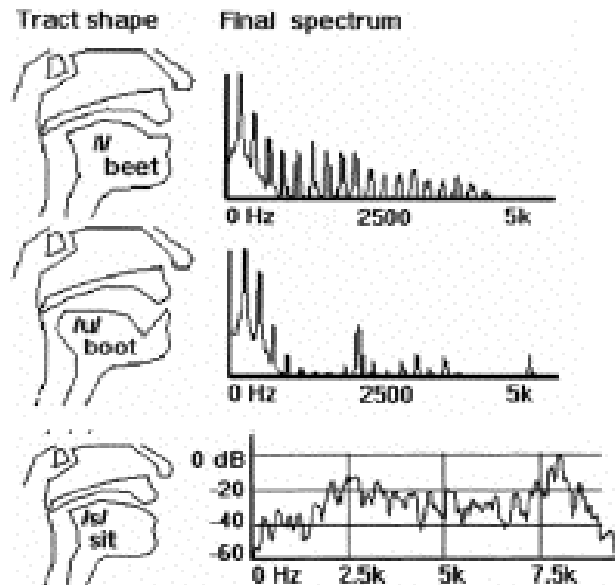


## Chapter 1 Section 5

### The Ear

Check out the pictures below (yes, we've seen 'em both before). The first picture along is the spectra of some common sounds, components of English words. Note where the peak frequency of the sounds sit.



The next picture is a curve, telling you how loud a sound has to be before it can be heard. The frequency of the sound is on the horizontal axis; the volume on the vertical axis. You see right away that some frequencies are easy to hear, some very hard.

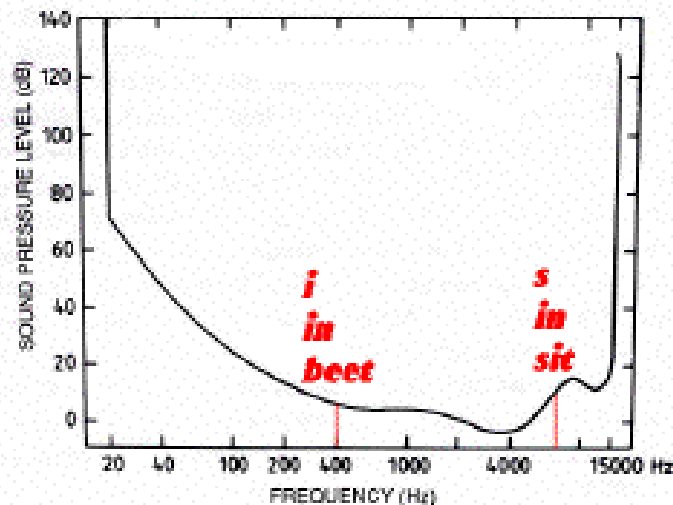
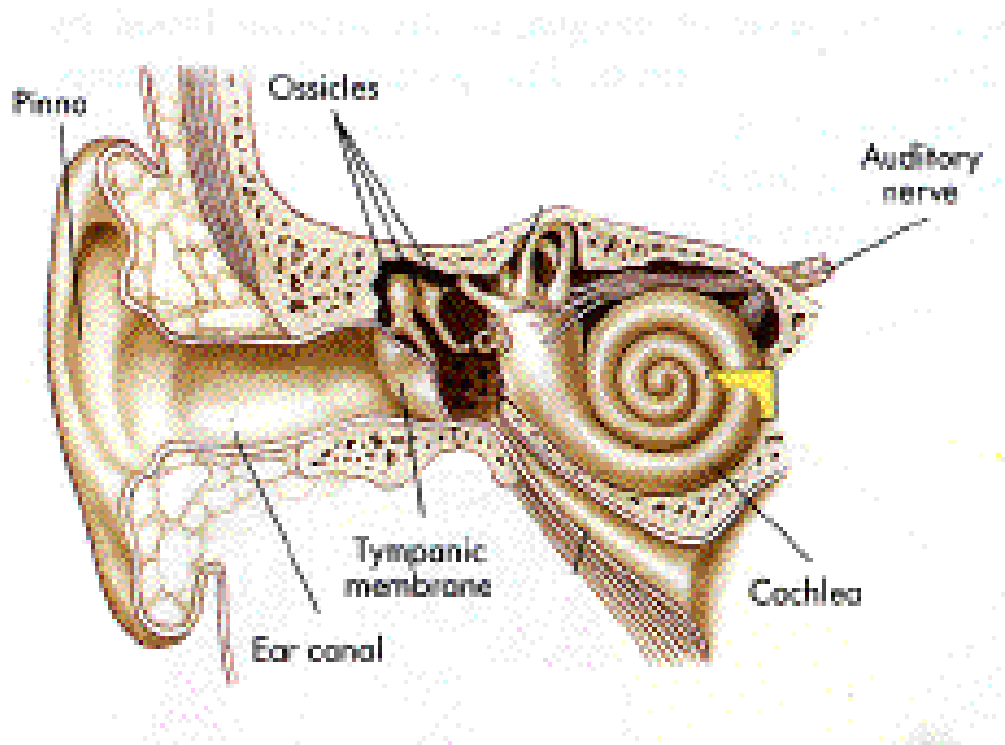


Figure 3.B Audible frequency and sound level range.

While I was at it, I plotted peak frequencies too. It's kind of interesting that the ear hears best, in the range where humans produce their speech sounds.

And that's what I want to talk about in this lecture: how the ear converts sound to sound. Or, to be serious, how the ear converts pressure in the air to something brains recognize as sound.

