www.youtube.com/watch?v=gJhyu6nlGt8`
An Ames room is a 3D room which is viewed in perspective. The room's walls and floor have a 3D trapezoidal shape, but the viewing position within the room is chosen so that the walls and floor have the same image projection as a 3D cube room. We perceive the room as a cube, even though it isn't. And this leads to some strange consequences when there are other objects within the room.

The video shows that two people who are in different places in the Ames room can have quite different perceived 3d sizes. In the first video above, people move in the room and seem to change size as they change position. It is remarkable that the visual system would interpret people as changing size rather than correctly perceive the actual (non-cube) shape of the scene.

Both of the previous examples seem to work because the visual systems prefers a regular shape (cube or room) over a non-regular one. Rather than trying to come with a theory of 'regularity', we will express this idea in terms of probabilities by saying that regular shapes occur more frequently than particular non-regular shapes that happen to look regular. Specifically we can capture this idea by considering the marginal probability $p(S)$ over scenes, and giving a larger value $p(S = s)$ for particular scenes s. This marginal scene probability is called the scene *prior*, and it plays a role in Bayes Rule (or Bayes Theorem) which I will now derive, and which most of you are familiar with since it is commonly taught in basic probability courses.

One can write the joint probability function $p(I, S)$ in terms of conditional and marginal probabilities in two ways:

$$\begin{aligned} p(I, S) &= p(I|S)\ p(S) \\ p(I, S) &= p(S|I)\ p(I). \end{aligned}$$

Equating right sides and isolating $p(S|I)$ gives us *Bayes Rule*:

$$p(S|I) = \frac{p(I|S)\ p(S)}{p(I)}$$

The function $p(S|I)$ is called the *posterior* probability function. It depends on the prior and on the likelihood. The posterior is really what we are interested in: we want to estimate the probability of a scene $S = s$, given an image $I = i$. One often solves for the maximum of the posterior – or *maximum a posteriori*, as its usually called.

Note that the posterior depends on the prior probability $p(I = i)$ of an image i occuring. One typically does not have a model for $p(I)$, and one does not care about $p(I)$. The reason is that one wants to estimate $S = s$, but $p(I = i)$ doesn't depend on any particular $S = s$. One often solves for the maximum of the posterior – or *maximum a posteriori*, as its usually called – one can ignore the dependence on $p(I = i)$. The reason is that one wants to know the probably of a scene $p(S = s)$ *given that* image i has already occurred.

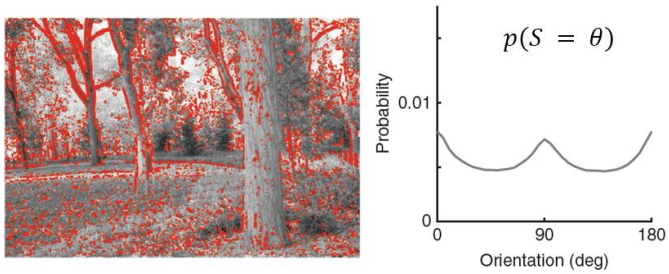
Also note that if the prior $p(S)$ is uniform over S then finding the maximum of the posterior is equivalent to finding the maximum of the likelihood. In many cases, one does not know the prior or one has reason to believe that the prior is relatively flat. In this case one can treat the prior is roughly constant over the region of the parameter S that one is considering.

Natural Image and Scene Statistics

Over the past two decades, researchers have begun to collect data and to make quantitative models of image and scene statistics in order to gain insight into the priors and likelihood functions that the visual system seems to use. Image statistics are relatively easy to come by: one takes many images and applies operators such as difference of Gaussian or Gabor filters or others, and fit models to the responses. Scene statistics are more challenging, since they require more sophisticated imaging devies for measuring 3D geometry. But these devices are now available e.g. lidar.

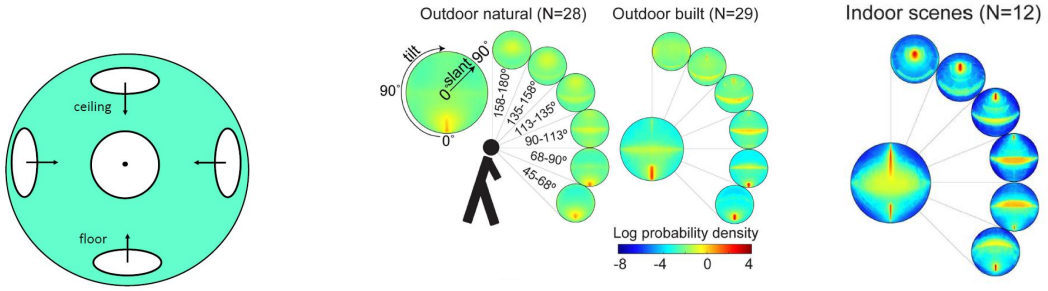
Here I discuss two examples. The first is a study[30] that examined line/edge orientations in natural images. They used computer vision methods to measure the frequency of lines/edges of different orientations. As shown on the right, there are about 50 percent more edges that are vertical or horizontal than are diagonal. This distribution was used to model human percepts of orientation in a psychophysical experiment. It has been known for many years that human observers are better at discriminating orientations that are near vertical or near horizontal than near oblique orientations. This study was able to relate the performance in such orientation tasks to a probability model of likelihoods and priors. (Details omitted.)

[30] Girshick et al



A second example is study[31] of the distributions of slants and tilts of surfaces in various enviroments including outdoor and indoor. Surface depth maps were imaged using lidar techniques (like radar, but uses light). They then fit planes to local surface patches. This gave them the frequencies of different slants and tilts.

One subtlety here is that slant and tilt are measured with respect to some XYZ coordinate system where Z is depth. If one is looking at the ground, then Z will be different than if one is looking upwards towards the ceiling. In the plots below, the slant and tilts are defined with respect to different viewing directionss, specifically elevations. For example, think of the $45-68°$ plot as have a Z axis centered at 56 degrees up from the gravity vector and considering the surfaces that are visible in 11 degree neighborhood around the Z axis. They calculate the slant and tilt at each surface point in that neighborhood and they do that for many different scenes. They then used a color map to plot the frequency distribution of slants and tilts. The camera was always at a height close to 2m above the ground so that the statistics correspond to what a typical adult will observe.



Examples of the data are shown above. For example, consider the 90-113 deg elevation which goes from the horizon (parallel to the ground) to 23 degrees above the horizon. For these viewing directions, there is a ridge of peaks for tilts of 0 or 180 deg and all different slants - see the yellow horizontal stripes in the outdoor scenes. This band is presumably due to trees and walls which are vertical surfaces and so the normal is always near perpendicular to the gravity direction. A similar stripe appears at other viewing elevations but the stripe is shifted to other slants and tilts because viewing direction is not horizontal.

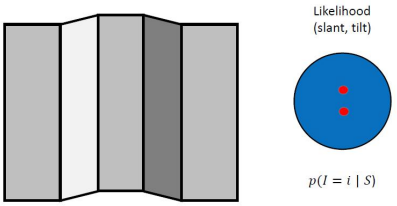
Another example is that in the indoor scenes at elevations above the horizon, there is a hot spot that the corresponds to points on the ceiling (tilt = 90). The slant and tilt of this hot spot shifts with the viewing elevation. There is also a hot spot for floor slants and tilts (tilt = -90 deg) when the viewing direction is in the 45-68 degree range which is below the horizon.

[31] Adams and Elder

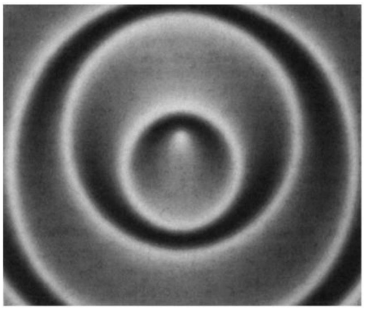
The main point here is that the distribution of slants and tilts in the world is highly non-uniform. The details might turn out to be very interesting for understanding perception, if it turns out that these details correspond to our preferences in perception. It has been shown that the visual system does prefer floor slopes over ceiling slopes, and we will discuss examples below. It hasn't (yet) been shown that there are differences in perception that correspond to the detailed probabilities differences shown the plots above. But perhaps one day that will be shown too.

Depth reversal ambiguity in shape from shading (on a sunny day)

Recall the corrugated plaid illusion from lecture 12. The figure below shows a simplified version of it which can either be interpreted as a ridge (convex) or valley (concave). When we perceive a ridge, the dominant lighting direction is from the left, and the surface is sloped slightly upwards (like a floor). When we perceive a valley, the dominant lighting direction is from the right, and the surface is sloped slightly downwards (like a ceiling). Both of these interpretations are consistent with the image information. We can think of a likelihood function with two corresponding peaks.



Below is a similar example (due to Reichel and Todd 1990). The center region of the shaded pattern can be seen either as a local hill or a local valley. These local curvature percepts depend on seeing the overall surface slant as slightly floor-like or slightly ceiling-like, respectively. Both percepts are valid for the given image. Again we can think of a likelihood function for the surface, which has two maxima corresponding to the two different surface interpretations for this image.



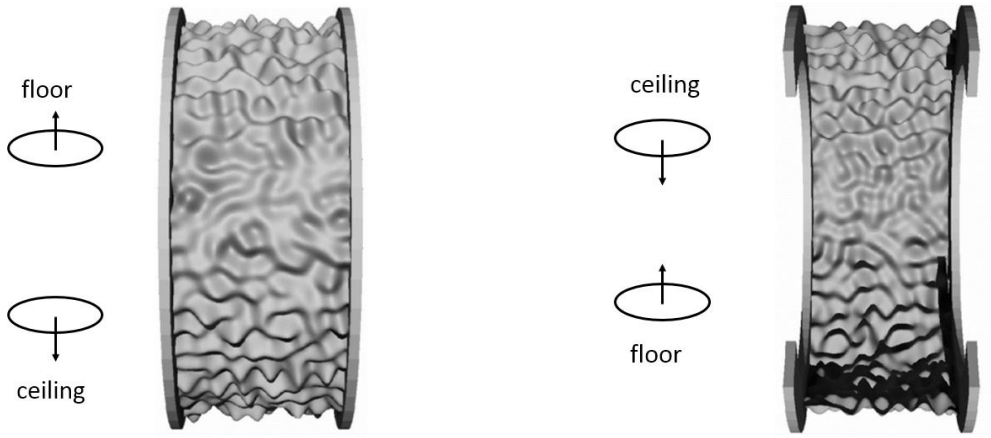
Both of the above are examples of a "depth reversal ambiguity". See the Exercises for which this ambiguity exists. As we will see next, the visual system often relies on prior assumptions to resolve such two fold ambiguities.

Priors for light from above and global convexity

It has been known for a few hundred years that two-fold ambiguities in shape from shading exist and that the visual system often resolves them by preferring a solution that is consistent with the light being from above. This can be demonstrated informally, as familiar objects such as faces look strange when illuminated from below. It can also be shown formally in shape from shading experiments. Subjects tend to perceive shapes that are consistent with light from above, rather than depth reversed shapes that are consistent with light from below. This prior from light from above is not surprising since, more often than not, scenes are illuminated from above.

Another prior that is well known is surface convexity. We prefer to see individual objects as having a solid shape, that is, overall convex rather than overall concave like a mask. Again this is not surprising since most objects have an overall shape that is solid and hence more convex than concave.

Before I came to McGill as a professor, I did experiments that investigated the prior assumptions $p(S)$ that we use to disambiguate the surface shape in situations of ambiguities. The image classes that I came up with pitted three priors against each other. See below. The surface on the left is overall convex and is illuminated slightly from above the line of sight, and the surface on the right is overall concave and is illuminated from slightly below the line of sight. Surface points that are either just above or just below the center of each image have an overall floor or ceiling slant.



I showed subjects many such images and marked single points on these images, and I asked them say if the points were on a 'hill' or in a 'valley'. Subject's percentage correct scores in each combination of conditions (light direction, floor or ceiling region, global shape) could be modelled as if they were using prior assumptions to disambiguate the two-fold ambiguities. In a nutshell, their percent correct scores were 50 percent plus or minus about 10 percent for each of the three priors. For example, in the floor region for the image on the left, subjects were about 80 percent correct (illuminated from above, floor, overall convex), whereas in the ceiling region in the figure on the right (illuminated from below, ceiling, overall concave) subjects were about 20 percent correct. I've looked at these stimuli thousands of times and I still tend to interpret the hills and valleys using these priors.

Convolution

Recall the definition of *cross-correlation* from lecture 4 which I write here for 1D functions:

$$f(x) \otimes I(x) \equiv \sum_u f(u-x)\ I(u)\ .$$

Convolution is defined slightly differently, namely:

$$f(x) * I(x) \ \equiv \ \sum_u f(x-u)\ I(u).$$

Note that the only difference here is that the argument of $f(\ )$ is now flipped. So whenever we have a cross-correlation, we can think of it as a convolution with a flipped function, and vice-versa. Also note that if $f()$ happens to be symmetric (like a Gaussian), then there is no difference between convolution and cross-correlation.

In general though, there is a difference in how we think of cross correlation and convolution. We think of cross-correlation as sliding a template function f across another function and taking the inner product. We think of convolution $f(x) * I(x)$ as adding up shifted versions of the function $f(x)$, namely $f(x-u)$. Each shifted version is weighted by the value $I(u)$, where u is the shift.

ASIDE: The question comes up of what to do when $f(x-u)$ is not defined for some value of $x-u$. This should be familiar to you, since a similar problem arose when we defined cross-correlation in lecture 4, and it has come up in assignments. Here we can do the same thing as we did there, and just 'zero-pad' the function $I()$ beyond the domain where it is defined. An alternative, which we will mention later is to treat $I()$ as periodic.

Algebraic properties of convolution

One surprising and useful property the convolution operation is that it *commutative*: one can switch the order of the two functions I and f in the convolution without affecting the result. The property does *not* always hold for cross-correlation.

To prove that convolution is commutative, we pad $I(x)$ and $f(x)$ with zeros. This allows us to take the summation from $-\infty$ to ∞.

$$I(x) * f(x) = \sum_{u=-\infty}^{\infty} f(x-u)I(u)$$

Using the substitution $w = x - u$, we

$$I(x) * f(x) = \sum_{w=-\infty}^{\infty} f(w)I(x-w) \ = I(x) * f(x)$$

If you think of I as a signal and f as a filter then you don't need to be concerned about order of writing $I * f$ or $f * I$ since they are the same.

 ©Michael Langer

A second important property of convolution is that it is *associative*:

$$I \ * \ (f_1 \ * \ f_2) \ = \ (I \ * \ f_1) \ * \ f_2$$

Again the proof is simple, and you should work it out for yourself.

Why are these properties useful? Often, in signal processing, we perform a sequence of operations. For example, you might average the pixels in a local neighborhood, then take their derivative (or second derivative). The algebraic properties just described give us some flexibility in the order of operations. For example, suppose we blur an image $I(x)$ and then take its local difference. We get the equivalent result if we take the local difference on the blur function and convolve the result with the image:

$$(D(x) * B(x) \ ) \ * I(x) = D(x) \ * (B(x) \ * \ I(x)).$$

One final property of convolution is that it is *distributive*:

$$(I_1 + I_2) \ * \ f \ = \ I_1 * f \ + \ I_2 * f$$

This is also simple to prove and I leave it to you as an exercise. This property is also useful. For example, if $I_1 = I(x)$ is an image and $I_2 = n(x)$ is a noise function added to the image, then if we blur the "image+noise," we get the same result as if we blur the image and noise separately, and then add the results together.

Impulse functions, and impulse response function

Define a "delta" function

$$\delta(x) = \begin{cases} 1, & x = 0 \\ 0, & \text{otherwise.} \end{cases}$$

$\delta(x)$ is also known as an *impulse function*. It is straightforward to show that, for any function $I(x)$,

$$I(x) = \delta(x) * I(x).$$

Another way to interpret the above equation is to think of a function $I(x)$ as a sum of delta functions

$$I(x) = \sum_{u=-\infty}^{\infty} \delta(x-u) I(u) \ ,$$

namely, if we put a delta function at each value of u and multiply that delta function by the value $I(u)$, then we get the original function.

Finally, suppose we have a mapping ("convolve with $f(x)$")

$$I(x) \rightarrow I(x) * f(x)$$

In this case, we often refer to $f(x)$ as an *impulse response function*. The reason is that if $I(x)$ were an impulse $\delta(x)$ then it would map to $f(x)$. That is, $f(x)$ is the response (output) when the stimulus (input) is $\delta(x)$.

Recall that convolution $f(x) * I(x)$ is defined by adding up shifted versions of the function $f(x)$, where each shifted version is weighted by a value $I(u)$ where u is the shift. Thus, thinking $I(x)$ as a sum of delta functions, we see now that $f(x) * I(x)$ can be interpreted a sum of impulse response functions $f(x-u)$ shifted by different amounts u and weighted by different amounts $I(u)$.

