Fundamentals of Hearing

A summary of phenomena, one view of mechanism, and considerations related to audio.

.

But first a Word on Audio and Acoustics.

The Origins of Audio

Audio, in the way we use it here, is the technology, science, and art of recording or creating something to be played back via transducers to the human auditory system. The musical issues are not included, and indeed warrant a separate and serious discussion by an expert in the subject.

There have been several varieties of recording and/or transmission methods, starting with none, i.e.:

Live Performance

Acoustic Recording of Live Performance

Electronic Recording of Complete Performance

Studio Recording of Individual **Sessions**

Direct Synthesis of Music

What do we hear in each of these formats or methods?

In the original venue, we hear whatever parts of the soundfield are present where we sit. As we move, turn our heads, move our heads, and focus our attention, we can learn a great deal about the concert hall, the performers, and everything else perceptible from the seat we are sitting in.

No method of recording to the present provides us with anything near this experience.

WHAT DID I JUST SAY?

Well, it’s not all bad news. There are things that both older and newer recording methods do very well. What are they?

1) Capture, at one or more points, the pressure, velocity, or 3d pressure/velocity of the sound at a given point in the atmosphere.

2) Store that in a secure, accurate, and easily reproduced and/or transmitted form.

3) Reproduce/amplify that recorded (set of) pressures, velocities, and such as electrical analogs.

What do we do an “ok” job with?

In a word, loudspeakers. While it’s no doubt that loudspeakers are far and away the weakest link in the array of equipment, they are still not too bad. They can reproduce most of the dynamic range, over most of the frequency range, and (if you spend your money wisely) at a level of distortion (in the waveform sense) that is at least sufferable, if not perfect.

What loudspeakers can’t do, however, is replace what was lost at the original acquisition, or replace what was never extant in the recording to begin with. Unfortunately, loudspeakers are very often called upon to do exactly that.

What are we missing?

In a real venue, the soundfield is very complex, and we sample it as we move and turn our heads. The soundfield in any reverberant performing venue will
change in less than one wavelength at any given frequency. That is to say, at 1 kHz, 1 foot, give or take, will yield a substantially different soundfield. In venues with highly dispersive reverberation (which is an ideal), in fact, the coherence length of a signal may be substantially less than a wavelength under some important and realistic conditions.

In such a diffuse soundfield, there is an enormous amount of analytic information present in the 1 meter space about one's head. Speaking purely from an analytical viewpoint, one would have to spatially sample the soundfield according to the Nyquist criterion, and capture an enormous number of channels.

For example, one would have to tile an imaginary sphere with microphones every .25" or so in order to correctly sample the surface of such a space at 20kHz. This sort of calculation leads directly to a catastrophic growth of number of channels and data rate.

What kinds of things will such a hypothetical microphone array pick up?

1) A set of coherent, plane/spherical waves passing through the space. (We will call these the **perceptually direct** parts of the soundfield for reasons that will be obvious later.)

2) A set of more or less uncorrelated (beyond the coherence length at any given frequency) waveforms from each of the microphones. These mostly uncorrelated microphones capture the *details* of the soundfield. (For reasons that will be obvious later, we will call these **diffuse** or **perceptually indirect** parts of the soundfield. The two terms are not *quite* synonyms.)

Basics of Hearing

Basic Hearing Issues

- Parts of the ear.
- Their contribution to the hearing process.
- What is loudness?
- What is intensity?
- How “loud” can you hear?
- How “quiet” can you hear?
- How “high” is high?
- How “low” is low?
- What do two ears do for us?

Fundamental Divisions of the Ear

- Outer ear:
  - Head and shoulders
  - Pinna
  - Ear Canal
- Middle Ear
  - Eardrum
  - 3 bones and spaces
- Inner ear
  - Cochlea
    - Organ of Corti
      - Basilar Membrane
      - Tectoral Membrane
      - Inner Hair Cells
      - Outer Hair Cells

Functions of the Outer Ear

- In a word, HRTF’s. HRTF means “head related transfer functions”, which are defined as the transfer functions of the body, head, and pinna as a function of direction. Sometimes people refer to HRIR’s, or “head related impulse responses”, which are the same information expressed in the time, as opposed to frequency, domain.