Let's start with the cross section of your basic ear:

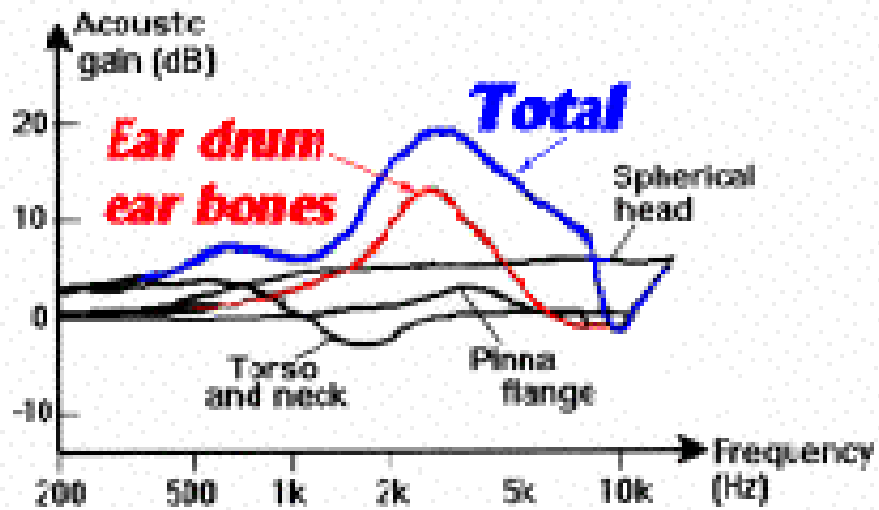


The parts have different names than in common speech. For example, the floppy things on the side of your face, called ears, are the pinnae. The thing your parents say you're going to pop isn't called the eardrum, it's called the tympanic membrane. The earbones all have technical names like malleus (hammer in Latin) and are collectively referred to as "ossicles". And the cochlea is the cochlea, which oddly enough you seldom hear a lot about. Notice that there are looping structures on top of the cochlea; these are the semicircular canals, and have to do with balance, not hearing.

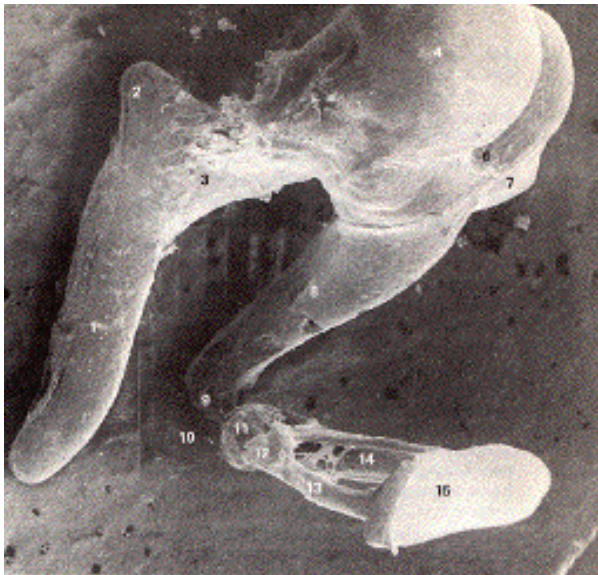
What happens to the varying air pressure is that it causes the eardrum to move in and out. The ear bones are attached to the eardrum, so they move as well, and the last bone rests on the cochlea. This transfers pressure to the cochlea.

It all seems kind of pointless. Why not have the ear open directly onto the cochlea?

What's accomplished by this scheme is that a large opening gathers the sound pressure waves, then transfers those waves to a much smaller opening. Going from large opening to small opening increases the force per square inch -- that is, it increases the pressure. So, the whole eardrum-ear bones-cochlea arrangement AMPLIFIES sound, makes it easier to hear.



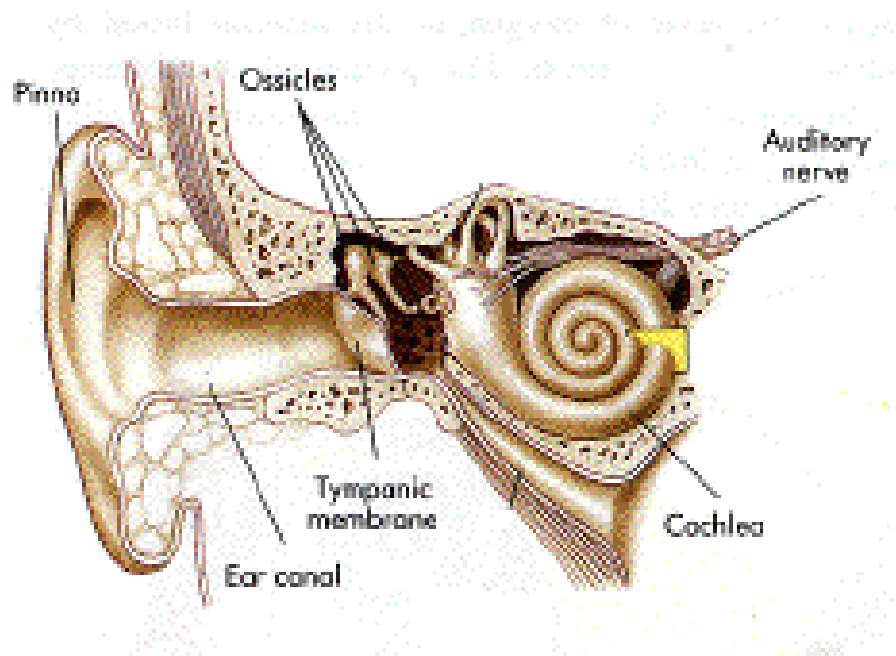
Actually, it does a bit more: it preferentially amplifies certain frequencies of sound. Above are the details. You see there is a lot of amplification, in blue, due to a lot of factors . . . the ears (pinnae flange), the body itself (torso) -- but a large chunk of the amplification, in red, is due to the eardrum-ear bones system. And you can see that the amplification starts around 500hz, and tapers off around 5000hz.



By the way, just because I can, a picture of the ossicles, the ear bones. they're neat, in a 'bones' kind of way.

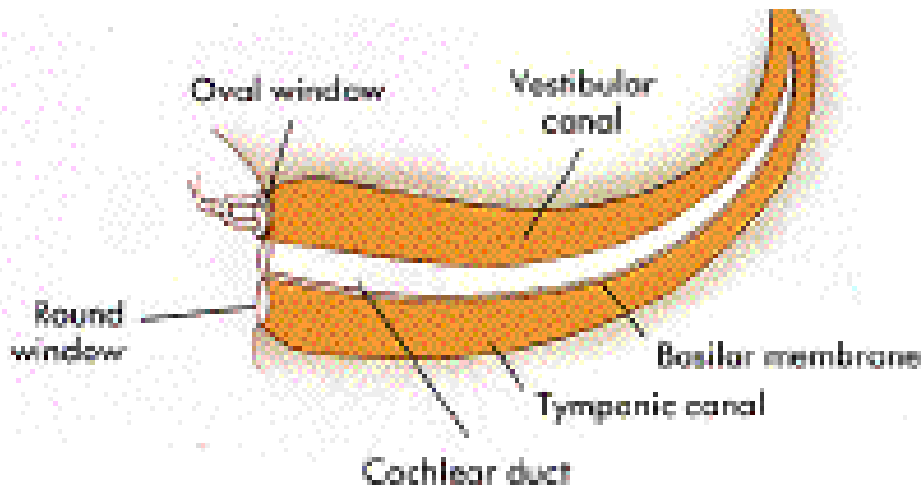


Here's another really neat picture, showing where all the ear structures are located, inside the human head. They're a lot deeper inside than I ever expected.