Sinusoids and convolution

We next show that sinusoids have a special behavior under convolution. Take a cosine function with k cycles from $x = 0$ to $x = N$, where k is an integer,

$$\cos(\frac{2\pi k}{N}\ x).$$

Note that this cosine function has the same value at $x = N$ as at $x = 0$. Suppose we were to convolve the cosine with a function $h(x)$ which is defined on $x \in 0, \ldots, N-1$:

$$h(x) * \cos(k\ \frac{2\pi}{N}\ x)\ =\ \sum_{x'=0}^{N-1} h(x')\ \cos(k\frac{2\pi}{N}(x - x'))$$

[**BEGIN ASIDE** (I did not include this in the lecture slides since it is just a calculation.)]

Recalling the trigonometry identity from Calculus 1,

$$\cos(\alpha + \beta)\ =\ \cos\alpha\cos\beta + \sin\alpha\sin\beta$$

we can expand the cos() in the summation on the right side of the above equation, and so the right hand side is just a sum of sine and cosine functions with variable x and constant frequency k. Thus, it can be written

$$h(x) * \cos(k\ \frac{2\pi}{N}\ x)\ \ =\ \ a\cos(\frac{2\pi}{N}\ kx) + b\sin(\frac{2\pi}{N}\ kx) \tag{12}$$

which was **Claim 1** in the lecture slides. The values of a and b depend on k and on the function $h(x)$ as follows:

$$a = \sum_{x'=0}^{N-1} h(x')\ \cos(k\frac{2\pi}{N}x')$$

$$b = \sum_{x'=0}^{N-1} h(x')\ \sin(k\frac{2\pi}{N}x')$$

which are just the inner products of the N dimensional vectors $h(\cdot)$ with a cosine or sine of frequency k, respectively. **[END ASIDE]**

Let's simplify Eq. (12). Let (a, b) be a 2D vector, and define angle ϕ such that

$$(\cos\phi, \sin\phi) = \frac{1}{\sqrt{a^2 + b^2}}(a, b).$$

Then

$$\begin{aligned} h * \cos(k\ \frac{2\pi}{N}\ x) \ \ &= \ \ \sqrt{a^2 + b^2}\ (\cos(\phi)\cos(\frac{2\pi}{N}\ kx) + \sin(\phi)\sin(\frac{2\pi}{N}\ kx)) \\ &= \ \ \sqrt{a^2 + b^2}\ \cos((\frac{2\pi}{N}kx) - \phi) \end{aligned}$$

which was **Claim 2** in the lecture slides. The quantity $\sqrt{a^2 + b^2}$ is called the *amplitude* and ϕ is called the *phase*. The amplitude and phase depend on frequency k and on the function $h(\cdot)$.

 ©Michael Langer

To briefly summarize, we have shown that convolving a cosine with an arbitrary function $h(x)$ gives you back a cosine of the same frequency k, but with possibly different amplitude and possibly phase shifted in position x. (Exactly the same argument can be made for a sine function.) These amplitude and phase changes turn out to be very important, as we'll see in the next few weeks when we discuss sound processing by the ear.

One final point: I made **Claim 3** in the slides, namely that any function $I(x)$ can be written as a sum of sine and cosine functions:

$$I(x) = \sum_{k=0}^{\frac{N}{2}} a_k \ \cos(\frac{2\pi}{N}kx) + \sum_{k=1}^{\frac{N}{2}-1} b_k \ \sin(\frac{2\pi}{N}kx)$$

Because of time constraints and because we will not use this representation, I won't prove that claim. Instead, what I will do (next lecture) is give you an alternative representation, called the *Fourier* representation, which is slightly different. The Fourier transform requires that we use complex numbers, so I will spend the rest of the lecture reviewing the basics.

Complex numbers (review)

To decompose functions into sines and cosines we are going to use complex variables. Recall that a complex number c consists of a pair of numbers, (a, b) called the "real" and the "imaginary" part. One often writes this pair using the notation

$$c = a + bi.$$

We define *addition* of two complex numbers by adding their real and imaginary parts separately:

$$c_1 + c_2 = (a_1 + a_2, b_1 + b_2)$$

or

$$c_1 + c_2 = (a_1 + a_2) + (b_1 + b_2)i.$$

We can define *multiplication* of two complex numbers by writing the two numbers in polar coordinates:

$$c_1 = r_1(\cos\theta_1 + i \ \sin\theta_1)$$
$$c_2 = r_2(\cos\theta_2 + i \ \sin\theta_2)$$

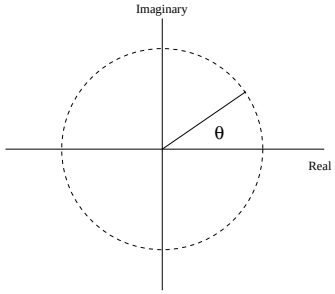
and defining the product $c_1 \ c_2$ to have a length $r_1 r_2$ and an angle $\theta_1 + \theta_2$:

$$c_1 \ c_2 = r_1 r_2(\cos(\theta_1 + \theta_2) + i \ \sin(\theta_1 + \theta_2)).$$

For example, take the case that $a = 0, b = 1$, or i. For this number, $r = 1, \theta = \frac{\pi}{2}$. So $c^2 = i^2$ has $r = 1$ and $\theta = \frac{\pi}{2} + \frac{\pi}{2} = \pi$ and so $i^2 = \cos(\pi) = -1$. Thus

$$i^2 = -1.$$

There is really nothing mysterious about this number i, once you understand that we are *defining* multiplication on pairs (a, b) of numbers in this special way.



Euler's equation

To multiply complex numbers, we often express the numbers using *Euler's equation*:

$$e^{i\theta} = \cos\theta \ + \ i \ \sin\theta$$

which represents a point on the unit circle in the complex plane.
Here are some examples:

$$e^{i0} = 1, \quad e^{i\pi/2} = i, \quad e^{i\pi} = -1, \quad e^{i\pi/4} = \frac{1}{\sqrt{2}}(1+i), \quad e^{i2\pi n} = 1 \ \text{ for any integer } n$$

More generally, consider what happens when we multiply two complex numbers $e^{i\theta_1}$ and $e^{i\theta_2}$. The definition of multiplication gives:

$$e^{i\theta_1} e^{i\theta_2} = e^{i(\theta_1+\theta_2)}$$

Using Euler's equation for the two terms on the left side gives:

$$(\cos\theta_1 + i \ \sin\theta_1)(\ \cos\theta_2 + i \ \sin\theta_2) \ = \ (\cos\theta_1\cos\theta_2 - \sin\theta_1\sin\theta_2) + i \ (\cos\theta_1\sin\theta_2 + \sin\theta_1\cos\theta_2)$$

Using Euler's equation for the right side gives:

$$\cos(\theta_1+\theta_2) \ + \ i \ \sin(\theta_1+\theta_2).$$

Thus,

$$\cos(\theta_1+\theta_2) = \cos\theta_1\cos\theta_2 - \sin\theta_1\sin\theta_2$$
$$\sin(\theta_1+\theta_2) = \cos\theta_1\sin\theta_2 + \sin\theta_1\cos\theta_2$$

which are familiar trig identies that you learned in Calculus.

Complex conjugate and inverse

The *complex conjugate* of $c = a + bi$ is defined

$$\overline{c} = a - bi.$$

The complex conjugate has the property that

$$c \ \overline{c} \ = \ \mid c \mid^2 \ = \ a^2 + b^2 \ .$$

In particular, $e^{-i\theta}$ is the *complex conjugate* of $e^{i\theta}$ and

$$e^{i\theta}\ e^{-i\theta}\ =\ 1\ .$$

The complex conjugate of c should not be confused with the inverse of c, namely the complex number c^{-1} which satisfies $cc^{-1} = 1$,

$$c^{-1} = \frac{1}{|c|}\overline{c}\ .$$

 ©Michael Langer

Last lecture I introduced the idea that any function defined on $x \in 0, \ldots, N-1$ could be written a sum of sines and cosines. There are two different reasons why this is useful. The first is a general one, that sines and cosines behave nicely under convolution and so we can sometimes understand better what filtering does if we understand its effects on sines and cosines. The second is more specific, that sines and cosines are a natural set of functions for describing sounds.

Today I will begin with the basic theory of Fourier analysis. This is a particular way of writing a signal as a sum of sines and cosines.

Discrete Fourier Transform

Consider 1D signals $I(x)$ which are defined on $x \in \{0, 1, ..., N-1\}$. Define the $N \times N$ *Fourier transform* matrix $\mathbf{F}$ whose k^{th} row and x^{th} column is:

$$\begin{aligned} \mathbf{F}_{k,x} &= \cos(\frac{2\pi}{N}kx) \; - \; i \; \sin(\frac{2\pi}{N}kx) \\ &\equiv e^{-i\frac{2\pi}{N}kx} \end{aligned}$$

Note that this matrix is symmetric since $e^{-i\frac{2\pi}{N}kx} = e^{-i\frac{2\pi}{N}xk}$. Also note that each row and column of the matrix $\mathbf{F}$ has a real part and an imaginary part. The real part is a sampled cosine function. The imaginary part is a sampled sine function. Note that the leftmost and rightmost column of the matrix ($x = 0$ and $x = N-1$) are not identical. You would need to go to $x = N$ to reach the same value as at $x = 0$, but $x = N$ is not represented. Similarly, the first and last row ($k = 0$ and $k = N-1$) are not identical.

Right multiplying the matrix $\mathbf{F}$ by the $N \times 1$ vector $I(x)$ gives a vector $\hat{I}(k)$

$$\hat{I}(k) \; \equiv \; \mathbf{F} \; I(x) = \sum_{x=0}^{N-1} I(x) e^{-i\frac{2\pi}{N}kx} \tag{13}$$

which is called the *discrete Fourier transform* of $I(x)$. In general, $\hat{I}(k)$ is a complex number for each k. We can write it using Euler's equation:

$$\hat{I}(k) \; = A(k) e^{i\phi(k)}$$

$|\hat{I}(k)| = A(k)$ is called the *amplitude spectrum* and $\phi(k)$ is called the *phase spectrum.*

Inverse Fourier transform

One can show (see Appendix A) that

$$\mathbf{F}^{-1} = \frac{1}{N}\overline{\mathbf{F}}$$

where $\bar{\mathbf{F}}$ is the matrix of complex conjugates of $\mathbf{F}$.

$$\overline{\mathbf{F}}_{k,x} \equiv e^{i\frac{2\pi}{N}kx}.$$

So, $\frac{1}{N}\mathbf{F}\overline{\mathbf{F}}$ is the identity matrix.

Periodicity properties of the Fourier transform

The Fourier transform definition assumed that the function was defined on $x \in 0, \ldots, N-1$, and for frequencies k in $0, \ldots, N-1$. However, sometimes we will want to be more flexible with our range of x and k.

For example, we may want to consider functions $h(x)$ that are defined on negative values of x such as the local difference function $D(x)$, the local average function $B(x)$, the Gaussian function which has mean 0, Gabor functions, etc. The point of the Fourier transform is to be able to write a function as a sum of sinuosoids. Since sine and cosine functions are defined over *all* integers, there is no reason why the Fourier transform needs to be defined only on functions that are defined on x in 0 to $N-1$.

We can define the Fourier transform of any function that is defined on a range of N consecutive values of x. For example, if we have a function defined on $-\frac{N}{2}, \cdots, -1, 0, 1, \frac{N}{2}-1$, then we can just write the Fourier transform as

$$\hat{I}(k) \;\equiv\; \mathbf{F}\; I(x) = \sum_{x=-\frac{N}{2}}^{\frac{N}{2}-1} h(x) e^{-i\frac{2\pi}{N}kx}$$

Essentially what we are doing here is treating this function $h(x)$ as periodic with period N, just like sine and cosine are, and compute the Fourier transform over a convenient sequence of N sample points. Later this lecture I will calculate the Fourier transform of $D(x)$ and $B(x)$, so look ahead to see how that is done.

The second aspect of periodicity in the Fourier transform is that $\hat{I}(k)$ is well-defined for *any* integer k (cycles per N pixels). The definition of the Fourier transform doesn't just allow k in 0 to $N-1$, but rather k can be any integer. In that case, $\hat{I}(k)$ may be considered periodic in k with period N,

$$\hat{I}(k) \;=\; \hat{I}(k + mN)$$

since, for any integer m,

$$e^{i2\pi m} = \cos(2\pi m) + i\sin(2\pi m) = 1$$

and so

$$e^{i\frac{2\pi}{N}kx} = e^{i\frac{2\pi}{N}k}\; e^{i\frac{2\pi}{N}mN} = e^{i\frac{2\pi}{N}(k+mN)x}$$

Thus, if we use frequency $k + mN$ instead of k in the definition of the Fourier transform, we get the same value.

Conjugacy property of the Fourier transform

It is a bit strange that our function $I(x)$ has N points and we will write it in terms of $2N$ functions, namely N cosines and N sines. I mentioned this point last lecture as well, and showed that indeed only N functions are needed, namely $\frac{N}{2}+1$ cosines and $\frac{N}{2}-1$ sines. This suggests that there is a redundancy in $\hat{I}(k)$ values. The redundancy is that $\cos(\frac{2\pi}{N}kx) = \cos(\frac{2\pi}{N}(N-k)x)$ and so taking the inner product with $I(x)$ will give the same value for frequency k as $N-k$. Similarly, $\sin(\frac{2\pi}{N}kx) = -\sin(\frac{2\pi}{N}(N-k)x)$ and so taking the inner product of $I(x)$ with these two functions will give the same value but with opposite sign.

 ©Michael Langer

Conjugacy property: If $I(x)$ is a real valued function, then

$$\overline{\hat{I}(k)} = \hat{I}(N-k).$$

The property does not apply if $I(x)$ has imaginary components. We will see an example later, namely if we take the Fourier transform of $e^{i\frac{2\pi}{N}k_0 x}$, for some fixed frequency k_0

For the proof of the Conjugacy Property, see Appendix B.

Linear Filtering

The visual and auditory systems analyze signals by *filtering* them into bands (ranges of different frequencies) of sines and cosines. The idea of a filter should be intuitive to you. You can imagine having a large bag of rocks and wanting to sort the rocks into ranges of different sizes. You could first pass the rocks through a fine mesh that has small holes only, so only the small rocks would pass through. Then take the bigger rocks that didn't pass through, and pass them through a mesh filter that has slightly larger holes so that now the medium size rocks pass through, but not the large rocks. This would give you three sets of rocks of a different range of sizes.

You are also intuitively familiar with filtering from color vision where the L, M, and S receptors selectively absorb the incoming light by wavelength[32]. There is some frequency overlap in the sensitivity functions, so we don't have a perfect separation of frequency bands by the three cones.

The figure below shows a more concrete example of the filtering that we will be considering. Here we have 1D signal in the upper left panel. We can write this signal as a *sum* of signals that have different ranges of frequencies. In this example, the original signal is exactly the sum of the other five signals. We will see shortly how this can be done.

Convolution Theorem

A very useful property of the Fourier transform is the *Convolution Theorem*: for any two functions $I(x)$ and $h(x)$ that are defined on 0 to $N-1$,

$$\mathbf{F}(I(x) * h(x))) = \mathbf{F}I(x)\ \mathbf{F}h(x)\ =\ \hat{I}(k)\ \hat{h}(k).$$

For the proof see Appendix C.

To prove this theorem, we need to deal with a similar issue that we mentioned before that the functions might be defined on values of x other than 0 to $N-1$. We do so by assuming the functions are periodic *i.e.* $I(x) = I(x+mN)$ and $h(x) = h(x+mN)$ for any integer m and we define the summation from 0 to $N-1$.

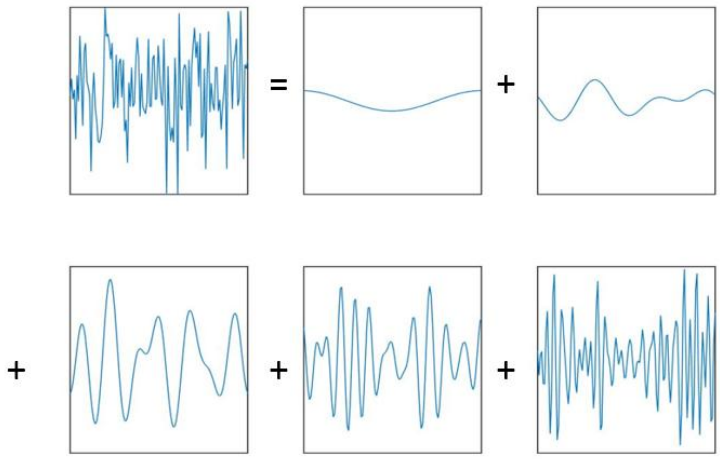
Filtering and bandwidth

Suppose we convolve an image $I(x)$ with a function $h(x)$. We have referred to $h(x)$ as an impulse response function. $h(x)$ is also called a *linear filter*. Recall that the Fourier transform of the filter $h(x)$ can be written

$$\hat{h}(k)\ =\ |\hat{h}(k)|\ e^{i\phi(k)}$$

[32] or frequency i.e. since light travels at a constant speed (called c), we can equivalently describe the sensitivity of L, M, and S cones to frequency (either spatial frequency λ or temporal frequency ω, where $c = \omega\lambda$.