- HRTF’s provide information on directionality above and beyond that of binaural hearing.
- HRTF’s also provide disambiguation of front/back, up/down sensations along the “cone of confusion”.

- The ear canal resonates somewhere between 1 and 4 kHz, resulting in some increased sensitivity at the point of resonance. This resonance is about 1 octave wide, give or take.

Middle Ear

- The middle ear acts primarily as a highpass filter (at about 700 Hz) followed by an impedance-matching mechanical transformer.
- The middle ear is affected by muscle activity, and can also provide some level clamping and protection at high levels.
  - ***You don’t want to be exposed to sound at that kind of level.***

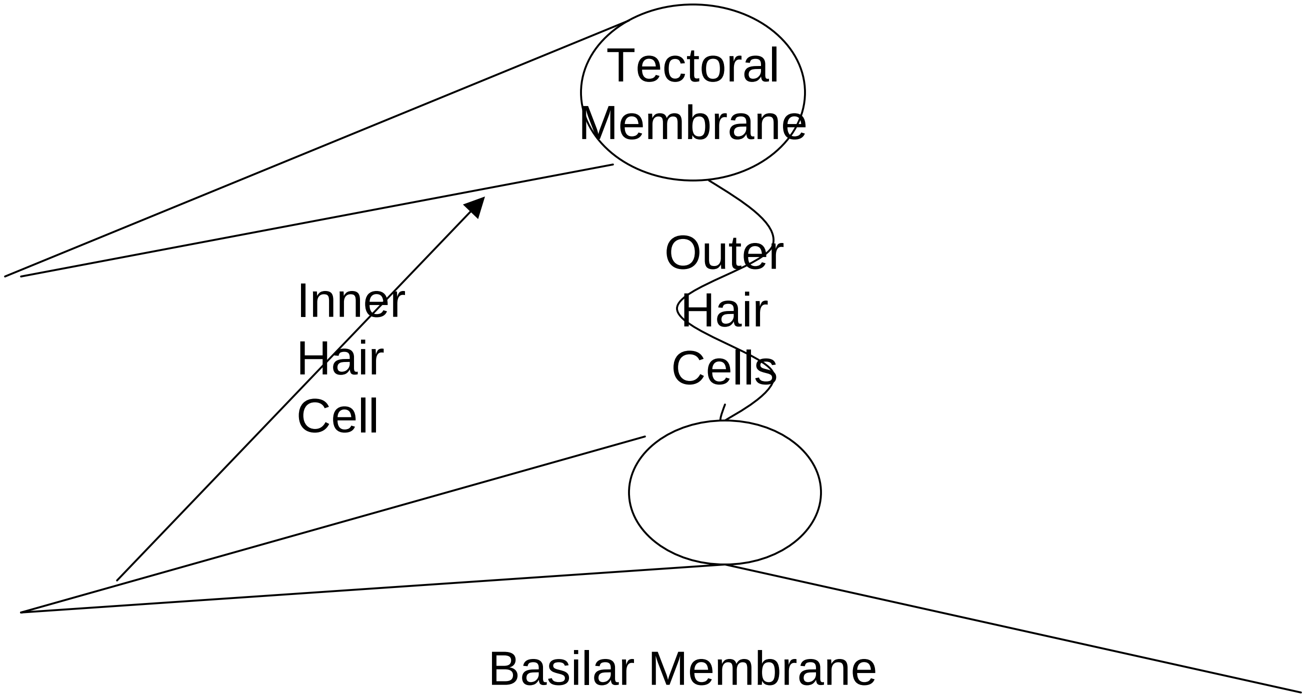
Inner Ear (Cochlea)

- In addition to the balance mechanism, the inner ear is where most sound is transduced into neural information.
- The inner ear is a mechanical filterbank, implementing a filter whose center-frequency tuning goes from high to low as one goes farther into the cochlea.
  - The bandpass nature is actually due to coupled tuning of two highpass filters, along with detectors (inner hair cells) that detect the difference between the two highpass (HP) filters.