The ear again, emphasizing the cochlea. The ossicles transfer pressure to the cochlea, amplifying certain frequencies, and the nerves lead away from the cochlea to the brain. So, somewhere inside the cochlea, pressure gets transferred to nerve impulses.

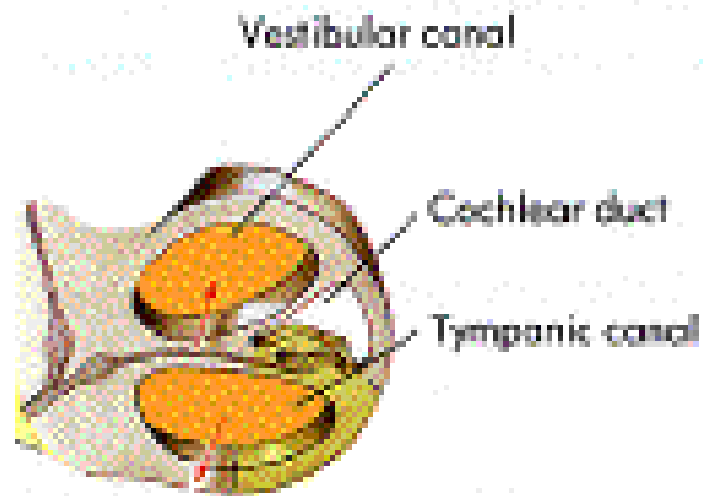
Though the cochlea is all coiled up like a snail, you sometimes see it drawn unrolled, like so:



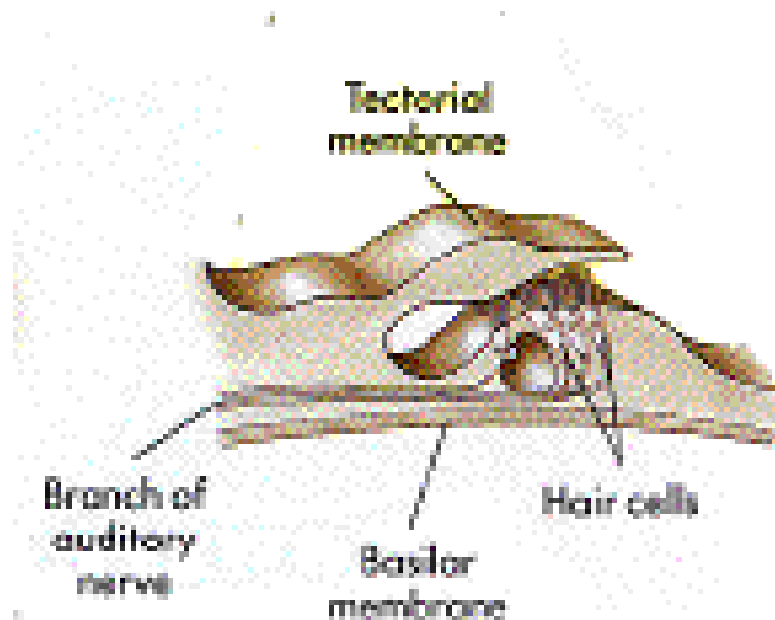
You see the ossicle pressing on a region of the cochlea called the oval window; actually that's a membrane like the eardrum. When pressed, it transfers pressure to a fluid (here colored mustard, due to an unfortunate genetic mishap responsible for my artistic abilities), and the fluid moves through canals, all the way around the white space, to the other side and another membrane, the round window. So, instead of waves of pressure in air, now we got waves of pressure in a fluid.

The white space sits like a hot dog inside a bun . . . a vibrating bun. This causes the lining of the white space to vibrate too. Except it isn't "space" or a hot dog, it's called the cochlear duct, and the lining is called the basilar membrane. However, it isn't just a membrane, it connects to a complicated sensory system, containing nerves, and all sorts of

micro-structures that do the actual hearing of the sound. Here's a picture, a cross-section of the cochlea:

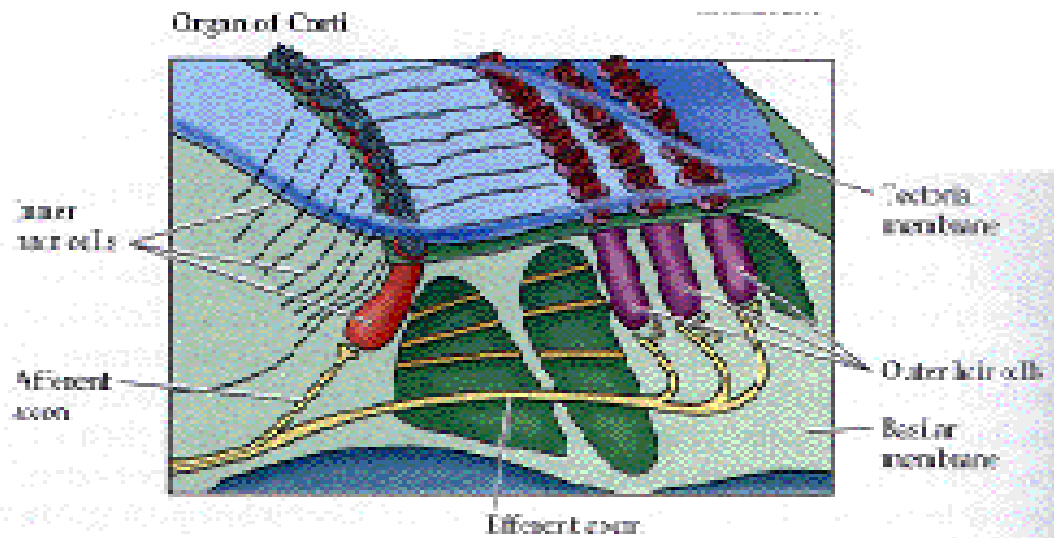


You see your mustard colored fluid, and you see the canals it flows through, and you see the bony structure of the cochlea that holds everything together. And you can also, finally, see the white space more clearly, and see that the white space has stuff inside. Here's the stuff:



As the pressure wave travels along the canals, it presses against the basilar membrane, and the membrane moves. You can see that the membrane has "hair cells" embedded in it; and those hairs push up against the tectorial membrane. The tectorial membrane doesn't vibrate very much, so when the pressure wave passes through, the basilar membrane pushes the hairs against the tectorial membrane. It's like the hairs on your arm being pushed, and when that happens, you feel it. When the ear hairs are pushed, you hear it.