 ©Michael Langer



where $|\hat{h}(k)|$ is called the *amplitude spectrum* and $\phi(k)$ is called the *phase spectrum*. By the convolution theorem,

$$\mathbf{F}I(x) = \mathbf{F}(I(x) * h(x)) = \hat{I}(k)\ |\hat{h}(k)|\ e^{i\phi(k)}$$

and $|\hat{h}(k)|$ amplifies or attenuates the frequency component amplitude $|\hat{I}(k)|$ and the phase $\phi(k)$ of the filter shifts each frequency component.

We can characterize filters by how they affect different frequencies. We will concern ourselves mainly with the amplitude spectrum for now. Let's first address the case of "ideal" filters. We say:

- $h(x)$ is an ideal *low pass filter* if there exists a frequency k_0 such that

$$\hat{h}(k) = \begin{cases} 1, & 0 \leq k \leq k_0 \\ 0, & k_0 < k \leq \frac{N}{2} \end{cases}$$

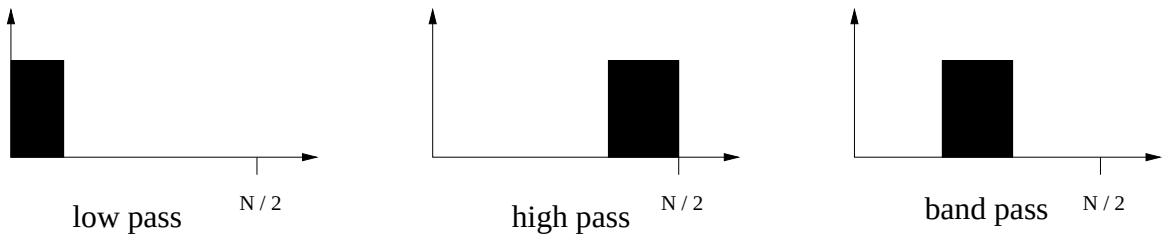
- $h(x)$ is an ideal *high pass filter* if there exists k_0 such that

$$\hat{h}(k) = \begin{cases} 0, & 0 \leq k < k_0 \\ 1, & k_0 \leq k \leq \frac{N}{2} \end{cases}$$

- $h(x)$ is an ideal *bandpass filter* if there exists two frequencies k_0 and k_1 such that

$$\hat{h}(k) = \begin{cases} 0, & 0 \leq k < k_0 \\ 1, & k_0 \leq k \leq k_1 \\ 0, & k_1 < k \leq \frac{N}{2} \end{cases}$$

Note that these definitions above only concern $k \in \{0, \ldots, \frac{N}{2}\}$. Frequencies $k < 0$ and frequencies $k > \frac{N}{2}$ are ignored in the definition because the values of $\hat{h}(k)$ of these frequencies are determined by the conjugacy and periodicity properties.



Non-ideal filters and bandwidth

We typically work with filters that are not ideal i.e. filters that only approximately satisfy the above definitions. If we have an approximately bandpass filter, then we would like to describe the width of this filter i.e. the range of frequencies that it lets through. One often does this by considering the frequencies at which $|\hat{h}(k)|$ reaches *half* its maximum value. The *bandwidth at half-height* is defined to be $k_1 - k_0$, where $k_0 < k_1$ and

$$|\hat{h}(k_0)| = |\hat{h}(k_1)| = \frac{1}{2} \max_{k \in [0, \frac{N}{2}]} |\hat{h}(k)|$$

Bandwidth can also be defined in terms of the *ratio* of k_1 to k_0, specifically, the *octave bandwidth* at half height is:

$$\log_2(\frac{k_1}{k_0}) = \log_2(k_1) - \log_2(k_0)$$

For example, a filter with a bandwidth of one octave means that the k_1 frequency is twice the k_0 frequency.

Examples of filters and their Fourier transforms

Let's look at some examples, starting with an impulse function, and the local difference and local average. Some of our calculations of Fourier transforms below will use Euler's formula, $e^{i\theta} = \cos(\theta) + i\sin(\theta)$. In particular, you can verify for yourselves that:

$$\cos(\theta) = \frac{1}{2}(e^{i\theta} + e^{-i\theta})$$

$$i\sin(\theta) = \frac{1}{2}(e^{i\theta} - e^{-i\theta})$$

We will often take $\theta = \frac{2\pi}{N} kx$.

Example 1: Impulse function

Recall

$$\delta(x) \equiv \begin{cases} 1, & x = 0 \\ 0, & \text{otherwise} \end{cases}$$

Its Fourier transform is

$$\begin{aligned}\hat{\delta}(k) &= \sum_{x=0}^{N-1} \delta(x) e^{\,-i\,(\frac{2\pi}{N}kx)} \\ &= 1 \cdot e^{i\frac{2\pi}{N}k\cdot 0} \\ &= 1\end{aligned}$$

This is rather surprising. It says that an impulse function can be written as sum of cosine functions over all frequencies $k \in 0, 1, \ldots, N-1$ and dividing by N, i.e.

$$\delta(x) = \frac{1}{N} \sum_{x=0}^{N-1} \hat{\delta}(k) e^{\,i\,(\frac{2\pi}{N}kx)}$$

Note that I write cosine functions, rather than cosine and sine functions, since $\hat{\delta}(k) = 1$ and so the phase is 0, i.e. $\phi(k) = 0$ for all k, i.e. purely real, and so there are no sine (imaginary) components. Basically, what happens is that all the cosine functions have the value 1 at $x = 0$, whereas at other values of x there are a range of values, some positive and some negative, and these other values cancel each other out when you take the sum.

To try to illustrate what is going on here, I have written a Matlab script

`http://www.cim.mcgill.ca/~langer/546/MATLAB/sumOfSinusoids.m`

which shows what happens when you add up all the cosines (top) and sines (bottom) of frequency $k = 0, \ldots, N-1$ for some chosen N.

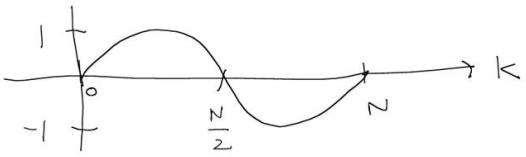
Example 2: local difference

Recall the local difference function $D(x)$ from last lecture. It has value -$\frac{1}{2}$ at $x = 1$ and value $\frac{1}{2}$ at $x = -1$. Let's compute its Fourier transform.

$$\begin{aligned}\hat{D}(k) &= \sum_x D(x) e^{-i\,(\frac{2\pi}{N}kx)} \\ &= \frac{1}{2}(-1 \cdot e^{-i\,\frac{2\pi}{N}k} + 1 \cdot e^{-i\,(\frac{2\pi}{N}k(-1))}) \\ &= \frac{1}{2}(-e^{-i\,\frac{2\pi}{N}k} + e^{i\,\frac{2\pi}{N}k}) \\ &= i\sin(\frac{2\pi}{N}k)\end{aligned}$$

The sketch below shows $sin(\frac{2\pi}{N}k)$.

The amplitude spectrum is $\mid \sin(\frac{2\pi}{N}k) \mid$. For the phase spectrum $\phi(k)$, notice that $\hat{D}(k)$ is purely imaginary. Thus $e^{i\phi(k)}$ is either i or $-i$. So the phase $\phi(k)$ is either $\frac{\pi}{2}$ or $-\frac{\pi}{2}$. In particular, $\phi(k) = \frac{\pi}{2}$ for $k \in (0, \frac{N}{2})$ and $\phi(k) = -\frac{\pi}{2}$ for $k \in (\frac{N}{2}, N)$.



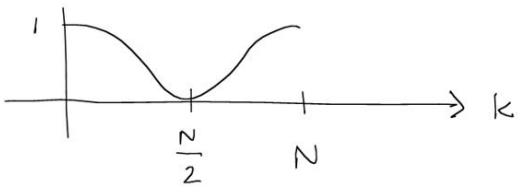
Example 3: local average

$$B(x) = \begin{cases} \frac{1}{2}, & x = 0 \\ \frac{1}{4}, & x = -1 \\ \frac{1}{4}, & x = 1 \\ 0, & \text{otherwise} \end{cases}$$

Taking its Fourier transform,

$$\begin{aligned} \mathbf{F}\ B(x) &= \frac{1}{2} + \frac{1}{4}(e^{-i\frac{2\pi}{N}k} + e^{-i\frac{2\pi}{N}k(-1)}) \\ &= \frac{1}{2} + \frac{1}{4}(e^{-i\frac{2\pi}{N}k} + e^{i\frac{2\pi}{N}k}), \\ &= \frac{1}{2}(1 + \cos(\frac{2\pi}{N}k)) \end{aligned}$$

Notice that $\hat{B}(k)$ is real, i.e. it has no imaginary component. Moreover it is non-negative. Thus, the phase spectrum $\phi(k)$ is 0.



Example 4: the "complex exponential"

Let $h(x) = e^{\,i\frac{2\pi}{N}k_0 x}$ for some integer frequency k_0. Then,

$$\mathbf{F}\ e^{\,i\frac{2\pi}{N}k_0 x} = N\delta(k - k_0).$$

See the Appendix A for a proof.

Is this result surprising. In hindsight, no. Taking the Fourier transform of a function amounts to finding out what are the coefficients on the complex exponentials $e^{\,i\frac{2\pi}{N}kx}$ for various k such that you can add these complex exponentials up and get the function. But if the function itself *is* a single complex exponential, then there is just one non-zero complex exponential needed!

We will use this result below when we compute the Fourier transforms of a cosine and sine function.

Example 5: constant function $h(x) = 1$

This is just a special case of the last example, namely if we take $k_0 = 0$. In this case,

$$\hat{h}(k) = N\ \delta(k).$$

Thus, the Fourier transform of the constant function $h(x) = 1$ is a delta function *in the frequency domain*, namely it has value N at $k = 0$ and has value 0 for all values of k in $1, \ldots, N-1$.

Examples 6 and 7: cosine and sine

We use Euler's formula to rewrite cosine and sine as a sum of complex exponentials.

$$\begin{aligned}
\mathbf{F}\cos(\frac{2\pi}{N}k_0 x) &= \sum_{x=0}^{N-1} \cos(\frac{2\pi}{N}k_0 x) e^{-i\ (\frac{2\pi}{N}kx)} \\
&= \sum_{x=0}^{N-1} \frac{1}{2}(e^{i\frac{2\pi}{N}k_0 x} + e^{-i\frac{2\pi}{N}k_0 x}) e^{-i\ \frac{2\pi}{N}kx} \\
&= \frac{N}{2}(\delta(k_0 - k) + \delta(k_0 + k))
\end{aligned}$$

$$\begin{aligned}
\mathbf{F}\sin(\frac{2\pi}{N}k_0 x) &= \sum_{x=0}^{N-1} \sin(\frac{2\pi}{N}k_0 x) e^{-i\ (\frac{2\pi}{N}kx)} \\
&= \sum_{x=0}^{N-1} \frac{1}{2i}(e^{i\frac{2\pi}{N}k_0 x} - e^{-i\frac{2\pi}{N}k_0 x}) e^{-i\ \frac{2\pi}{N}kx} \\
&= -\frac{Ni}{2}(\delta(k_0 - k) - \delta(k_0 + k))
\end{aligned}$$

Example 7: Gaussian

If we sample a Gaussian function

$$G(x, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}$$

on integer values of x, and take the Fourier transform, we get the following approximation:

$$\hat{G}(k, \sigma) \approx e^{-\frac{1}{2}(\frac{2\pi}{N})^2 \sigma^2 k^2}$$

This approximation becomes exact in the limit as $N, \sigma \to \infty$, with $\frac{\sigma}{N}$ held constant. (This amounts to taking the continuous instead of discrete Fourier transform. The proof of these claims are beyond the scope of this course.)

If you wish to see this approximation for yourself, run the Matlab script

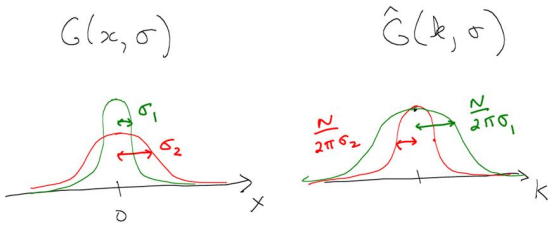
`http://www.cim.mcgill.ca/~langer/546/MATLAB/plotFourierTransformGaussian.m`

which generates the figure

`http://www.cim.mcgill.ca/~langer/546/MATLAB/plotFourierTransformGaussian.jpg`

A few key properties to notice are:

- If the standard deviation of the Gaussian in the space (x) domain is σ then the standard deviation of the Gaussian in the frequency (k) domain is proportional to $\frac{1}{\sigma}$
- $\hat{G}(k, \sigma)$ has a Gaussian shape, but it does not integrate to 1, namely there is no scaling factor present. The max value occurs at $k = 0$ and the max value is always 1.
- The Fourier transform is periodic, with period N. This is always true.



Example 8: Gabor

To compute the Fourier transform of Gabor, we use a property which is similar to the convolution theorem:

$$\mathbf{F}(I(x)h(x))) = \frac{1}{N}\mathbf{F}I(x) * \mathbf{F}h(x) \ .$$

See Appendix B for a proof, if you are interested (not on exam).

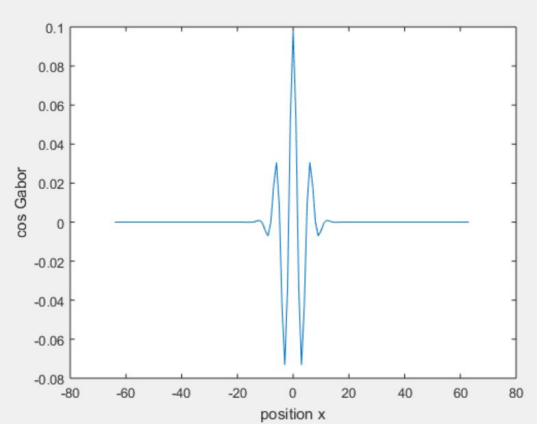
Thus the Fourier transform of a cosine Gabor is the convolution in the frequency domain of the Fourier transforms of a Gaussian and the Fourier transform of a cosine:

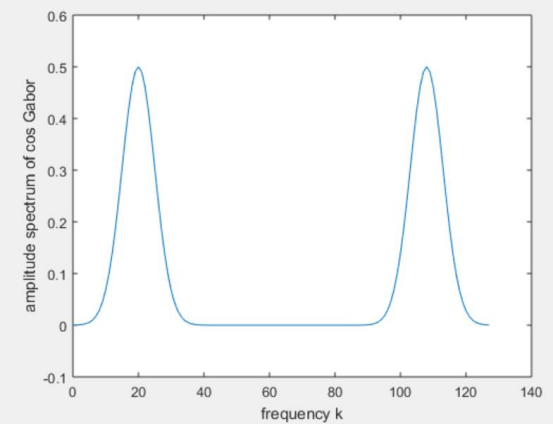
$$\begin{aligned} \mathbf{F}\ cosGabor(x, k_0, \sigma) &= \mathbf{F}\ \{\ G(x, \sigma)\cos(\frac{2\pi}{N}k_0 x)\ \} \\ &= \frac{1}{N}\ e^{-\frac{1}{2}(\frac{2\pi\sigma\ k}{N}\ )^2} * \frac{N}{2}(\delta(k_0 - k) + \delta(k_0 + k)) \\ &= \frac{1}{2}\{e^{-\frac{1}{2}(\frac{2\pi\sigma\ (k-k_0)}{N}\ )^2} + e^{-\frac{1}{2}(\frac{2\pi\sigma\ (k+k_0)}{N}\ )^2}\ \} \end{aligned}$$

which is the sum of two Gaussians, centered at $k = \pm k_0$.

[ADDED: April 12, 2018]

An example is shown below which was computed using Matlab. The cosine Gabor is defined on a vector of size $N = 128$ and has a central frequency of 20 cycles and a Gaussian with a standard deviation of 5. The amplitude spectrum has a peak at $k_0 = \pm 20$. In the amplitude spectrum plot on the right below, I plot the frequency range from 0 to $N-1$ instead of $-\frac{N}{2}$ to $\frac{N}{2}+1$. The Fourier transform of a sine Gabor can be calculated similarly. (See Exercises.)





The convolution theorem tells us that convolving a function $I(x)$ with a cosine (or sine) Gabor will give you a function that has only a band of frequencies remaining, namely the frequencies near the center frequency k_0 of the Gabor. The width of the band depends on the σ of the Gaussian of the Gabor. We will return to this idea of filtering a signal into bands of different frequencies (different Gabor filters can be used, or other bandpass filteres) when we discuss audition.

 ©Michael Langer

Appendix A

We will use the following claim to show what the inverse Fourier transform is (bottom of page).

Claim (Example 4): For any frequency k_0,

$$\mathbf{F}\ e^{\ i\frac{2\pi}{N}k_0x}\ =\ N\delta(k-k_0).$$

That is,

$$\sum_{x=0}^{N-1} e^{\ i\ \frac{2\pi}{N}k_0x}\ \ e^{-i\frac{2\pi}{N}kx} = \begin{cases} N, & k=k_0 \\ 0, & k\neq k_0 \end{cases}$$

Note that this claim essentially is essentially equivalent to saying that two cosine (or sine) functions of different frequencies are orthogonal; their inner product is 0.