Some Web Views of the Basilar Membrane and Cochlear Mechanics.

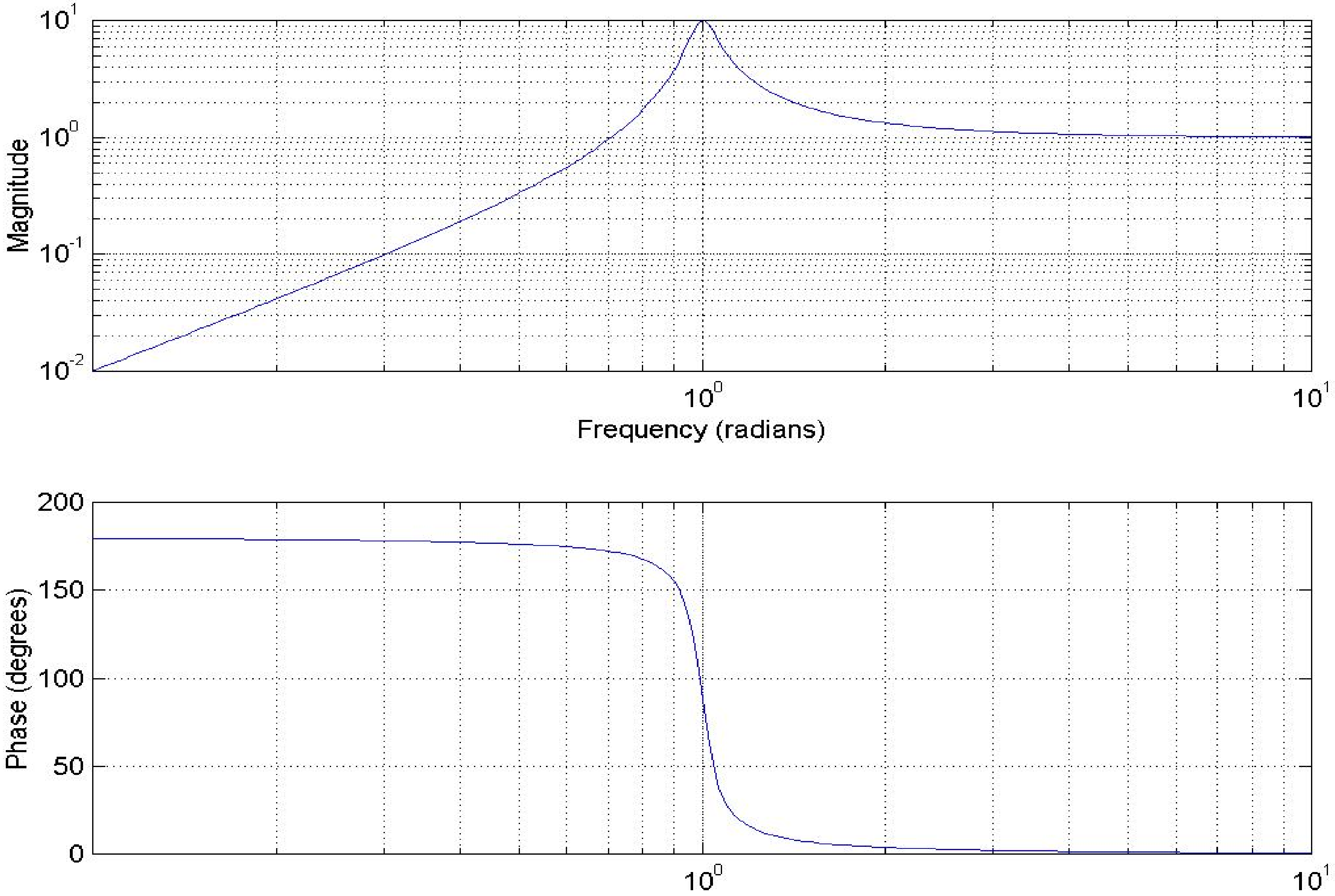
- http://www.medizin.fu-berlin.de/klinphys/topics/ear2_anatomy_e.htm
- http://www.enchantedlearning.com/subjects/anatomy/ear/