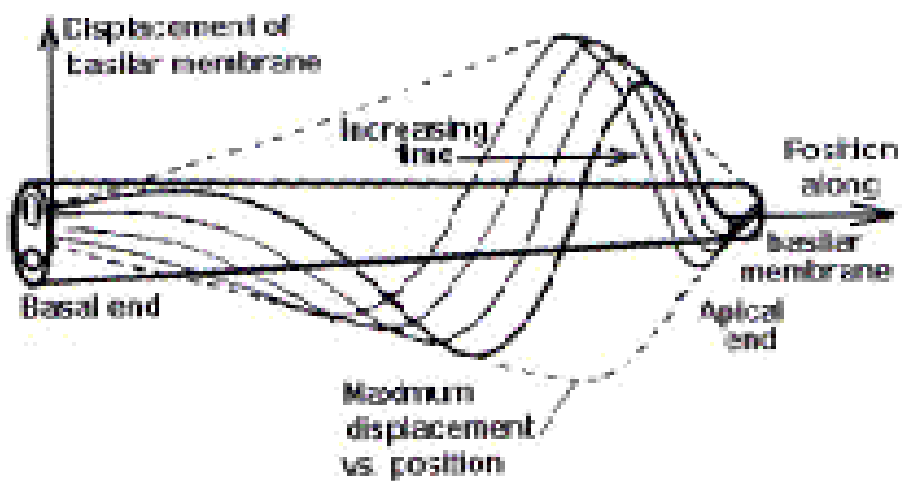
Finally, here's an ultra-close up of the hair cells, embedded in the basilar membrane, squashed from above by the tectorial membrane, with nerves leading to and from the hair cells.



Where we are now: the nerve impulses that the brain will recognize as sounds, are produced when the hair cells move, causing a small electrical impulse to be released from the cell. We say sound is encoded in the firing of the nerve cells. That's the next question, how the coding works.

The theory of how that works is called place theory. I'm using "theory" here because it still is just a theory, and no-one really knows everything about hearing. There's already been one Nobel Prize awarded for work on hearing theory and I bet there's some more for the people who finally figure it out.

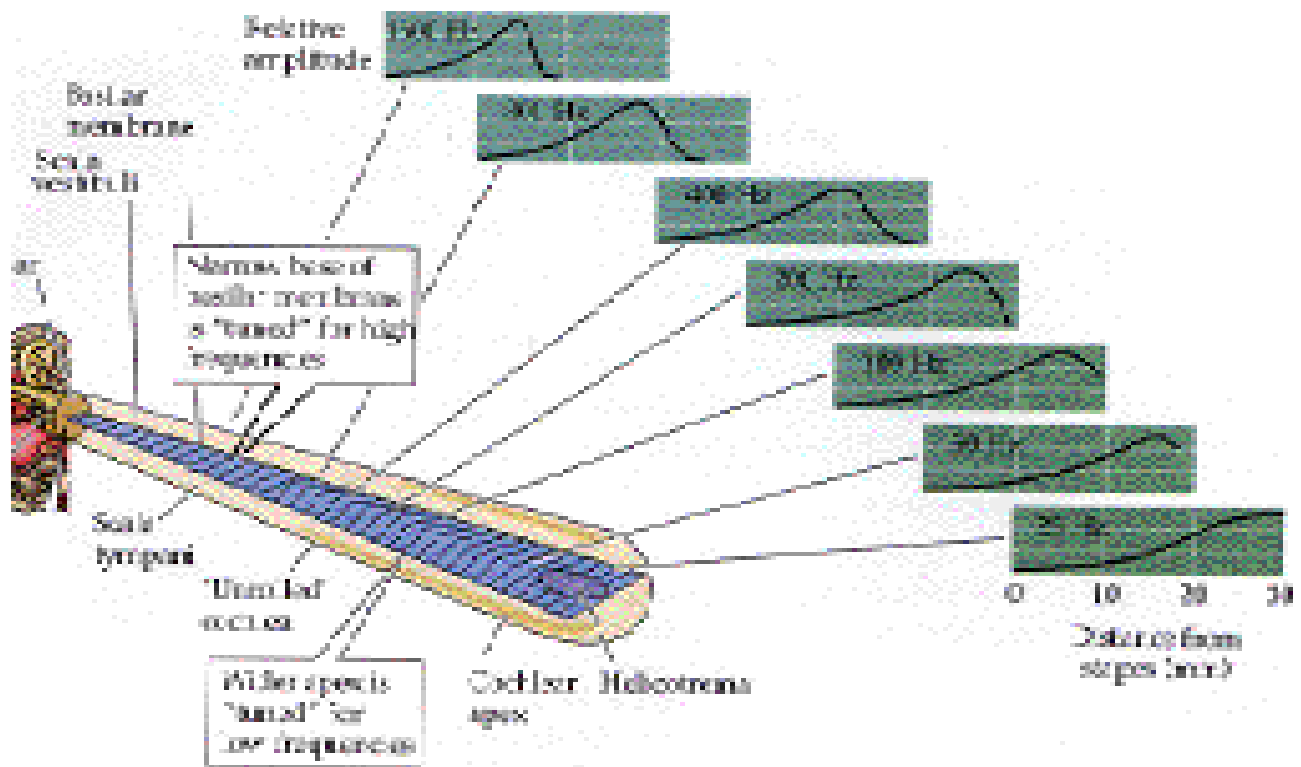
When a wave of pressure rolls through the canals of the cochlea, the basilar membrane responds by moving up and down, like a flag waving in the wind. Here's a sketch of what it looks like:



The waving of the basilar membrane is exaggerated a little, for effect, by the way . . . by about a million! But what you can see is that a wave runs through the basilar membrane, the amplitude of the wave peaking at one place, then falling off sharply.

If you run a sine wave through the ear, then look closely at the basilar membrane, the place where the amplitude peaks depends on the frequency. The basilar membrane turns out to be built to do just that; it varies in thickness and in stiffness along its length, just so different frequencies can peak at different places. Almost like it was trying to find frequency, huh? This is called the tonotopic organization of the cochlea -- organization by tone.