Proof: Rewrite the left side of the above summation as

$$\sum_{x=0}^{N-1} e^{\ i\ \frac{2\pi}{N}(k_0-k)x}\ \ . \tag{14}$$

If $k=k_0$, then the exponent is 0 and so we are just summing $e^0=1$ and the result is N.

That doesn't yet give us the result of the claim, because we still need to show that the summation is 0 when $k\neq k_0$. So, for the case $k\neq k_0$, observe that the summation is a finite geometric series and thus we can use the following identity which you know from Calculus:[33]: let γ be any number (real or complex) then

$$\sum_{x=0}^{N-1}\gamma^x = \frac{1-\gamma^N}{1-\gamma}.$$

Applying this identity for our case, namely $\gamma = e^{\ i\ \frac{2\pi}{N}(k-k_0)}$, lets us write (14) as

$$\sum_{x=0}^{N-1} e^{\ i\ \frac{2\pi}{N}(k-k_0)x} = \frac{1-e^{\ i\ 2\pi(k-k_0)}}{1-e^{\ i\ \frac{2\pi}{N}(k-k_0)}}\ \ . \tag{15}$$

The numerator on the right hand side vanishes because $k-k_0$ is an integer and so

$$e^{\ i2\pi(k-k_0)}\ =1\ .$$

What about the denominator? Since k and k_0 are both in $0,\ldots,N-1$ and since we are considering the case that $k\neq k_0$, we know that $|k-k_0|<N$ and so $e^{\ i\ \frac{2\pi}{N}(k-k_0)}\neq 1$. Hence the denominator does not vanish. Since the numerator is 0 but the denominator is not 0, we can conclude that the right side of Eq. (15) is 0. Thus, the summation of (14) is 0, and so $\mathbf{F}\ e^{\ i\frac{2\pi}{N}k_0x}=0$ when $k\neq k_0$. This completes the derivation for the case $k\neq k_0$.

Claim (inverse Fourier transform): $\mathbf{F}^{-1}=\frac{1}{N}\overline{\mathbf{F}}$

Proof:

The matrix $\frac{1}{N}\mathbf{F}\bar{\mathbf{F}}$ is $N\times N$. The above example says that row k_0 and column k of this matrix is 1 when $k_0=k$ and 0 when $k_0\neq k$ and hence this matrix is the unit diagonal.

[33] If you are unsure where this comes from, see equations (1)-(6) of `http://mathworld.wolfram.com/GeometricSeries.html`.

Appendix B: Conjugacy property of the Fourier transform

Claim: If $I(x)$ is a real valued function, then

$$\overline{\hat{I}(k)} = \hat{I}(N-k).$$

Proof: (not on final exam)

$$\begin{aligned}
\hat{I}(N-k) &= \sum_{x=0}^{N-1} I(x) e^{-i\frac{2\pi}{N}\ (N-k)x} \\
&= \sum_{x=0}^{N-1} I(x)\ e^{-i\ 2\pi x}\ e^{i\ \frac{2\pi}{N}\ kx} \\
&= \sum_{x=0}^{N-1} I(x)\ e^{i\ \frac{2\pi}{N}\ kx}, \quad \text{since } e^{i2\pi x}\ = 1 \quad \text{for any integer } x \\
&= \sum_{x=0}^{N-1} I(x)\ \overline{e^{-i\frac{2\pi}{N}\ kx}} \\
&= \sum_{x=0}^{N-1} \overline{I(x)}\ \overline{e^{-i\frac{2\pi}{N}\ kx}} \quad \text{, if } I(x) \text{ is real} \\
&= \sum_{x=0}^{N-1} \overline{I(x)\ e^{-i\frac{2\pi}{N}\ kx}} \\
&= \overline{\hat{I}(k)}
\end{aligned}$$

Appendix C: Convolution Theorem

Claim: For any two functions $I(x)$ and $h(x)$ that are defined on N consecutive samples e.g. 0 to $N-1$,

$$\mathbf{F}(I(x) * h(x))) = \mathbf{F}I(x)\ \mathbf{F}h(x)\ =\ \hat{I}(k)\ \hat{h}(k).$$

Proof: (not on final exam)

$$\begin{aligned}
\mathbf{F}\quad I * h(x) &= \sum_{x=0}^{N-1} e^{-i\frac{2\pi}{N}kx} \sum_{x'=0}^{N-1} I(x-x')h(x'), \quad \text{by definition} \\
&= \sum_{x'=0}^{N-1} h(x') \sum_{x=0}^{N-1} e^{-i\frac{2\pi}{N}kx}\ \ I(x-x'), \quad \text{by switching order of sums} \\
&= \sum_{x'=0}^{N-1} h(x') \sum_{u=0}^{N-1} e^{-i\frac{2\pi}{N}k(u+x')} I(u)\ , \quad \text{where } u = x - x' \\
&= \sum_{x'=0}^{N-1} h(x')\ e^{-i\frac{2\pi}{N}\ kx'} \sum_{u=0}^{N-1} e^{-i\frac{2\pi}{N}ku}\ I(u) \\
&= \hat{h}(k)\ \hat{I}(k)
\end{aligned}$$

Appendix D (another Convolution Theorem)

We will often work with filters such as Gabor functions that are the product of two functions. Suppose we have two 1D functions $I(x)$ and $h(x)$ and we take their product. What can we say about the Fourier transform? The answer is similar to the convolution theorem, and indeed is just another version of that theorem:

$$\mathbf{F}\ (I(x)h(x)) \ = \ \frac{1}{N}\ \hat{I}(k) * \hat{h}(k)$$

or, in words, the Fourier transform of the product of two functions is the convolution of the Fourier transforms of the two functions. Note that the convolution on the right hand side is between two complex valued functions, rather than real valued functions. But the same definition of convolution applies.

To prove the above property, we take the inverse Fourier transform of the right side and show that it gives $I(x)h(x)$. Note that the summations and functions below are defined on frequencies k, k', k'' mod N, since the Fourier transform of a function has period N.

$$\begin{aligned}
\mathbf{F}^{-1}\ \ \hat{I}(k) * \hat{h}(k) \ &= \ \frac{1}{N}\sum_{k=0}^{N-1} e^{i\frac{2\pi}{N}kx}\ \ \sum_{k'=0}^{N-1} \hat{h}(k')\ \hat{I}(k-k') \qquad \text{...and rearrange...} \\
&= \ \frac{1}{N}\sum_{k'=0}^{N-1} \hat{h}(k')\ \ \sum_{k=0}^{N-1} e^{i\frac{2\pi}{N}kx}\ \ \hat{I}(k-k') \qquad \text{...and multiply by...} e^{i\frac{2\pi}{N}k'x}e^{-i\frac{2\pi}{N}k'x} \\
&= \ \frac{1}{N}\sum_{k'=0}^{N-1} \hat{h}(k')e^{i\frac{2\pi}{N}k'x}\ \ \sum_{k=0}^{N-1} e^{i\frac{2\pi}{N}(k-k')x}\ \ \hat{I}(k-k') \\
&= \ h(x) \sum_{k''=-k'}^{N-1-k'} e^{i\frac{2\pi}{N}(k'')x}\ \ \hat{I}(k''), \quad \text{where } k'' = k - k' \\
&= \ Nh(x)\ I(x)
\end{aligned}$$

Dividing both sides by N and we're done.

[Most of this lecture was finishing up the linear systems material from lecture 17. See the lecture 17 notes.]

Introduction to sound

A few lectures from now, we will consider problems of *spatial hearing* which I loosely define as problems of using sounds to compute where objects are in 3D space. There are two kinds of spatial hearing problems in perception. One involves sounds that are *emitted* by objects in the world, which is the spatial hearing problem we are used to. The other involves sounds that are *reflected* off objects in the world. This is the spatial hearing problem used by echolocating animals e.g bats. This year we will (probably) just have time to cover the first.

Emitted sounds are produced when forces are applied to an object that make the object vibrate (oscillate). For example, when one object hits another object, kinetic energy is transformed into potential energy by an elastic compression [34]. This elastic compression results in vibrations of the object(s) which dampen out over time depending on the material of the object. These object vibrations produce tiny pressure changes in the air surrounding the object, since as the object vibrates it bumps into the air molecules next to it. These air pressure changes then propagate as waves into the surrounding air.

There are three general factors that determine a sounds. The first is the energy source which drives some object into oscillatory behavior. The second is the object that is oscillating – its shape and material properties. The third factor is the space into which or through which the sound is emitted. This space cavity can attenuate or enhance certain frequencies through resonance. The varying shapes of musical instruments are an obvious example. Voice is another, and we will return to both.

Emitted sounds are important for hearing. They inform us about events occurring around us, such as footsteps, a person talking, an approaching vehicle, etc. Here we have a major difference between hearing and vision. Nearly all the visible surfaces reflect light rather than emit light.[35] However, light sources themselves are generally not informative for vision but rather their 'role' is to illuminate other objects, that is, shining light on other objects so that the visual system can use this reflected light to estimate 3D scene properties or recognize the object by the spatial configuration of the light patterns. By contrast, emitting sound sources *are* informative. They tell us about the location of objects and their material properties.

What about reflected sounds? To what extent are they useful? Blind people seem to use reflected sounds (echos) to navigate. Blind people can walk through an environment without bumping into walls and to do so they use the echos of their footsteps and the echos of the tapping of their cane. They hear the reflections of these sounds off walls and other obstancles. People like us that have normal vision probably use reflected sound also but not nearly as much as blind people. This hasn't been studied much.

[34] Another component of the kinetic energy is transformed into non-elastic mechanical energy, namely a permanent shape change. This happens when the object cracks, chips, breaks or is dented

[35] Typically only hot objects emit light. Electronic displays/lights e.g. LEDs are obvious exceptions to this statement. Not only are they non-hot light emitters, but the light patterns they emit are often meant to be informative – perhaps the light pattern you are seeing right now.

 ©Michael Langer

Pressure vs. intensity

Sound is a set of air pressure waves that are measured by the ear. Air pressure is always positive. It oscillates about some mean value I_a which we call *atmospheric pressure*. The units of air pressure are atmospheres and the mean air pressure around us is approximately "one atmosphere".

Sounds are small variations in air pressure about this mean value I_a. These pressure variations can be either positive (compression) or negative (rarefaction). We will treat the pressure at a point in 3D space as a function of time,

$$P(X, Y, Z, t) \;=\; I_a + I(X, Y, Z, t)$$

where $I(X, Y, Z, t)$ is small compared to atmospheric pressure I_a. Note that we are using "big" X, Y, Z rather than little x, y, z, since we are talking about points in 3D space.

I emphasize that sounds are quite small perturbations of the atmospheric pressure. The quietest sound that we can hear is a perturbation of 10^{-9} atmospheres. The loudest sound that we can tolerate without pain is 10^{-3} atmospheres. Thus, we are sensitive to 6 orders of magnitude of pressure changes. (An *order of magnitude* is a factor of 10.)

Loudness: SPL and dB

To refer to the loudness of a sound, one can refer either to *sound pressure* $I(X, Y, Z, t)$ or to the square of this value, which I will loosely refer to as *intensity*. Sound pressure oscillates about 0, whereas intensity is of course always positive. Think of intensity as the energy energy per unit volume, namely, the work done to compress or expand a unit volume of air to produce the deviation from I_a.

When describing the loudness of a sound, we typically don't care about the instantaneous pressure or intensity, but rather the average over some time. Averaging the sound pressure $I(X, Y, Z, t)$ over time makes no sense, since the average is I_a which has nothing to do with the loudness of a sound. Instead one averages the intensity over some time T. Let I be the root mean square of the sound pressure:

$$I \equiv \sqrt{\frac{1}{T} \sum_{t=1}^{T} I(t)^2}$$

The loudness will then be defined in terms of the ratio of the RMS sound pressure to the RMS sound pressure of some standard I_0 which is very quiet sound called the "threshold of hearing". I_0 is a specific extremely soft sound level which has been determined in careful experiments in acoustically isolated rooms. It is the quietest sound below which one cannot discriminate.

The range of sound pressures that we can hear comfortably is very large. Moreover, psychophysical studies have shown that people are sensitive to ratios of RMS sound pressures rather than differences. For these reasons, one defines the loudness by the log of this ratio:

$$\text{Bels} = \log_{10} \frac{I^2}{I_0^2} = 2 \log_{10} \frac{I}{I_0}$$

It is common to use a slightly different unit, namely ten times Bels. That is, the common units for defining the loudness of a sound is the *sound pressure level* (SPL) in SPL in decibels (dB):

$$10 \log_{10} \frac{I^2}{I_0^2} \;=\; 20 \;\log_{10} \;|\frac{I}{I_0}|$$

 ©Michael Langer

Multiplying by a factor 10 is convenient because the human auditory system is limited in its ability to discriminate sounds of different loudnesses, such that we can just discriminate sounds that different from each other by about $\frac{1}{10}$ of a Bel, or 1 dB. So, we can think of 1 dB as a just noticable difference (JND).

As an example, suppose you were to double the RMS sound pressure from I to $2I$. What would be the increase in SPL (dB) ? To answer this, note

$$20 \log_{10} \frac{2I}{I_0} = 20 \log_{10} 2 + 20 \log_{10} \frac{I}{I_0}$$

So the increase in dB is $20 \log_{10} 2 \approx 6$ dB.

Here are a few examples of sound pressure levels:

Sound	dB
jet plane taking off (60 m)	120
noisy traffic	90
conversation (1 m)	60
middle of night quiet	30
recording studio	10
threshold of hearing	0

 ©Michael Langer

Sound waves

Last lecture we considered sound to be a pressure function $I(X, Y, Z, t)$. However, sound is not just any function of those four variables. Rather, sound obeys the *wave equation*:

$$\frac{\partial^2 I(X,Y,Z,t)}{\partial X^2} + \frac{\partial^2 I(X,Y,Z,t)}{\partial Y^2} + \frac{\partial^2 I(X,Y,Z,t)}{\partial Z^2} = \frac{1}{v^2}\frac{\partial^2 I(X,Y,Z,t)}{\partial t^2}$$

where v is the speed of sound. This equation says that if you take a snapshot of the pressure function at any time t, then the spatial derivatives the pressure function at each point XYZ tell you how the pressure at the point will change as time varies. Note that this equation contains the consant v which is the speed of sound.

The speed of sound in air is about $v = 340$ meters per second, or 34 cm per millisecond. This is quite slow. (If you go to a baseball game and you sit behind the outfield fence over 100 m away, you can easily perceive the delay between when you *see* the ball hit the bat, and when you *hear* the ball hit the bat.) Amazingly, the speed of sound is so slow that our brains can detect differences in the arrival times of sounds at the left and right ear, and we use this difference to help us perceive where sound sources are. (We'll discuss this in the following few lectures.)

Also notice that the wave equation is linear in $I(X, Y, Z, t)$. If you have several sources of sound, then the pressure function I that results is identical to the sum of the pressure functions produced by the individual sources in isolation.

Today we will examine two types of sounds that are of great interest: music and speech. We will see how a frequency domain analysis is fundamental to both.

Musical sounds

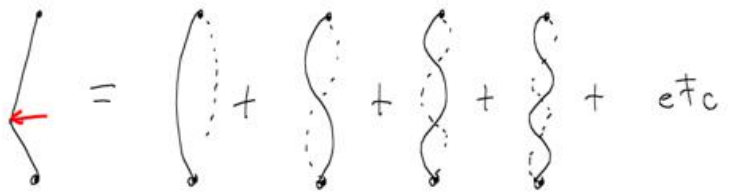
Let's begin by briefly considering string instruments such as guitars. First consider the vibrating string. When we pluck the guitar string, we are setting its initial shape to something different than its resting state. This initial shape and the subsequent shape as it vibrates always has fixed end points. The initial shape can be written as a sum of sine functions, specifically sine functions with value zero at the end points. This is summation is similar a Fourier transform, but here we only need sine functions (not sines and cosines), in particular,

$$\sin(\frac{\pi}{L}xm)$$

where $m \geq 0$ is an integer and L is the length of the guitar string. We have π rather than 2π in the numerator since the sine value is zero when $x = \frac{L}{m}$ for any m.

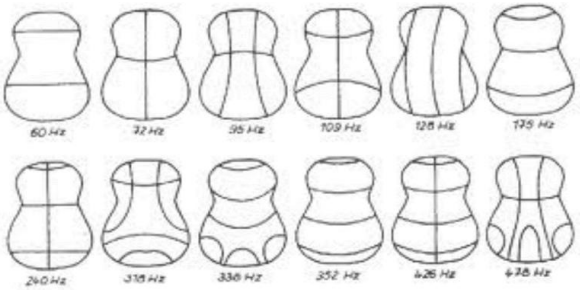
Physics tells us that if a string is of length L then its mode $\sin(\frac{\pi}{L}x)$ vibrates at a temporal frequency $\omega = \frac{c}{L}$ where c is a constant that depends on the properties of the string such as its material, thickness, tension. Think of each mode m of vibration as dividing the string into equal size parts of size $\frac{L}{m}$. For example, we would have four parts of length $\frac{L}{4}$. (See sketch in slide). You can think of each of these parts as being little strings with fixed endpoints.