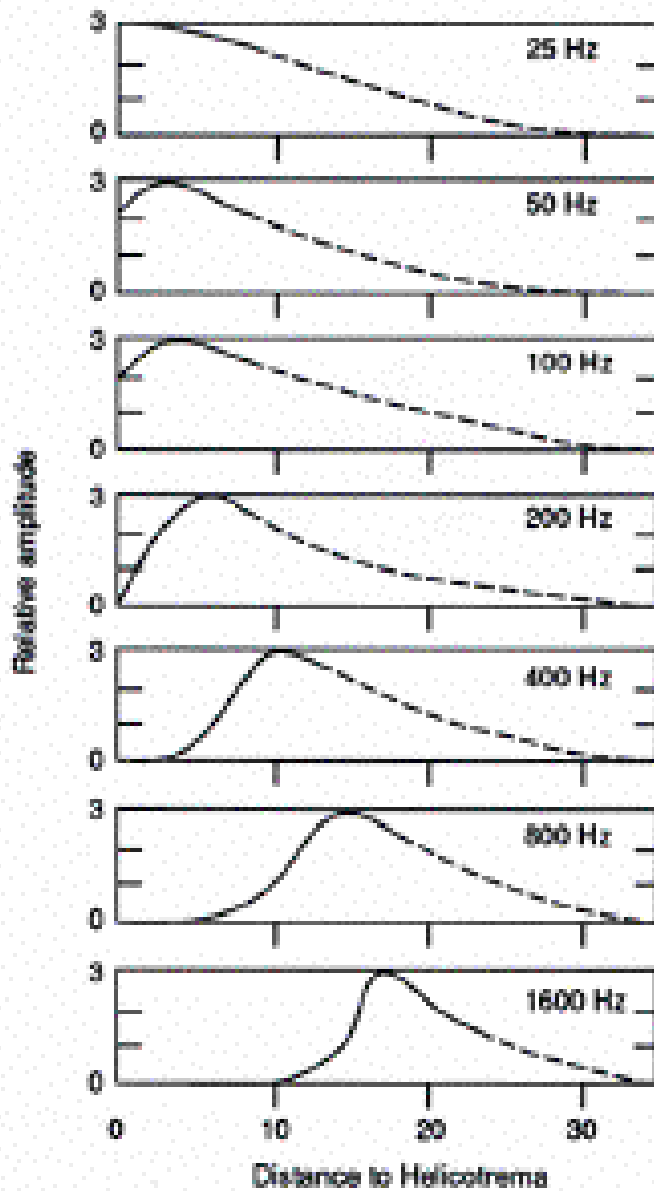
Here's a picture of some frequencies and their envelopes, with peaks:



One of the really neat things, if you look at the frequencies listed: as you move from the helicotrema to the ear, the frequencies double --- from 25hz to 50 to 100 to 200. BUT -- the place of maximum displacement does not double. Instead, these doubled frequencies are spaced apart equally.

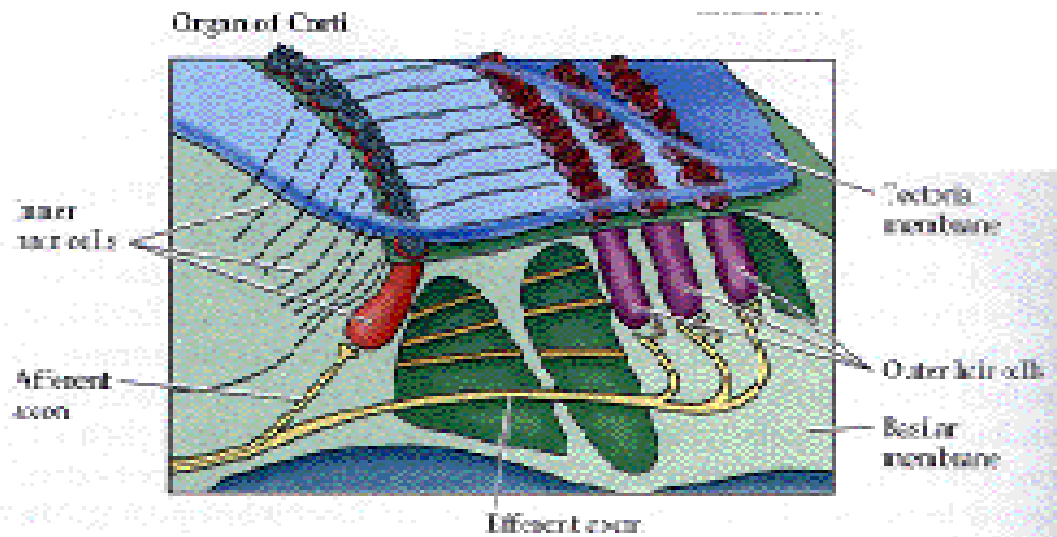
That kind of spacing is logarithmic . . . the cochlea is organized logarithmically. This is the reason that musical scales are organized into octaves; the phrase "go up an octave" means "double your frequency". But it isn't heard as a doubling, just equal spacing. You can check this out in the labs.

Here's more of the same thing:

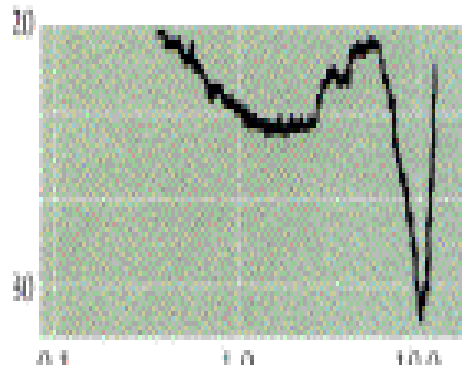


What you see here is a wave traveling down the basilar membrane, and, while the amplitude is going to peak at some place, still the wave will be setting off the hair cells all along the basilar membrane. Because the wave is too spread out, you think it'd be hard for the cochlea to figure out exactly what frequency the wave is.

And, left all to themselves, it would be. But remember the details of the organ of Corti



The individual hair cells join together to form the nerve fibres leading out of the cochlea. And when you test these nerve fibres, you see that they won't conduct an electrical impulse for any old wave coming down the basilar membrane. Each fiber is highly tuned; it conducts the electrical impulse only for a very narrow range of frequencies. Here's a sample:



The spike peaks downward instead of upward, showing how little energy is required to make the nerve conduct - if that energy is focused in just the right band of narrow frequencies.

OK . . . that looks like it: a pressure wave moves the basilar membrane, and neurons fire just exactly where the wave peaks, which corresponds to the frequency -- the log of the frequency, anyway.