Frequency m is called the *m-th harmonic*. The frequency $\omega_0 = \frac{c}{L}$ i.e. $m = 1$ is called the *fundamental* frequency. Frequencies for $m > 1$ are called *overtones*. Note harmonic frequencies have a linear progression $m\omega_0$. They are multiples of the fundamental.



Note that the *definition* of harmonic frequencies is that they are an integer multiple of a fundamental frequency. It just happens to be the case that vibrating strings naturally produce a set of harmonic frequencies. There are other ways to get harmonic frequencies as well, for example, voiced sounds as we will see later.

For stringed instruments such as a guitar, most of the sound that you hear comes not from the vibrating strings, but rather the sound comes from the vibrations of the instrument body (neck, front and back plates) in response to the vibrating strings. The body has its own vibration modes as shown below. The curved lines in the figure are the nodal points which do not move. Unlike the string, the body modes do not define an arithmetic progression.

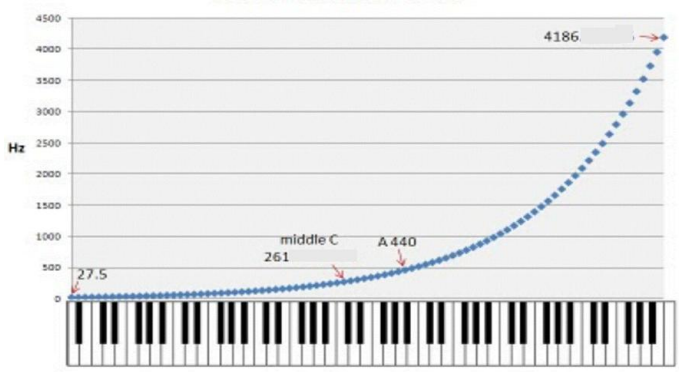


For another example, see
`http://www.acs.psu.edu/drussell/guitars/hummingbird.html`

In western music, notes have letter names and are periodic: A, B, C, D, E, F, G, A, B, C, D, E, F, G, A, B, C, D, E, F, G, etc. Each of these notes defines a fundamental frequency. The consecutive fundamental frequencies of the notes for any letter (say C) are separated by one octave. e.g. A, B, C, D, E, F, G, A covers one octave. Recall from the linear systems lecture that a difference of one octave is a doubling of the frequency, and in general two frequencies ω_1 and ω_2 are separated by $\log_2 \frac{\omega_2}{\omega_1}$ octaves.

An octave is partitioned into 12 intervals called *semitones*. The intervals are each $\frac{1}{12}$ of an octave, i.e. equal intervals on a log scale. A to B, C to D, D to E, F to G, and G to A are all two semitones, whereas B to C and E to F are each one semitone. (No, I don't know the history of that.) It follows that the number of semitones between a note with fundamental ω_1 and a note with fundamental ω_2 is $12 \log_2 \frac{\omega_2}{\omega_1}$. To put it another way, the frequency that is n semitones above ω_1 is $\omega_1 2^{\frac{n}{12}}$. The notes on a piano keyboard are shown below, along with a plot of their fundamental frequencies.

 ©Michael Langer



Notice that the frequencies of consecutive semitones define a geometric progression, whereas consecutive harmonics of a string define an arithmetic progression. When you play a note on a piano keyboard, the sound that results contains the fundamental as well as all the overtones - which form an arithmetic progression. When you play multiple notes, the sound contains the fundamentals of each note as well as the overtones of each. [ASIDE: The reason why some chords (multiple notes played together) sound better than other has to do – in part – with the distribution of the overtones of the notes, namely how well they align. Details omitted.]

Speech sounds

Human speech sounds have very particular properties. They obey certain physical constraints, namely our anatomy. Speech sounds depend on several variables. One is the shape of the *oral cavity*, which is the space inside your mouth. This shape is defined by the tongue, lips, and jaw position which are known as *articulators*. The sound wave that you hear has passed from the lungs, past the vocal cords, and through the long cavity (pharynx + oral and nasal cavity) before it exits the body. The shape of the oral cavity is determined by the position of the tongue, the jaw, the lips.

Consider the different vowel sounds in normal spoken English "aaaaaa", "eeeeeee", "iiiiiiii", "oooooo", "uuuuuuu". Make these sounds to yourself and notice how you need to move your tongue, lips, and jaw around. These variations are determined by the positioning of the articulators. Think of the vocal tract (the volume between the vocal cords and the mouth and nose) as a resonant tube, like a bottle. Changing the shape of the tube by varying the articulators causes different sound frequencies that are emitted from you to be amplified and others to be attenuated.

Voiced Sounds

Certain sounds require that your vocal cords vibrate while other sounds require that they do not vibrate. When vocal cords are tensed, the sounds that result are called *voiced*. An example is a tone produced by a singing voice. When the vocal cords are relaxed, the sounds are called *unvoiced*. An example is whispering. Normal human speech is a combination of voiced and unvoiced sounds.

Voiced sounds are formed by regular pulses of air from the vocal cords. There is an opening in the vocal cords called the *glottis*. When the vocal cords are tensed, the glottis opens and closes at a regular rate. A typical rate for glottal "pulses" for adult males and females are around 100 and 200 Hz *i.e.* about a 10 ms or 5 ms period, although this can vary a lot depending on whether one has

a deep versus average versus high voice. Moreover, each person can change their glottal frequency by varying the tension. That is what happens when you sing different notes.

Suppose you have $n_{glottal}$ glottal pulses which occur with period $T_{glottal}$ (time between pulses). The total duration would be $T = n_{glottal}T_{glottal}$ time samples. We can write the sound source pressure signal that is due to the glottal pulse train as:

$$I(t) = \sum_{j=0}^{n_{\text{glottal}}-1} g(t - jT_{glottal})$$

where $g()$ is the sound pressure due to each glottal pulse. We can write this equivalently as

$$I(t) \;=\; g(t) * \sum_{j=0}^{n_{\text{glottal}}-1} \delta(t - jT_{glottal}).$$

Each glottal pulse gets further shaped by the oral and nasal cavities. The oral cavity in particular depends on the positions of the articulators. If the articulators are fixed in place over some time interval, each glottal pulse will undergo the same waveform change in that interval. Some people speak very quickly but not so quickly that the position of the tongue, jaw and mouth changes over time scales of the order of say 10 ms. Indeed, if you could move your articulators that quickly, then your speech would not be comprehensible.

One can model the transformed glottal pulse train as a convolution with a function $a(t)$, so the final emitted sound is:

$$I(t) \;=\; a(t) * g(t) * \sum_{j=0}^{n_{\text{glottal}}-1} \delta(t - jT_{glottal})$$

So you can think of $a(t) * g(t)$ as a single impulse response function. The reason for separating them is that there really are two different things happening here. The glottal pulse $g(t)$ is not an impulse function and it is different from the effect $a(t)$ of the articulators. Each glottal pulse produces its own $a(t) * g(t)$ pressure wave and these little waves follow one after the other.

Let's next briefly consider the frequency properties of voiced sounds. If we take the Fourier transform of $I(t)$ over T time samples – and we assume the articulators are fixed in position so that we can define $a(t)$ and we assume $T_{glottal}$ is fixed over that time also – we get

$$\hat{I}(\omega) \;=\; \hat{a}(\omega)\; \hat{g}(\omega)\; \mathbf{F} \sum_{j=0}^{n_{\text{glottal}}-1} \delta(t - jT_{glottal}).$$

You can show (in Assignment 3) that

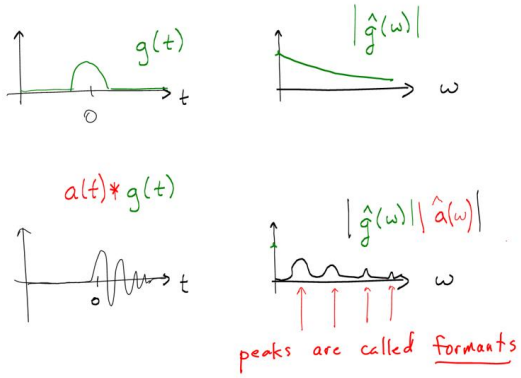
$$\mathbf{F} \sum_{j=0}^{n_{\text{glottal}}} \delta(t - jT_{glottal}) = n_{glottal} \sum_{j=0}^{T_{glottal}-1} \delta(\omega - jn_{glottal})$$

So,

$$\hat{I}(\omega) \;=\; \hat{a}(\omega)\; \hat{g}(\omega)\; n_{glottal} \sum_{j=0}^{T_{glottal}-1} \delta(\omega - jn_{glottal})$$

This means that the glottal pulses cancel out all frequencies except other than those that are a multiple of $n_{glottal} = \frac{T}{T_{glottal}}$, that is, the number glottal pulses per T samples. I emphasize here that this clean mathematical result requires that the sequence of glottal pulses spans the T samples, and the period is regular and the articulators are fixed during that interval.

Measurements show that the glottal pulse $g(t)$ is a low pass function. You can think of it as having a smooth amplitude spectrum, somewhere between a Gaussian amplitude spectrum which falls off quickly and an impulse amplitude spectrum which is constant over ω.



The effect of the articulators is to modulate the amplitude spectrum that is produced by the glottal pulses, namely by multiplying by $\hat{a}(\omega)$. This amplifies some frequencies and attenuates others. (It also produces phase shifts which we will ignore in this analysis, but which are important if one considers the wave shape of each pulse.) The peaks of the amplitude spectrum $|\hat{g}(\omega)\ \hat{a}(\omega)|$ are called *formants*. As you change the shape of your mouth and you move your jaw, you change $a(t)$ which changes the frequencies of the formants. I will mention formants again later when I discuss spectrograms.

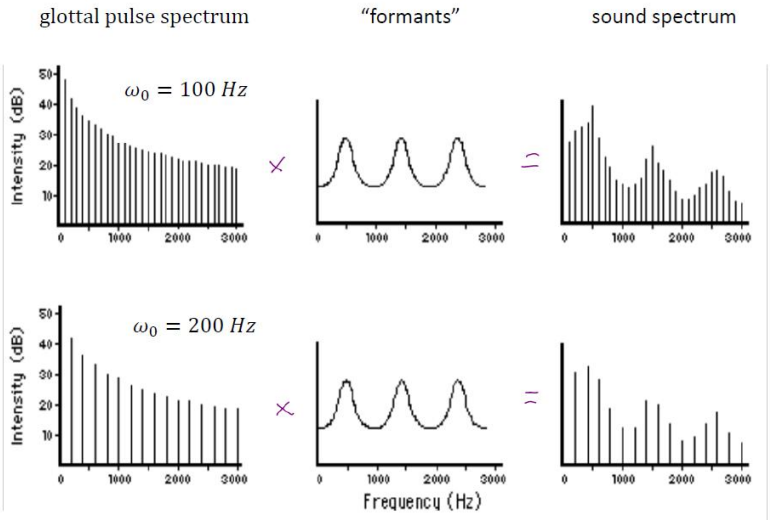
As mentioned above, the sum of delta functions nulls out frequencies except those that happen to be part of an arithmetic progression of fundamental frequency $\omega_0 = n_{glottal} = \frac{T}{T_{glottal}}$, that is, $n_{glottal}$ samples per T time steps. However, we often want to express our frequencies in cycles per second rather than cycles per T samples. The typical sampling rate used in high quality digital audio is 44,100 samples per second, or about 44 samples per ms.[36] To convert from cycles per T samples to cycles per second, one should multiply by 44,100 samples per second.

This sampling rate is not the only one that is used, though. Telephone uses a lower sampling rate, for example, since quality is less important.

The frequency 44,100 * $n_{glottal}$ is the fundamental frequency in cycles per second, which correponds the glottal pulse train. As mentinoed earlier, in adult males this is typically around 100 Hz for normal spoken voice. In adult females, it is typically around 200 Hz. In children, it is often higher than 250 Hz.

The two rows in the figure below illustrate a voiced sound with fundamental 100 and 200 Hz. The left panels shows just amplitude spectrum of the glottal pulse train. The center panels illustrate the amplitude spectrum of the articulators for several formants. The right panel shows the amplitude spectrum of the resulting sound.

[36] One often uses 16 bits for each of two channels (two speakers or two headphones).



Unvoiced sounds (whispering)

When the vocal cords are relaxed, the resulting sounds are called *unvoiced*. There are no glottal pulses. Instead, the sound wave that enters the oral cavity can be described better as noise. The changes that are produced by the articulators, etc are roughly the same in voiced versus unvoiced speech, but the sounds that are produced are quite different. You can still recognize speech when someone whispers. That's because there is still the same shaping of the different frequencies into the formants, and so the vowels are still defined. But now it is the noise that gets shaped rather than glottal pulses.

I mentioned in the lecture that the noise $n(t)$ produced by expelling air from the lungs has a flat amplitude spectrum, hat is, prior to the reshaping of the spectrum by the articulators. The sound that comes out the mouth is $n(t) * a(t)$ and that sound is shaped by the articulators.

Consonants

Another important speech sound occurs when one restricts the flow of air, and force it through a small opening. For example, consider the sound produced when the upper front teeth contact the lower lip. Compare this to when the lower front teeth are put in contact with the upper lip. (The latter is not part of English. I suggest you amuse yourself by experimenting with the sounds you can make in this way.) Compare these to when the tongue is put in contact with the front part of the palate vs. the back part of the palate.

Most *consonants* are defined this way, namely by a partial or complete blockage of air flow. There are several classes of consonants. Let's consider a few of them. For each, you should consider what is causing the blockage (lips, tongue, palate).

- fricatives (narrow constriction in vocal tract):
  - voiced: z, v, zh, th (as in *the*)
  - unvoiced: s, f, sh, th (as in θ)

- stops (temporary cessation of air flow):
  - voiced: b, d, g
  - unvoiced: p, t, k

  These are distinguished by where in the mouth the flow is cutoff. Stops are accompanied by a brief silence

- nasals (oral cavity is blocked, but nasal cavity is open)
  - voiced: m, n, ng

  You might not believe me when I tell you that nasal sounds actually come out of your nose. Try shutting your mouth, plugging your nose with your fingers, and saying "mmmmm". See what happens?

Spectrograms

When we considered voiced sounds, we took the Fourier transform over T samples and assumed that the voiced sound extended over those samples. One typically does not know in advance the duration of voiced sounds, so one has to arbitrary choose a time interval.

Often one analyzes the frequency content of a sound by partitioning $I(t)$ into blocks of B disjoint intervals each containing T samples – the total duration of the sound would be BT. For example, if $T = 512$ and the sampling rate is 44000 samples per second, then each interval would be about 12 milliseconds.

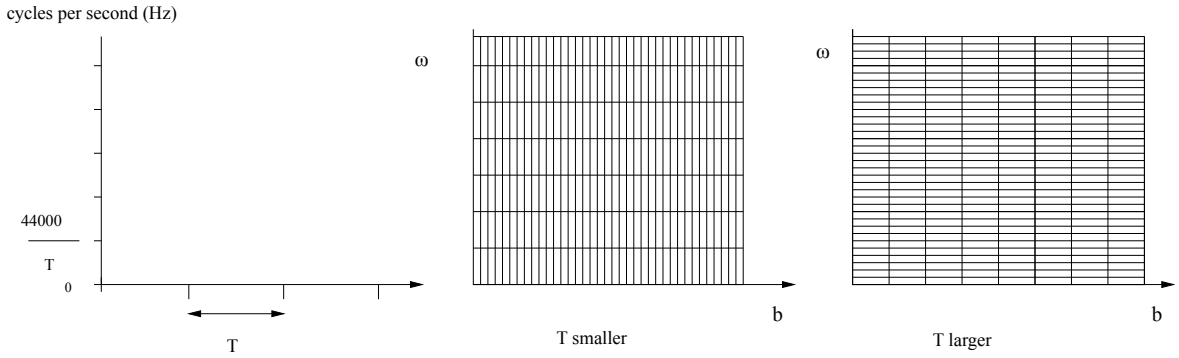
Let's compute the discrete Fourier transform on the T samples in each of these block. Let ω be the frequency variable, namely cycles per T samples, where $\omega = 0, 1, \ldots, T-1$. Consider a 2D function which is the Fourier transform of block b:

$$\hat{I}(b, \omega) \;=\; \sum_{t=0}^{T-1} I(\; b\; T + t)\; e^{-i\frac{2\pi}{T}\omega t}.$$

Typically one ignores the phase of the Fourier transform here, and so one only plots the amplitude $\mid \hat{I}(b, \omega) \mid$. You can plot such a function as a 2D "image", which is called a *spectrogram*.

The sketch in the middle shows a spectrogram with a smaller T, and the sketch on the right shows one with a larger T. The one in the middle is called a "wideband" spectrogram because each 'pixel' of the spectrogram has a wide range of frequncies, and the one on the right is called a narrowband spectrogram because each 'pixel' has a smaller range of frequencies. For example, if $T = 512$ samples, each pixel would be about 12 ms wide and the steps in ω would be 86 Hz high, whereas if $T = 2048$ samples, then each pixel would be be 48 ms wide and the ω steps would be 21 Hz.

Notice that we cannot simultaneously localize the properties of the signal in time and in frequency. If you want good frequency resolution (small ω steps), then you need to estimate the frequency components over long time intervals. Similarly, if you want good temporal resolution (i.e. when exactly does something happen?), then you can only make coarse statements about which frequencies are present "when" that event happens. This inverse relationship is similar to what we observed earlier when we discussed the Gaussian and its Fourier transform.



Examples (see slides)

The slides show a few examples of spectrograms of speech sounds, in particular, vowels. The horizontal bands of frequencies are the formants which I mentioned earlier. Each vowel sound is characterized by the relative positions of the three formants. For an adult male, the first formant (called F1) is typically centered anywhere from 200 to 800 Hz. The second formant F2 from 800 to 2400 Hz, F3 from 2000 to 3000 Hz.

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 Δr 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.

 ©Michael Langer

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 τ is the time delay, α 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 α, τ explicitly, of course. Rather it has to estimate them. We can formulate this estimation problem as finding α, τ that minimizes the sum of squared errors:

$$\sum_{t=1}^{T} \; (I_l(t) \;-\; \alpha I_r(t-\tau))^2 \;. \tag{16}$$

Intuitively, we wish to shift and scale the right ear's signal by τ 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 τ, the auditory system only needs to consider τ 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 (16) is equivalent to minimizing

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

The summations in the first two terms are over slightly different intervals because of the shift τ in the second term. However, if τ is small relative to T, then the second summation will vary little with τ. The third term in the summation is the one that depends heavily on τ, 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 τ that maximizes

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

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

 ©Michael Langer

To estimate α, one could use the model:

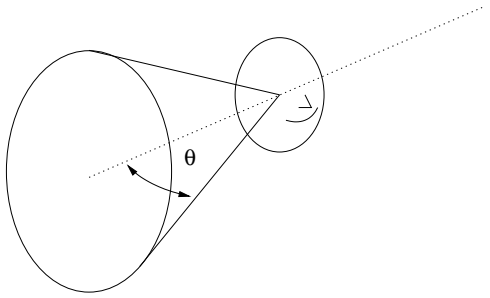
$$I_l(t) \approx \alpha I_r(t-\tau)$$

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 (θ, ϕ) 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 (θ, ϕ), 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 θ and ϕ define a spherical coordinate system, with the poles being directly above and below the head. The angle θ is the *azimuth* and goes from 0 to 360 degrees (front, left, behind, right). The angle ϕ 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 (θ, ϕ), 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 (ϕ, θ). 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 ϕ, θ. 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

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

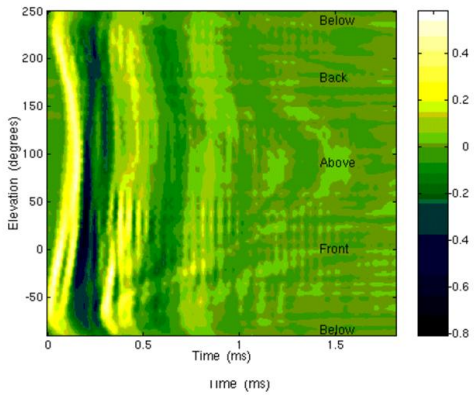
where $\theta = 0$ is the straight ahead azimuth.

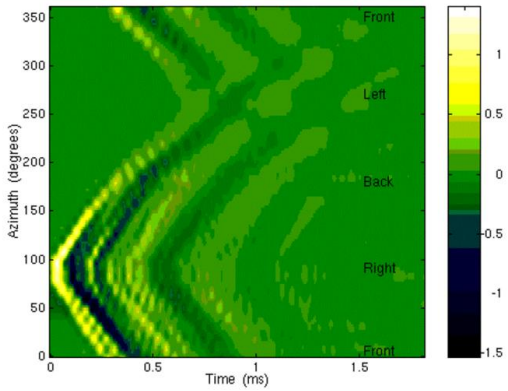
 ©Michael Langer

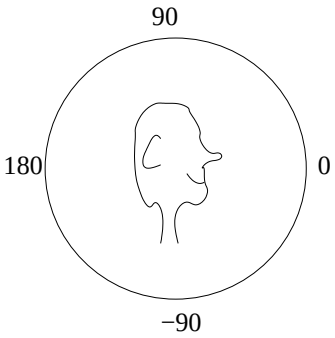
To measure the function $h(t;\phi,\theta)$ for a given person, one can place a tiny microphone inside the person's auditory canal and then record the sound produced by an impulse source function at some distance away in direction (ϕ,θ), and repeat the experiment for many different directions ϕ,θ. Another 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`

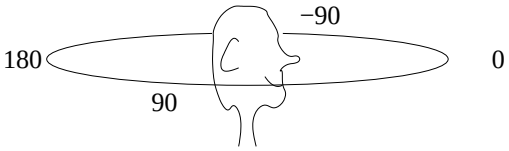
Examples of measurements using a KEMAR mannequin are shown below. Data taken from `http://interface.cipic.ucdavis.edu/`. That web site also has some nice tutorials).







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 μs (1 μs $= 10^{-6}$ 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.

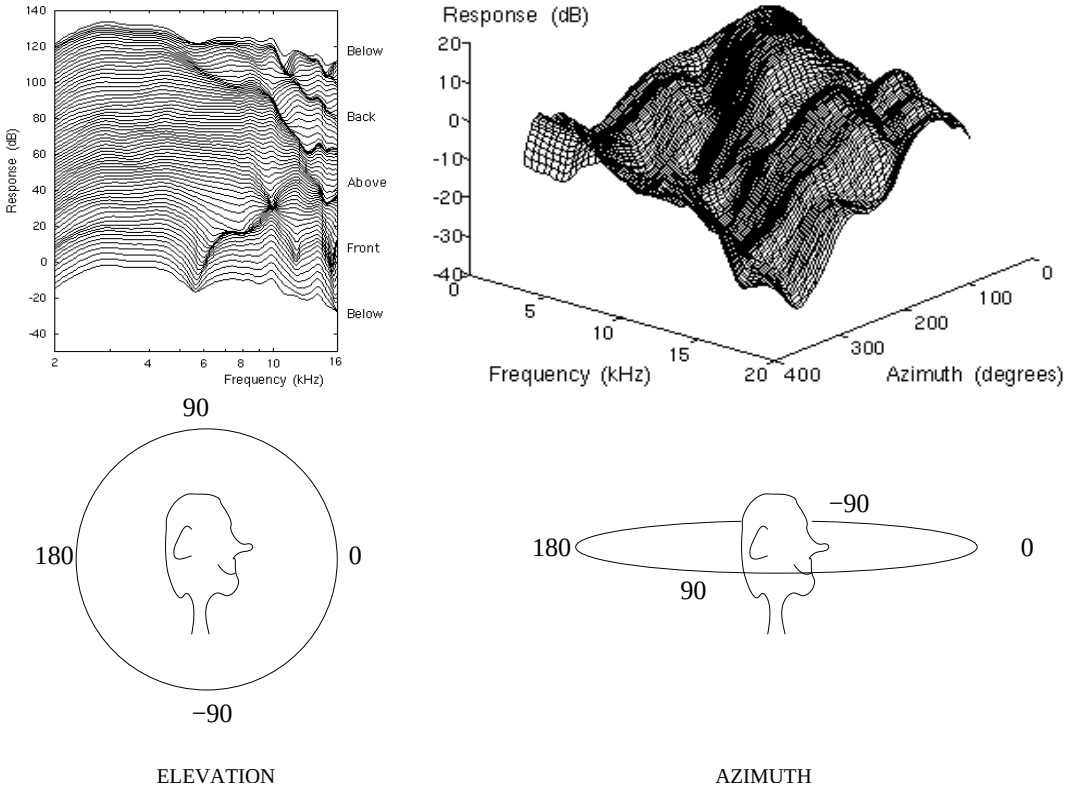
Essentially what we are doing here is decomposing the incoming sound $I_{src}(t;\phi,\theta)$ which is arriving from just one direction into its frequency components and looking at how each frequency component gets transformed by the head and ear. The HRTF $\hat{h}_l(\omega;\theta,\phi)$ is a complex number which specifies a gain (amplitude) and phase shift for each temporal frequency ω 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 ω, θ, ϕ. The above plots showed 2D slices for a fixed θ or a fixed ϕ. You can see examples of a 2D HRTF slice which is a function of θ, ϕ for two different frequencies ω here: `https://auditoryneuroscience.com/topics/acoustic-cues-sound-location`



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.

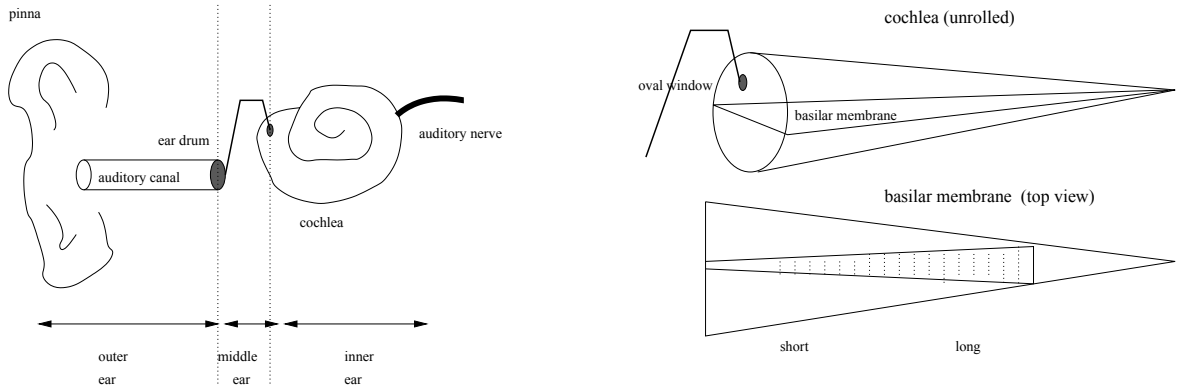
Beyond the ear drum is middle ear. The middle ear is an air filled cavity [37] behind the ear drum. The middle ear contains a rigid chain of three small bones (*ossicles*). One end of the chain is attached to the ear drum and the other end is attached to the *cochlea* which is part of the *inner ear*. 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[38], 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

[37] 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.

[38] 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.

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

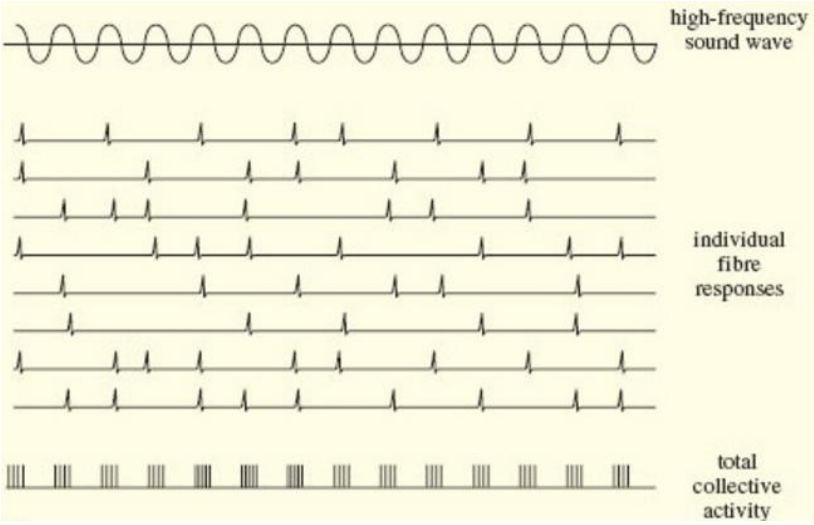
For each of the two cochleas (left and right), there are only about 3,000 hair cells over the entire basilar membrane, and about 30,000 ganglion cells. So think of each hair cell mapping to about 10 ganglion cells. The reason for this 1-to-many mapping is that the ganglion cells cannot spike at rates 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* [39] 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

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

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.



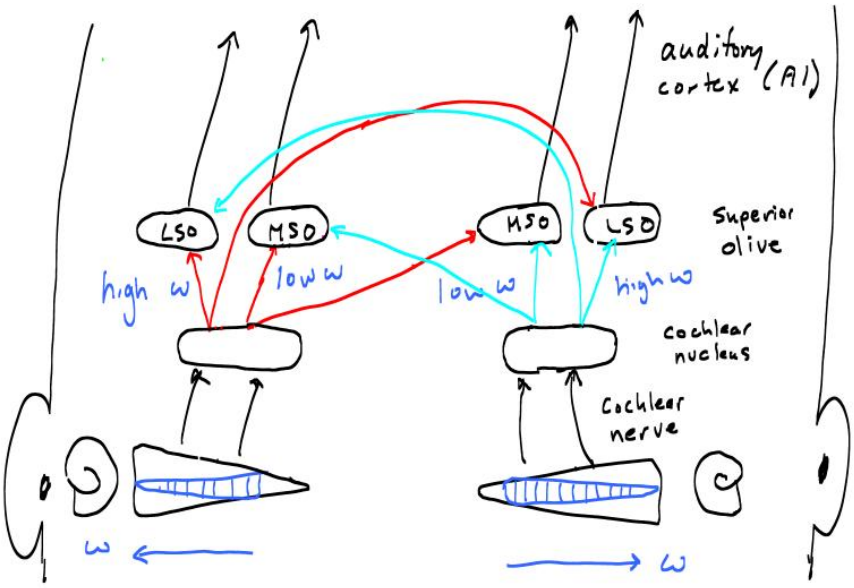
Phase locking only occurs up to a few kHz. At higher frequencies than that, the exact timing of the neurotransmitter release by the hair cells cannot follow the BM motion exactly. Instead, the amount of neurotransmitter released depends simply on the amplitude of BM motion at that location. This amplitude 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 vestibulo-cochlear 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/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.[40] 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.

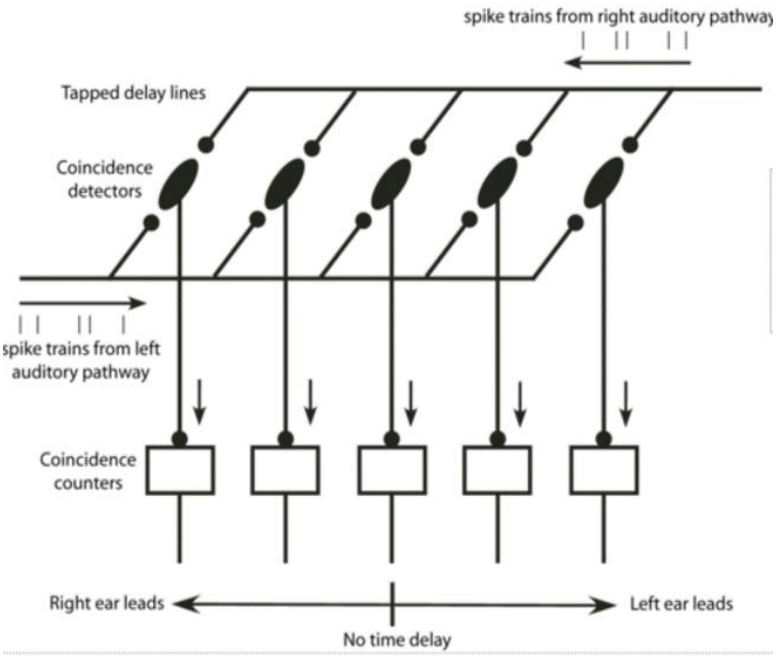
Because they carry different information, low and high frequencies are separated by the auditory system and processed separately. As mentioned above, the LSO receives the high frequency components and computes the level differences between left and right ears. Cells in the LSO are excited by inputs from 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

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

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`.



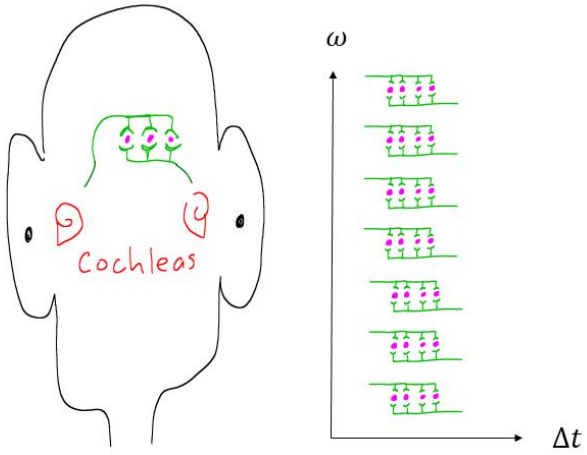
For any sound source in some direction in space and for any frequency band, one of the coincidence detectors for that band will have the greatest response. This greatest response will occur when the signals from the two ears arrive at the coincidence detector at the same time. How exactly this 'coincidence' 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.[41]



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

[41] 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.

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.[42] 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:

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

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.