There's some problems with this picture. You'll notice it's all about frequency; doesn't say a thing about amplitude, how the ear recognizes and encodes amplitude. Big problem.

Second problem: look back at the picture on the previous page. While each frequency peaks at its own special place, the peak is spread around a bit. This means that the cochlea can't pin down the exact frequency; there's going to be a little spread, a little uncertainty. The amount of uncertainty is called the critical band around the frequency. There's even a nice empirical formula for predicting it, in hertz:

$$\text{critical bandwidth} = 24.7(.00437[\text{frequency}] + 1)$$

So, at 220hz, you get a critical band of about 48.5 hz. But of course the ear doesn't expect to get just one frequency at a time; it gets many frequencies, and if the frequencies are close, you can expect the critical frequencies will overlap. And you can expect the ear will then have problems with the sounds.

## Beats

I want to start with basesound.wav, a 220hz sine wave; this is just here so you can recognize it when you hear it. Where you're supposed to recognize it in is sounds.wav, which is the 220hz frequency played against a mix of gradually higher frequencies. You hear the base sine very clearly, as well as some science-fictiony higher frequency sounds.

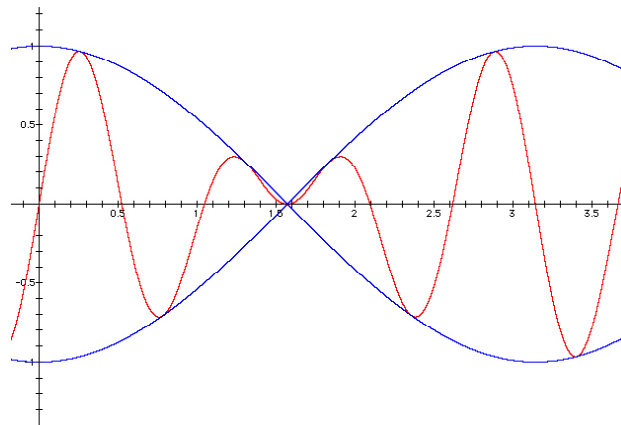
It's going to be important to remember: this is the way two frequencies together are supposed to sound. This is what the ear does, telling me which two frequencies are present. It's the ear-version of spectrograms.

Now play sounds1.wav. This consists of two sines, at frequencies 220hz and 221 hz, each played separately then played together. The difference 1hz is well inside the critical band, and you expect some distortion. And what you hear is . . . different. The ear doesn't hear two separate frequencies, and it doesn't hear one frequency blended together; it hears a combination sound. The combo sound is easy to explain mathematically:

$$\begin{aligned} \sin(A + B) &= \sin A \cos B + \sin B \cos A \\ + &= + \\ \sin(A - B) &= \sin A \cos B - \sin B \cos A \\ &= 2 \sin A \cos B \end{aligned}$$

So that  $\sin(\phi) + \sin(\psi) = .5 \cos\left(\frac{\phi - \psi}{2}\right) \sin\left(\frac{\phi + \psi}{2}\right)$

And here the sum is, graphed: the sine in red, the cos in blue.



In our case, the sum frequency is 200.5hz, the difference is .5hz. What you get is an envelope function: there's a sine playing at 200.5hz, but its amplitude is being controlled by the cosine. Although the cosine has frequency 1/2 hz, it goes through a max and a min .5 times a second. But when it forms an envelope, it moves the amplitude through a max twice as fast, at once per second. If you play sounds.wav, you can count out the dips.

These dips are called beats, and the 1 cycle per second is the beat frequency. Of course it all makes perfect sense, especially since you can explain it with a nice mathematical formula and a pretty graph. Makes it hard to remember that this is not supposed to be how sounds work. Beats, envelopes, beat frequencies . . . none of those are supposed to be audible. What you should be hearing is two simple sines, together. Remember that, in sounds.wav?

The fact that you don't hear two simple tones is a result of the non-linearity of the ear. Recall we said a system is linear if the presence of one sound doesn't affect how another sound is perceived. But as long as there are critical bands, the ear won't be completely linear.

The next thing to try is sounds2.wav: two sines, at 220hz and 225 hz. The difference frequency is 2.5, but again this produces an envelope modulating the frequency at 5 cycles per second. And you can hear the faster modulation if you play sounds2.wav.

The last experiment is in sounds3.wav, with sines at 220hz and 270 hz. They have critical bands 48.5hz and 53.4hz. This means that the sounds overlap, just a bit. And you can hear the overlap, just a bit, in the "fuzziness" of the resultant sound.

There are several morals to this story: first, the ear has a kind of uncertainty principle: the ear can't find frequency to 100% accuracy. Second, the ear is non-linear; two sounds can interact and produce a completely new and different sound. And third . . . if you're a musician, you would probably want to avoid writing parts where the critical frequencies overlapped. Because instead of some great harmony, you'd get a big "wubba-wubba" in your song. Tacky.

In fact, it turns out you can use frequencies and spectra to detect non-linearities. Say you have some non-linear function  $f(x)$  and you feed a frequency  $\sin(\omega t)$  into it. What comes out? I'll start with  $f(x) = x^2$ . Then  $f(\sin \omega t) = \sin^2(\omega t) = \frac{1}{2} [1 - \cos(2\omega t)]$ . I'll choose to write this as  $\frac{1}{2} [\cos(\omega - \omega)t - \cos(\omega + \omega)t]$ . The nonlinearity presents itself in the formation of new frequencies, sum and difference frequencies, that weren't in the original signal.

Now try inputting two frequencies . . . try the sum,  
 $f(\sin \omega_1 t + \sin \omega_2 t)$ . Squaring out gives me

$$\sin^2(\omega_1 t) + 2 \sin(\omega_1 t) \sin(\omega_2 t) + \sin^2(\omega_2 t)$$

The first and last terms give me the difference frequencies I already know; the middle term though . . . hmm!

$$\begin{array}{rcl} \cos(A + B) & & \cos A \cos B - \sin A \sin B \\ + & = & + \\ \cos(A - B) & & \cos A \cos B + \sin A \sin B \end{array}$$

So that

$$\sin(A) \sin(B) = \frac{1}{2} [\cos(A - B) - \cos(A + B)]$$

This means that the cross term  $2 \sin(\omega_1 t) \sin(\omega_2 t)$  will also generate sum and difference frequencies  $\omega_1 \pm \omega_2$ .