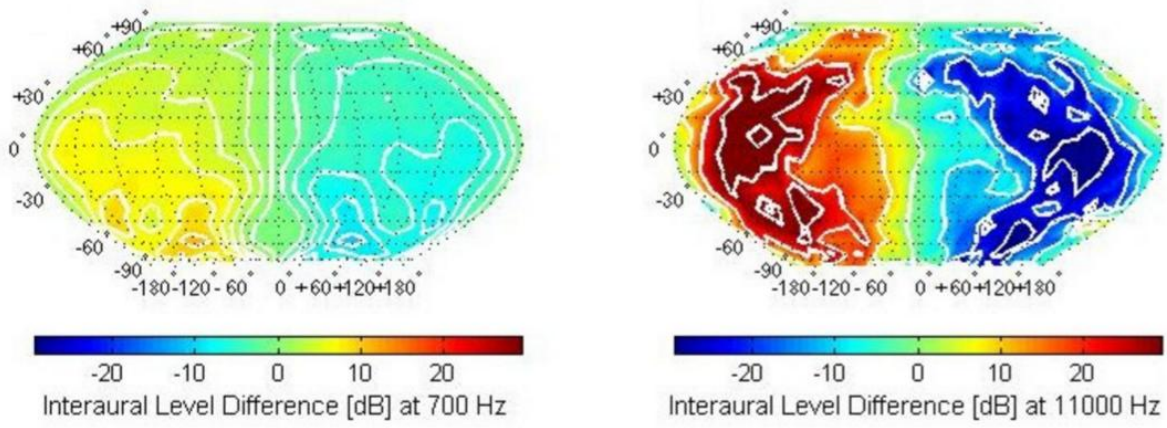
[42] This assumption is an approximation only. Recall the HRIR function from last lecture, where in the medial plane there was some variability in the HRIR over elevation angles ϕ. This suggests that there also *would be* some timing difference in the filtered signals $I_l^j(t,\phi,\theta=0)$ versus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

 ©Michael Langer

Using a mathematical result known as Parseval's theorem[43] 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.

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.

[43] 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

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 and it is believed that the pinnal notch in particular is used. 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.

 ©Michael Langer

Spectrograms (revisited)

We begin the lecture by reviewing the units of spectrograms, which I had only glossed over when I covered spectrograms at the end of lecture 19. We then relate the blocks of a spectrogram to auditory filters and spend the remainder of the lecture on the latter.

Recall that spectogram partitions a signal $I(t)$ into B blocks of length T samples, each and then takes the amplitude spectrum of each block. Each block is typically 10-100 ms. The spectrogram is meant to capture events at that time scale, such as the glottal pulses or parts of speech sounds (vowels versus consonants, voiced vs unvoiced, etc).

The units of spectrograms need to be treated carefully. The Fourier transform of a block uses frequency ω in units of cycles per block, that is, cylces per T samples. These frequency units can be converted to cycles per second by multiplying by ω_0, which is the number of blocks per second. We can think of ω_0 as the fundamental frequency that is represented by the spectrogram.

The block number b can be converted to time in seconds by multiplying by seconds per block, or $\frac{1}{\omega_0}$. The number of samples per block is then $\frac{1}{\omega_0}$ times the number of samples per second. High quality audio signals usually have 44,100 samples per second. To put this another way, if you choose T samples for each block, then dividing T by 44,100 samples per second gives the number of seconds per block.

Putting those conversions aside, it is important to realize that time scales of 10 ms to 100 ms are quite large, relatively to the time scales that we were discussing last lecture when we considered spatial localization. Sound travels at 340 ms^{-1} and so 10 ms sound duration corresponds to 3.4 meters. If a block of a spectrogram is 10 ms long, then this covers 3.4 m of a snapshot of sound. The component of the sound at such a wavelength does not play a role in spatial hearing since the two ears would be at nearly the same phase of the wave at any time and the shadowing by the head and the pinna effects are negligable for such long waves.

Auditory Filters

We have discussed filtering of sound by the outer ear, and last lecture we discussed filtering by the basilar membrane. Researchers have also examined the frequency response properties of ganglion cells in the cochlea by measuring spikes of axons in the cochlear nerve, and researchers have also measured cell responses in the brainstem of various animals. These experiments often use pure tone stimuli. An example of a plot showing different cells and their thresholds for responding to pure tone stimuli was shown in the slides. Typically cells in the cochlear or brainstem have a peak (or center) frequency to which they are tuned. Indeed this is what we meant last lecture when we discussed the tonotopic map of cells along the basilar membrane and in areas such as the cochlear nucleus and MSO and LSO.

Masking and Critical bands

It is also possible to measure and model auditory filters using human or animal psychophysics experiments. A common experiment is to ask how good are we at discriminating two different frequencies. For example (not discussed in class), consider an experiment in which two tones are played, one following the other, and the listener is asked to say whether the tones are the same or different. Another example is *masking experiments*: one tone is presented twice (called the masking tone) one after the after, and another tone is presented just once, namely at the same time as one of the two masking tones. The question is, how loud does the second (called *test*) tone need to be for you to hear it i.e. to say which of the two intervals contains the test. One typically holds the test tone at some frequency and sound pressure level, and varies the frequency and loudness of the masking tone. We say that the masking tone *masks* the test tone.

 ©Michael Langer

Many masking experiments have been done, and consistently show that similar frequencies mask each other much more than different frequencies mask each other. This is consistent with the fact that the cochlea decomposes sounds into bands and then encodes the bands independently. If two frequencies are coded in different bands (or frequency "channels"), then they tend to mask each other less. One often speaks of *critical bands* that cover the range of temporal frequencies that our auditory system is sensitive to.

Models of auditory filters of sound that are based on masking experiments have characterized the bands as follows:

- Below 1000 Hz, human can discriminate two frequencies reliably when they differ by more than about 100 Hz. For this reason, many models of auditory processing begin by filtering the sound below 1 kHz by using about 10 channels, each 100 Hz wide.
- Above 1000 Hz, humans can discriminate two frequencies reliably when they differ by more than $\frac{1}{3}$ of an octave. For this reason, many models of auditory processing filter the sound from 1,000 Hz to 22 kHz using about 14 channels. i.e. $3 \log_2 22 \approx 14$ since there are 3 bands per octave and $\log_2 22$ octaves from 1 to 22 kHz.

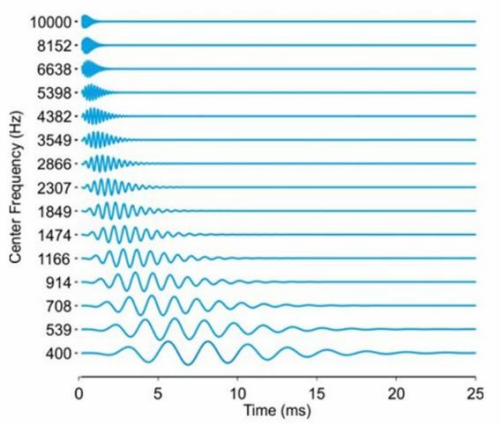
When we refer to *critical bands*, we often think of a partioning up of the frequency range. Note, however, that a 'partition' (mutually exclusive ranges of frequencies) is a convenient model, but does not describe the coding that occurs. There is no partition or boundary between frequency bands, but rather the bands form a continuum of frequencies.

Gammatone filters

The frequency behavior of auditory filter models are similar – whether we are referring to a basilar membraine mechanical response, a ganglion cell or brainstem cell response, or even a psychophyscal response namely critical bands. For this reason, one often conceptually does not distinguish which mechanism we are talking about.

Keeping it general, therefore, let's think of auditory filters as defining an impulse response function (or its Fourier transform, a transfer function). We can model these filters using Gabor functions of various center frequencies and bandwidths. One limitation with gabor function is that they have (Gaussian) tails that go off to infinity. The filter will have some peak sensitivity at some time in the past, but the tail of the filter will reach into the future which of course is impossible since a cell cannot respond to a sound that hasn't occured yet. (This same issue of "causality" came up with motion cells in vision.) The usual way around this is to use a slightly different window than a Gaussian, namely one which is asymmetric and goes to 0. In audition, one often uses a *gammatone filter*. See `https://en.wikipedia.org/wiki/Gammatone_filter` for the formula.

 ©Michael Langer



Examples are shown above. The lowest curve shows a filter with center frequency 400 Hz, so it is most sensitive to sine component whose period is 2.5 ms. You can verify for yourself that the lowest curve has roughly this period for its waves. As the center frequency increases for different curves shown, the period of the waves decreases. In addition, note that for lower frequency filters, the peak of the envelope occurs at a greater time in the past. One way to think of this is to imagine the cochlea and remember that the low frequency components are represented at the far end (the apex) and high frequencies are represented at the near end (base). If you think of the sound as a wave travelling through the cochlea, then this corresponds qualitatively[44] to the response at the near end occuring before that of the far end.

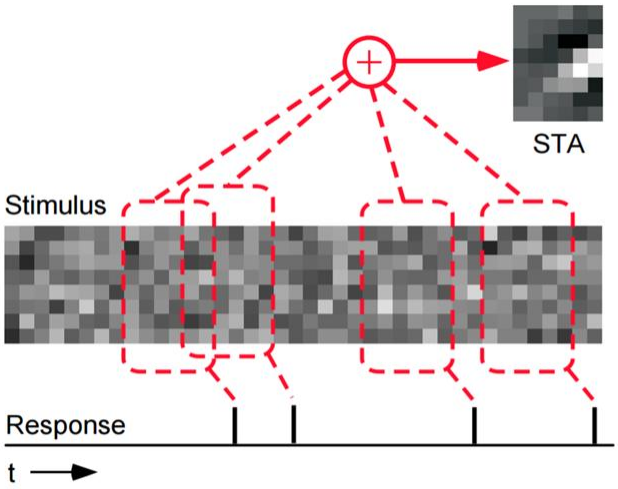
Spike triggered averaging: How to measure a cell's receptive field profile?

Several times in the course I have referred to a cell's receptive field profile. In vision, we saw center-surround cells in the retina and LGN, and we saw oriented cells in V1. In audition, I just mentioned gammatone filters. When I discussed the vision experiments from the 1950's of retinal cells and Hubel and Wiesel's measurement of V1 cells, I described their process as 'trial and error'. Present a stimulus over diffent positions in the visual field and perhaps at different sizes and orientations, and by hand determine which is the preferred stimulus. Then mark out the excitatory and inhibitory regions. This method is fine for some experiments. However, more systematic approaches have also been developed too.

One common method is the *spike triggered average*. The idea is use a random noisy signal as input, and to examine what specific values the signal takes which leads to the cell responding to the noise. The idea is that noise will occasionally by chance present a structure close to what the cell is tuned for, and when it does the cell will be more likely to spike. Spike triggered averaging takes two signals: the noise stimulus signal and the spike train response of the cell. For each spike, it considers a fixed block of time (say 300 ms) in the source signal *prior to that spike*. It then sums up these source signals. The idea is that if *something* in the signal caused the cell to spike at that time, then this something should be revealed by the spike triggered average. This approach has been quite successful.

[44](This 'travelling wave' turns out only to be qualitative, however, as the delay in the peaks of the curves shown doesn't correspond to the speed at which the sound wave propagates in the cochlear; rather it has more to do with the mechanics of the basilar membrane.

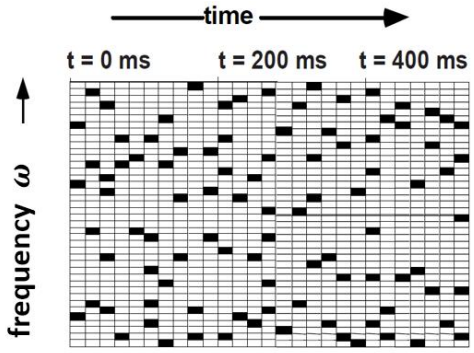
The example below shows an XT stimulus. The spike triggered average (STA) over four spikes is shown. In general one takes the average of the stimulus over thousands of spikes. A real example (for a V1 neuron) is shown in the slides.



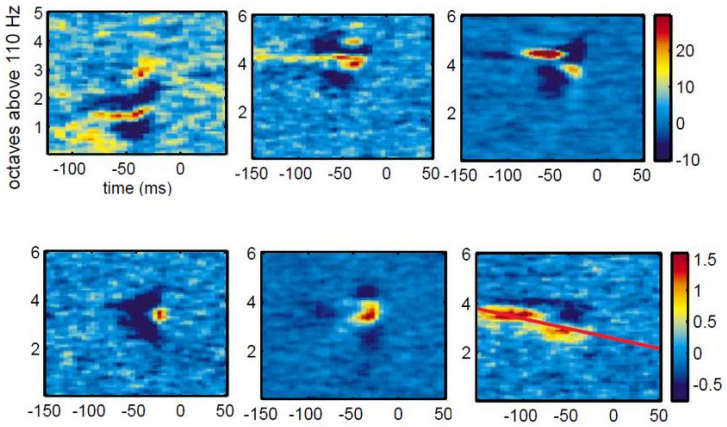
Auditory Cortex (A1)

Can spike triggered averaging be used to discover receptive fields in auditory cortex? (A1 is the audition analogue of V1, namely it is where the auditory signals are first processed in the cortex.) In principle, yes. However, in practice it has been difficult to do because many cells do not respond well to pure tones, regardless of the frequency. Moreover, spike triggered averaging doesn't work well either, if one uses a white noise stimulus e.g. the sound 'sssssssss'.

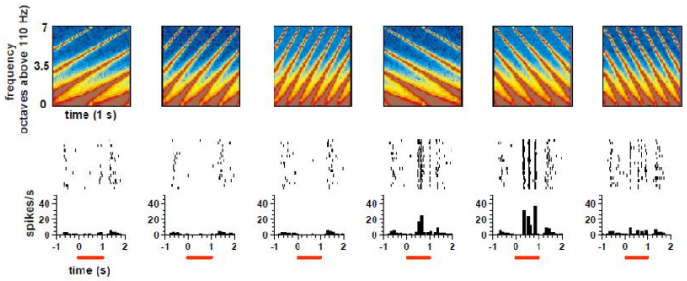
In the late 1990's, another idea for using spike triggered average was tested. Rather than using pure noise ("sssssss"), instead random 'chords' were used which consisted of a sequence of short duration intervals of bandlimited noise. See illustration of a spectrogram of this random chord noise.



Examples of spike-triggered averages of some A1 cells are shown below. The axes are frequency versus time. Negative values of time indicate that the spikes (which were at time 0) responded to the parts of the sound that occured before the spike.



Notice that these cells are not simply excitatory for some particular band of frequencies. Rather, these A1 cells have both excitatory and inhibitory regions of the receptive field. The cell in the bottom right corner, in particular, seems to be sensitive to *frequency modulation* as indicated by the diagonal lines.



For this cell, the authors then confirmed that it was indeed sensitive to frequency modulation. They examined the responses of the cell to various FM modulated stimuli (see above). The plot shows six 'orientations' of FM modulated stimuli. The black dots below show rows of cell responses, namely spike trains. Each row is one trial i.e. one example where the sound plays. Each black dot is a spike. The plots the bottom show histograms where the rows of spikes are summed up – called a *peristimulus time histogram*. The main point here is that you get more spikes from the cell when the sound is FM modulated such that the frequencies decrease over time, as in the spike triggered average shown above (bottom right receptive field profile, with red line drawn on it).

At the end of the lecture, I briefly mentioned a few applications, namely cochlear implants and MP3 compression. Both of these applications are based on the theory of auditor filtering. I am leaving that discussion out of these lecture notes for now.

Echolocation

Suppose that you wished to judge the 3D position of objects around us by clapping your hands and listening for the echo. The time between hand clap and echo in principle can tell you how far you are from object. This is the method of *sonar* (sound navigation and ranging). Note that the *distance* to the object is determined by time delay between the hand clap and the arrival of the echo, namely the distance is $\frac{1}{2}v\ \tau$ where v is the speed of sound and τ is the delay. The reason for the factor $\frac{1}{2}$ is that the sound has to go to the object and back again. Also note that the detailed structure of the echo could tell you the direction, for example the timing and level differences in your two years.

One issue that arises with this simple method is that the reflected sound (the echo) is much weaker than the original sound. The hand clap sends off a spherical wave of sound in all directions. For any small cone of directions, the energy remains in that cone and travels radially outward from the source. (There is also a *loss* of energy due to friction/attenuation in the air, which is greater for high frequencies). As discussed a few lectures ago, the area of each wavefront of the expanding sphere is $4\pi r^2$, and the sound energy per unit area of the sphere must fall off as $\frac{1}{r^2}$, and the sound pressure level (SPL) falls off as $\frac{1}{r}$. If the sound of the hand clap reflects off a small object in the scene, the reflection that arrives back at the source will be weak. Take a small flat surface of area A which faces the source. This surface receives about $\frac{A}{4\pi r^2}$ of the energy of the hand clap. Some fraction of this energy is reflected back, though the reflection occurs in many directions: the small flat surface acts as a small sound source and the wave it reflects radiates back as a sphere. By the time the reflected wave reaches the original source (the ears of the person that clapped hands), the reflected echo energy is proportional to $\frac{A}{4\pi r^2}\frac{1}{r^2}$ which is proportional to $\frac{1}{r^4}$. Thus we see that although timing of the echo carries information about distance, it is not obvious that this echo can be reliably measured since it may be too weak.

Let's now examine echolocation in bats, and how they deal with this problem.

Bat sonar

Bats are among the strangest of the mammals. They have horrific pointed ears and large flaring nostrils. But perhaps what makes bats most frightening and alien is that they can navigate in the dark. Humans fear the dark. Bats thrive in it.

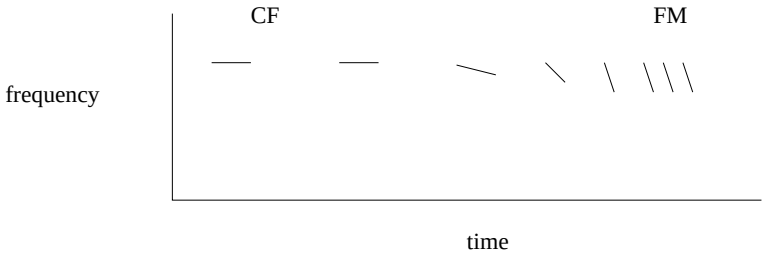
How bats manage to live in darkness was a mystery for centuries. Bats do have eyes so obviously they can see. It was long believed that bats had much more sensitive visual systems than humans and other mammals, and that bats can see in the dark because they only need very little light and they can adapt to lower light levels than people can. Others believed that bats sensed the location of objects by "feeling" the reflection of pressure waves which are produced by the bat's beating wings. Both of these commonly held beliefs turned out to be completely wrong.

Over 200 years ago, an Italian named Spallanzani carried out experiments that showed bats use hearing to navigate in the dark. He captured a set of bats and blinded them, and verified they could navigate fine – avoiding very small obstacles and even continuing to feed on flying insects. This ruled out the hypersensitive vision hypothesis. Ruling out the wing beat theory was more difficult. Instead, he proposed another hypothesis, namely that bats rely on hearing to navigate. To test this hypothesis, he inserted small hollow cones into a blinded bat's ears. The bat could still navigate just fine. He then filled the cones with wax, blocking out most of the sound. With the cones blocked, the bats could not navigate at all. They crashed into walls, objects, and were unable to feed on flying insects.

Spallanzani's experiments were carefully done but were largely rejected by the scientific community. The reason is that he had not explained *how* bats use hearing to navigate and locate insects. It wasn't until the late 1930's that this puzzle was solved. The key insight came from experiments done by Donald Griffin. Griffin was the first to be able to measure the cries of bats. This had been difficult to do previously

because bat cries are at frequencies much higher than what people can hear (i.e. ultrasonic). There were few devices before Griffin's time for recording ultrasonic frequencies and those that did exist were expensive and so were not used to record bat sounds.

Griffin was the first to record the sounds made by a bat as they navigate (in a big cage in his lab) and hunt for food. If you compute a spectrogram of these sounds, you often find a pattern such as shown in the figure below. The time and frequency scales are not shown, but here are some rough numbers. Bat cries are from 12 kHz up to 200 kHz. Most of this range is non-audible for humans. I will explain what CF and FM mean below.



If we take 34 kHz, for example, this is 34 cycles in 1 ms which (according to $\omega\lambda = v$) corresponds to a wavelength of about 1 cm. Similarly, a high frequency such as 170 kHz corresponds to a wavelength of 2 mm. These wavelengths are behaviorally relevant for the bat. They are the size of objects such as tree branches, and edible insects such as moths which the bats eat.

There is also a large range in the durations of bat cries. A single cry can be as long as 200 ms and as short as a fraction of 1 ms. In the spectrogram above, the horizontal lines on the left have a larger time duration than the highly slanted lines on the right.

There are over 700 species of bat. Different species live in different habitats, have different shaped ears, eat different foods, and have evolved different mechanisms for using echolocation. Some bats cry through their noses, others through their mouths. Despite the variations, there are general echolocation principles that have been discovered which seem to explain much of bat echolocation behavior.

Consider again the spectrogram above. In order to explain such a spectrogram, we need to consider what is involved in hunting for food – what problems need to be solved? Suppose you are a bat flying through the air and you are hungry for flying insects. You need to solve three problems, in the following order:

1. *Detect:* is there something out there?
2. *Localize:* where is it ? (distance and direction)
3. *Recognize*: what is it? (shape, motion, material)

Roughly speaking, there are two kinds of bat cries that are used to solve these problems: constant frequency (CF) and frequency modulation (FM). These correspond to the horizontal and highly slanted lines in the spectrogram above, respectively.

Constant frequency (CF)

The frequency composition of a CF cry does not change throughout the cry. CF cries are composed of a very small range of frequencies. In order to achieve this property, the cry must to have a long duration. Constant frequency signals are typically over 10 ms, and are often well over 100 ms. To understand why a CF cry must have a long duration, recall the properties of a time Gabor functions: if the Gabor has a small

bandwidth $\Delta\omega$ i.e. a Gaussian with a small standard deviation, then the Gabor must have a large Gaussian window in time.[45] To make a CF cry, presumably the bat must use a voiced sound, but the glottal pulses can only be so fast. The high frequency presumably arises from the articulation which amplifies a small set of frequencies and attenuates most of the others.

What are the advantages and disadvantages of CF cries, with respect to solving the three problems mentioned above? First consider detection. Bats are mammals and their auditory system is like our own in that it encodes the sound using bandpass filters of varying bandwidth ("critical bands"). So, CF is good for detection because it puts alot of sound energy within a single critical band, and so the echos of a CF cry also lie within one critical band. The concentration of energy in a single critical band makes it easier to detect this signal in the presence of other sounds in the environment and noise in the auditory system in that band.

One important difference between the bat's critical bands and our own is that the bat's critical bands contain an *acoustic fovea*. Recall from our discussion of vision that the human retina packs a high percentage of photoreceptors into one small area, i.e. the direction in which we are looking. The bat's acoustic fovea devotes more cells to a particular range of positions on the basilar membrane. (An example is shown in the lecture slides.) This allows the bat very good frequency discrimination at these frequencies, as well as the ability to detect relatively quiet sounds at these frequencies. These are the frequencies near which the bat makes its CF cries. This specialization is helpful because the environment may contain sound energy at many frequencies, and the bat wishes to only hear the echos of sounds that it generates.

How does the bat hear the echo of its CF cries, in the presence of the cries themselves? There are two answers to this. First, the bat leaves a gap of silence between its CF cries, which allows time for the cry to propogate through space, reflect, and then return to the bat. By the time the echoed cry returns, the cry should be over. The longer the gap, the less of a forward masking[46] occurs.

Second, note that if the bat is flying forward, then the emitted cry will undergo a *Doppler shift*. The bat will be chasing the sound as it emits the sound, which will lead to an increase in the frequency of sounds that are received at the reflecting surface. These higher frequency sounds are then reflected back and the bar flies toward these sounds which results in another Doppler shift.

Suppose the bat were flying forward with velocity v_{bat} and emitting sound at some frequency ω_{emit}. Once can show that the frequency of the echo observed by a forward flying bat is:

$$\omega_{observed} = \omega_{emit} \frac{v_{sound} + v_{bat}}{v_{sound} - v_{bat}}.$$

So, for example, if the bat's speed is say $\frac{1}{100}$ of the speed of sound, then we get an approximately 2% increase in the frequency from the Doppler shift. e.g. if the bat emits a cry at 100 kHz, then the shift can be about 2 kHz. This is not a lot but it is enough to put the echo into a different critical band. The idea here is that if the bat emits a sound just below the frequency of the acoustic fovea which has a very small bandwidth, then the reflected sound will fall in the fovea and masking is avoided.

Once the bat has detected an object and has a rough estimate of the depth (based on delay between cry and echo), it needs to estimate the location. For this it can use the monocular and binocular cues we have discussed in the previous lecture, namely frequency based cues arising from the HRTF, level and timing differences.

Are these cues useful for CF cries? The level differences are of little use since the bat directs most of the sound generally forward and there is presumably not much level difference between the two ears near the forward direction. The timing differences are also not very useful since the CF cries are high frequency (recall the duplex theory).

[45] Careful: The converse is not true. I am not saying that a long duration sound always has a small bandwidth.

[46] Masking of sound B by sound A can occur even if the sounds are not played simultaneously. Foreward masking means that the mask occurs before the test. Backwards masking means that the mask occurs after the test.

 ©Michael Langer

One might wonder if the *envelope* of the CF cry could be used for a timing differences. Probably not. The CF cry has a gradual envelope and so it doesn't have a well defined starting and ending point. (To understand this, think of the smooth Gaussian envelope of a Gabor function. Where does a smooth signal start?)

In order to get well defined timing signals, the bat instead needs to modulate the frequency over time using FM, as I will describe below. Before we do so, let's consider one more aspect of CF signals, namely how could be used for *recognition*. Suppose the bat would like to decide whether to pursue a flying insect, based on what kind of insect it is. Bats have tastes, just like we do. How can the bat do so? Suppose a particular species of moth beats its wings at a rate of about 40 beats per second, and so the wingbeat period is around 25 ms. When the wing is perpendicular to the direction of the bat's cry, the echo back toward the bat is maximal, and when the wing is parallel to the direction of the cry, the echo is minimal. Thus for long CF cries (100 ms), the echo contains a periodic structure – on, off, on, off, etc. Different moths have different wing beat rate, and this cue can be used for recognition!

Frequency modulated (FM)

Frequency modulated cries are roughly of the form $\sin(\omega t)$ where ω is itself a function of t, for example, $\omega = \omega_0 - \beta t$. (A slightly different form of this equation was given in the lecture slides.) The idea is that the frequency near the beginning of the cry is different than near the end of the cry. The bigger is β, the faster the frequency drops over time and the steeper the slope in a spectrogram representation. Such a signal is also called a *chirp*.

What are the advantages and disadvantages of FM cries? One disadvantage is that FM cries are poor for detecting objects at a distance. The cry (and hence the echo) sweeps through each critical band for a short time only. The bat has to use the signal within each critical band to detect the echo amid environmental noise, so if the bat is still far from the reflecting object then the echo will have little energy in each band. Think of an analogy in vision: consider a 2D sine grating with noise. If the 2D sine grating covers the whole display of your monitor then it will be easier to detect in the noise than if the sine component (of the same frequency) just covers a small window say 50 x 50 pixels in the the display.

The advantage of FM comes when the bat is close to its target. FM cries are loud enough in each channel to be heard. Moreover, since the duration of the bat's cry and the echo within each critical band is short, the timing difference between the cry and echo can be computed more precisely than what we had with the CF signal where the duration of the cry was large. In this sense, FM cries and their echos carry accurate information about distance. Note that FM cries do not need to be separated by long silent gaps, as did CF cries. Provided the temporal duration within each frequency channel is short enough that the echo in that frequency channel doesn't return before the component of cry in that channel is complete, then there is no overlap in that channel. This is a big advantage with FM.

The timing differences just mentioned were between the cry and the echo and gave information about the distance. There also will be timing differences for arrival of the echo at the two ears, provided that the target is not straight ahead (in the medial plane). Note that cries are high frequencies, so the bat brain is not matching individual spikes. Rather it is measuring time differences between envelopes. Also, lots of frequency channels are activated by the FM sweep and so the auditory system can combine signals across channels to get timing differences, as humans do. The same argument can be made for level differences. So, with FM signals bats can use binaural hearing cues just as humans do.

What about recognition? Earlier I mentioned that the wing beats of a moth can be used as a cue for recognition for CF cries. For FM cries, the moth wingbeat information is useless because the duration of the cry is much shorter than the period of a wingbeats. Can FM cries be used for recognition? Yes they can! When an FM cry echos off an insect such as a moth, there is rarely a single echo, but rather multiple

echos[47]: the pulse bounces off a wing and also off the body – or off the head and the wing – and these two reflecting surfaces may be at slightly different depths. To keep the analysis simple and just get the basic idea, let's look at the overlap of two echos, and we'll use a toy model where the moth impulse response function is

$$m(t) = a\delta(t) + b\delta(t-\tau).$$

where $a > 0, b > 0$. So there are two echos and the second is shifted in time by τ relative to the first. Letting ω have units cycles per second[48] and let τ be in units of seconds, we have

$$\mathbf{F}\ \{\ m(t)\ \}\ =\ a\ +\ b\ e^{-i2\pi\omega\tau}\ .$$

Observe that *constructive interference* occurs when $\omega = \frac{1}{\tau}, \frac{2}{\tau}, \frac{3}{\tau}, \ldots$ etc and *destructive interference* occurs when $2\omega\tau$ is an odd integer, so $\omega = \frac{1}{2\tau}, \frac{3}{2\tau}, \frac{5}{2\tau}, etc.$ See the Exercises for an example.

[47] called *glints,* like the specular reflections off water waves

[48] If you prefer, we could use the familiar units of samples t and cycles ω per T samples, and in that case we would write

$$\mathbf{F}\ \{\ m(t)\ \}\ =\ a\ +\ b\ e^{-i\frac{2\pi}{T}\omega\tau}\ .$$

 ©Michael Langer

The main idea here is that two echos separated by a small distance can produce a systematic interference pattern. If an FM cry had a roughly constant level (SPL) over a range of frequencies that it sweeps through, then the echo would *not* have constant level over frequency. In general, objects such as moths or flying beetles will have more complex echos than the simple toy model above. But whatever the echo pattern is, it will be a signature of the shape/orientation of the moth. Think of the object that is reflecting the sounds as having a transfer function, similar to what the HRTF does to a sound arriving at the head. By detecting which bands receive sound and which do not, it is possible to infer the something about the shape of the reflecting surface. Amazingly, behavioral experiments with bats have shown that bats can indeed discriminate between various spectral patterns in echos.

To briefly summarize, the bat uses the CF signals to detect and recognize the object producing the echo. If there is an object and it is worth pursuing (wing beat frequency corresponds to edible moth species), then it gradually switches to an FM cry. The FM cry is a shorter duration signal and the length of the signal within each band is much shorter. This provides better timing information which allows accurate distance estimation, as well as direction estimation: more channels are active which allows binaural and monaural spectral cues to be used for localization. It also allows spectral cues to be used for recognition, since the object's size and shape determine the constructive and destructive interference in the reflected echos.

How dolphins and porpoises use echorecognition

Dolphins and porpoises and other marine mammals also use echolocation and echorecognition. These animals are very sociable with people and so they can be trained to perform many behavioral tasks. A key difference between dolphin and bat sounds is that porpoises do not use FM cries, but rather they use "clicks". They are called clicks because that is what they sound like to a human listener. The center frequency is often in the 60-150 kHz range and there are 2-3 cycles within the envelope's half height which corresponds to about half an octave. For example, for a 120 kHz center frequency, you typically have about a 60 kHz bandwidth. (We can't hear these clicks, but porpoises also emit clicks at much lower frequencies that do fall within the human range.[49])

Dolphins can echolocate objects and also use echos to recognize them. They can distinguish the shapes and materials of the objects that produce the echos. This allows them to distinguish different types of fish, for example, some of which are easier to catch or to digest.

Let's just look at one aspect of this, namely the constructive and destructive interference patterns in the echo. The idea is similar to what we saw with the bat, but the click is different so new issues arise. Suppose a target fish is aligned so its body axis is perpendicular to the line between the dolphin and the fish. When dolphin click is reflected off a fish, the sound reflects off both the front surface as well as the back surface – the sound passes through the fish and back out. (There are multiple reflections within the fish, but let's keep things simple and just consider two echos, as we did above i.e. two echos using a formula such as written above.) The reason that the sound passes through the fish is that the fish is made mostly of water. (Discussion of 'impedence' omitted.)

The speed of sound in water is about 1500 m/s which is about four times faster than in air. Suppose we take center frequency of 75 kHz. Verify using the $v = \omega\lambda$ formula that the wavelength is $\lambda = 2cm$. Consider possible thicknesses of the fish and whether these produce constructive or destructive interference for a given wavelength. Note that the echo that passes through the fish must pass through it twice (go and come back). Thus, if the fish thickness is 1/4 wavelength of the sound of some frequency, then twice the fish thickness (aller-retour) will be 1/2 a wavelength which will give destructive interference, whereas if the fish thickness is 1/2 wavelength then we get constructive, etc

[49] `https://www.dolphins-for-kids.com/dolphin-clicks-and-whistles`

fish thickness

.25 *λ	destructive
.5 *λ	constructive
.75 *λ	destructive
1 *λ	constructive
1.25 *λ	destructive
1.5 *λ	constructive
1.75 *λ	destructive
etc	

Keep in mind that the above is just a toy model, meant to illustrate some basic properties of echose. Hopefully this is enough for you to appreciate how the reflected sound might be quite different than the emitted click.

Like all hearing mammals, the dolphin processes the reflected sound by using bandpass auditory filters. The frequencies are far too great for the details of the sound wave to be captured by the timing of spikes. Rather, auditory nerve cells that encode the echos hear will respond to those frequencies that are present and will compare those that are present with those that are absent. As with bats, this can be used not just to locate objects but also to recognize them. [50]

Human echolocation

Humans can use echolocation as well. Blind people use their cane not just to feel their way through the world, but also to making tapping sounds and listen for the echos of these sounds. Some blind people have taught themselves to make clicking sounds as well. See many videos of Daniel Kish, an advocate for blind people, e.g. `https://www.ted.com/talks/daniel_kish_how_i_use_sonar_to_navigate_the_world`

[50] `https://books.google.ca/books/about/The_Sonar_of_Dolphins.html?id=Q3MIsrPDA5EC&redir_esc=y`