The very same thing happens with  $f(x) = x^3$ . Let's try a typical term of  $f(\sin \omega_1 t + \sin \omega_2 t)$  . . . say  $3 \sin^2(\omega_1 t) \sin(\omega_2 t)$ . We resolve the square as above, giving  $\frac{3}{2} [1 - \cos(2\omega_1 t)] \sin(\omega_2 t)$ .

I'm starting to run out of trig tricks. I think a cos times a sin would give me . . . ah,  $\cos(A) \sin(B)$  comes from  $\sin(A \pm B)$ , is it? Something like  $\frac{1}{2} [\sin(A + B) + \sin(A - B)]$ . Hence, our cross-term  $\frac{3}{2} [1 - \cos(2\omega_1 t)] \sin(\omega_2 t)$  yields terms like  $\sin(2\omega_1 + \omega_2)t$  and  $\sin(2\omega_1 - \omega_2)t$ . By the time you've sorted everything out . . .

When a nonlinearity gets frequencies  $\omega_1, \omega_2$  as an input, it gives frequencies  $k\omega_1 \pm j\omega_2$  as output. So you can detect non-linearities by the presence of sum and difference frequencies.

I bring up this whole non-linearity issue because . . . well, not to put too fine a point on it . . . the ear itself is non-linear. Say you play tones at frequencies  $\omega_2$ ,  $\omega_1$ , where  $\omega_2 > \omega_1$ . Then the ear always produces a false tone, at  $2\omega_1 - \omega_2$ , as well as  $(k + 1)\omega_1 - k\omega_2$ .

One example is to play frequencies 523.3hz and 660hz. You can check the combination tones are at 386.6, 249.9 and 113.3hz. For those with good musical ears, they should be audible, even though they aren't really "there".

## Masking

Everyone has her fave singer or movie star. I have my favorite nonlinearity (posh? ginger?). What makes it my favorite is that it's so built into our experience, that we don't even think there's anything unusual about it.

Here's the experience: go to a concert -- a nice loud one. You'll notice that it's hard to talk to the person next to you. Next go to a waterfall -- now it's just about impossible to talk to the person next to you. The sound of music can mask the sound of conversation, but pure noise can mask it even better.

It works in reverse too. If there's some annoying low-level street noises, you can turn on some music -- and the noise seems to go away: music can mask noise.

We're so used to this happening that we expect it: a loud sound is supposed to "drown out" quieter ones. But think about it. Say you're looking at a picture, and there's a brilliant red flower in the center. In fact, imagine it's so very red that you can't even see any of the rest of the picture.

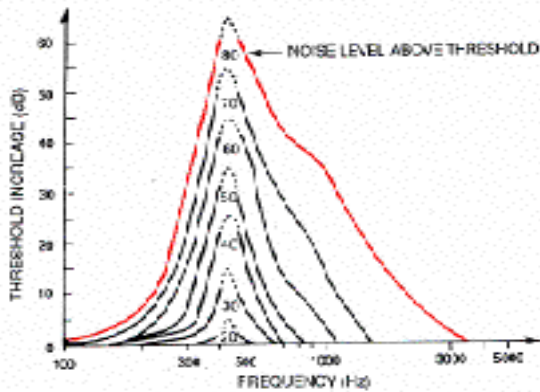
Doesn't happen, does it? However, this kind of thing does happen with the way we hear sound. It's a nonlinearity, in the technical sense: the presence of one sound affects how a second sound is perceived.

Masking plays an important role in something called MPEG Layer 3, also known as MP3. Online music, Napster . . . it's all based on masking. The idea is this: if a piece of music has two sounds, one masking the other, then you could skip the second sound entirely, and the music wouldn't sound any different. You could compress music.

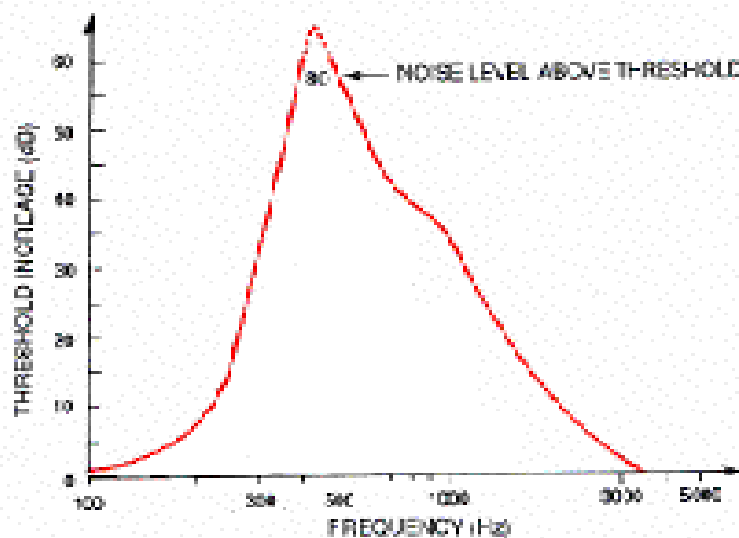
It turns out to be hard to skip sounds entirely, but what you can do is change them to make 'em just a little noisier. If the music is loud enough, it will mask that extra noise.

So it's worth spending a little time talking about masking. At the very least, anyone designing a music compression scheme would need to know how loud the music has to be, to mask noise. Or, how loud you can allow the noise to get.

Here's a typical "masking curve", from a book on hearing:



It's a nice, happy, complicated collection of curves. Let me simplify it a bit, then talk about what it means.



OK. First of all, the horizontal axis measures frequency, and right away you see there's a peak frequency. This is the frequency of the very loud sound we're going to play, the one that's supposed to drown out all the other sounds. The sound known on the street as "the masker." And secretly you're being told that the masker is a simple sine wave, with the frequency . . . umm, it looks about 500hz to me (actually 415hz).

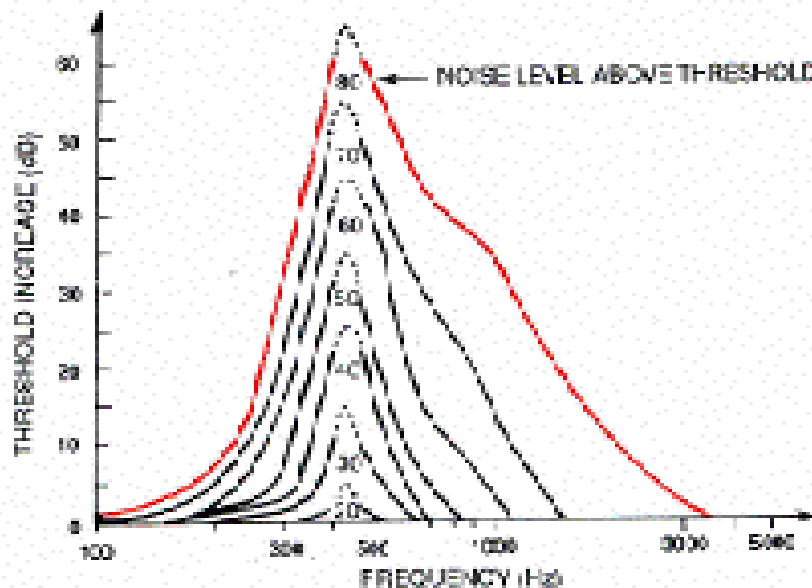
Given it's a sine wave, the only other thing I need to know about it is amplitude. Measured in decibels, like all good amplitudes are measured in this course. That's the extra number written next to the masker: 80 decibels "noise level above threshold," as the picture says.

Now “noise level above threshold” is not a unit I’m used to talking about. I’m used to decibels SPL =  $20\log(A/.000002)$  -- remember Section 3. “Threshold” tells you that the reference level (the denominator) for the decibels is not .000002, but “threshold,” that is, the SPL needed so you can just hear the 415hz sound.

All that is about the peak point of the curve. The rest of the curve is about the sounds that get drowned out. Again, the masked sounds are modeled really simplistically: they’re sine waves, with just a frequency and an amplitude. The frequency is on the horizontal axis, the amplitude on the vertical. And again the amplitude is measured with units that I haven’t seen before: “threshold increase.” This measures how loud the masked frequency must be, before you can even hear it over the sound of the masking frequency. And the reference level here is how much louder the masked sound has to be, compared to the threshold( if the masking frequency weren’t there at all).

This model of masking says that the effect of the masker is to raise the threshold required before you can hear the masked sound. It’s kind of like my intuitive idea about one sound drowning out another. I want to know how much louder I need to talk, to be heard over the sound of a concert; that’s what the curve measures: not ‘how loud’ but ‘how much louder than normal.’ There’s more interesting things about the red graph, too. First, you can see that a loud masker really masks all kinds of frequencies . . . some serious masking going on, from about 300hz to over 1000hz.

You can also see that the masking curve isn’t symmetric; it masks higher frequencies more than lower frequencies. Lastly, if you put the red curve back into its original context:



This gives the masking curves for a number of different amplitudes of the masker. One thing it shows clearly is that the quieter the masker is, the fewer frequencies it masks. And that the quieter the masker is, the less you have to turn up the volume on the masked frequencies. It’s what we always hear happening, when one sound drowns out another. It does tell me that the curves are in touch with reality.

There's some more reality checks here. The curves show that one frequency will mask out frequencies close to it, but won't mask out every frequency. That's what I expect: a concert doesn't make conversation impossible, just difficult. And I bet . . . if I went to a concert where all the sounds were very high frequency -- say a bunch of violins -- I bet I could carry on a conversation more easily than if there were a rock band playing at the same volume. Interesting experiment? but it's the waterfall experience again. A waterfall is noise, which means that it has all kinds of frequencies in its spectrum. That means it will mask every other frequency and make conversation impossible.

So the curves aren't exactly rocket science; they just confirm my intuitions about how sounds work. OK, I'm making fun of the masking curves. But there are some issues that I don't expect, and don't have intuition about.

First, a loud sound can mask a quieter sound of a different frequency: I know that. But it can also mask a quieter sound played a tiny time afterward. This is called forward time masking. What's truly strange is that that a masker can mask a sound played a tiny time *before*. We also got backward time masking.

Whoa! What is this, time travel? No . . . it's another peculiarity of the way the ear processes sound. It tells us that the ear doesn't respond instantaneously to each sound, and send that on to the brain. precisely what the ear does do, is unknown at this point.

If you think of the ear as averaging the sound, a little, then sending the averaged information on to the brain, you can see how time masking could work. If you have a loud sound at the beginning, it'll mask what follows in the average, and you get forward time masking. If the loud sound is near the end, it masks what came before in the average, and you get backward time masking.

BUT . . . this is just a model; the ear could do different things. for example, it might accumulate information in blocks, then send that on. Or, maybe the ear identifies louder sounds faster than it processes and identifies quieter sounds. No-one really knows.

However, time-masking can turn out to be important in manipulating sound. It means that you have a little bit of leeway. If you introduce noise, there's a chance it might not get heard, provided it lasts a short enough time to be masked by louder parts of the signal.

Here's s'more: everything I've been saying is about how one sine wave masks another sine wave. But music isn't sine waves. What happens when the sounds start to get complicated?

Well, obviously why there's still the big dollars to be made in designing music compressors. However, there's one issue that's relevant to MPEG technology -- MPEG doesn't rely on music masking music; it relies on music masking noise.

So that's a block of experiments I ask you to look at in the labs. We've done "sine masks sine" curves. What do the "sine masks noise" curves look like?

Notes Say you have a  $\sin(100x)$  and a  $\sin(105x)$ , 100 cycles per second and 105 cycles per second. The two sounds will act like they're in and out of phase most of the time, hence they'll be reinforcing and canceling constantly. Now, Sundberg remarks that they reinforce five times a second, and cancel out five times a second. Therefore this creates an envelope effect, a big waving, five times a second. Play it, count out the beats.

It's interesting to do the analysis.  $\sin(100x)$  maxes/mins out when  $\cos(100x) = 0$  this gives a max at  $100x = 2k\pi$  and min at  $100x = (2k+1)\pi$ . For the two to re-inforce you have to have  $x = (2k\pi)/100 = (2j\pi)/105$  or  $k = (105/100)j$  with integer solutions  $k, j$ ! Rewrite as  $k = (1 + 5/100)j = k = (1 + 1/20)j$ . You'll get an integer everytime  $j$  is a multiple of 20 that is  $j = 20, 40, 60, 80, 100$ . Five times in one second.

This doesn't work with incommensurable frequencies like 111 vs 107, BUT the mathematics of cosine and sine do work. Each representation gives you a little something.

There's a detailed explanation of how the ear processes beats and distinguishes frequencies, in Roederer Physics and Psychophysics of Music 1994, page 23.

In discussing mpeg, play some badly compressed sounds: tones.wav and badtones.wav