One point along the Membranes:



This View is Controversial (As are others)

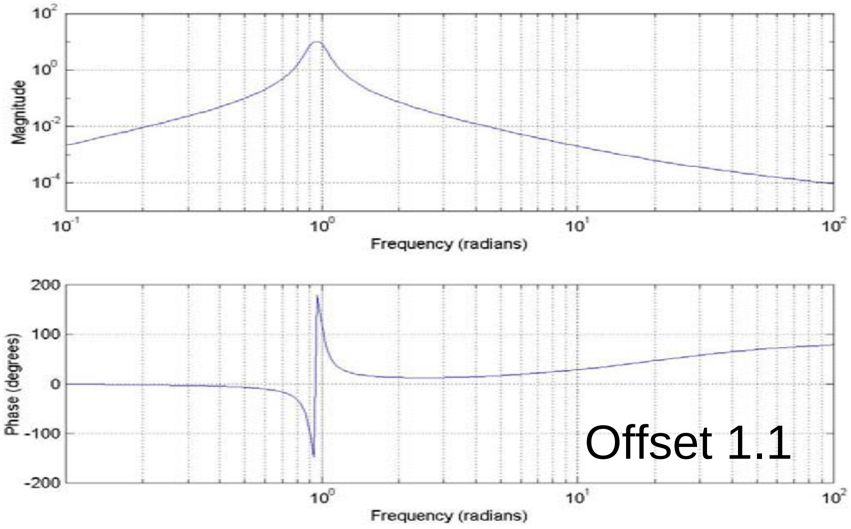
Example HP filter
(This filter is synthetic, NOT real)



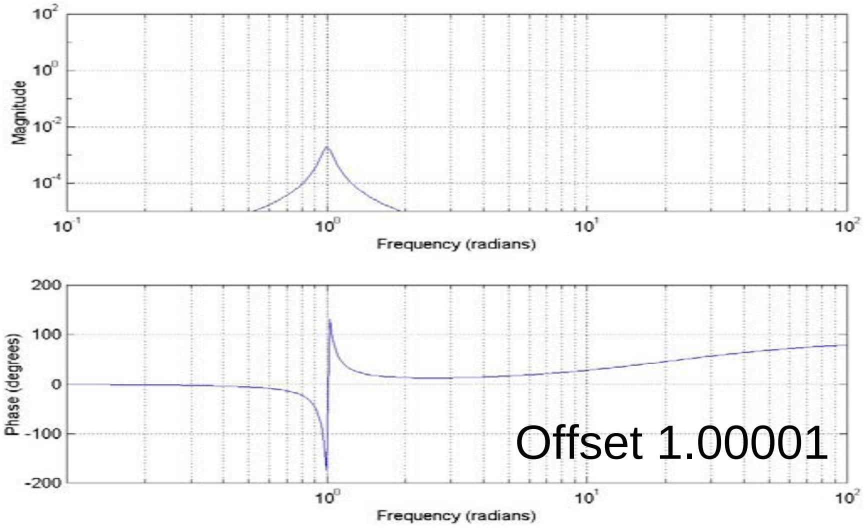
Features of a HP filter

- At the frequency where the amplitude is greatest, the phase is changing rapidly.
  - This means that two filters, slightly offset in frequency, will show a large difference between the two center frequencies, providing a very big difference in that region.
- When two nearly-identical filters are coupled, their resonances “split” into two peaks, slightly offset in frequency.
- As the coupling decreases, the two resonances move back toward the same point.
- The ear takes the difference between two filters.

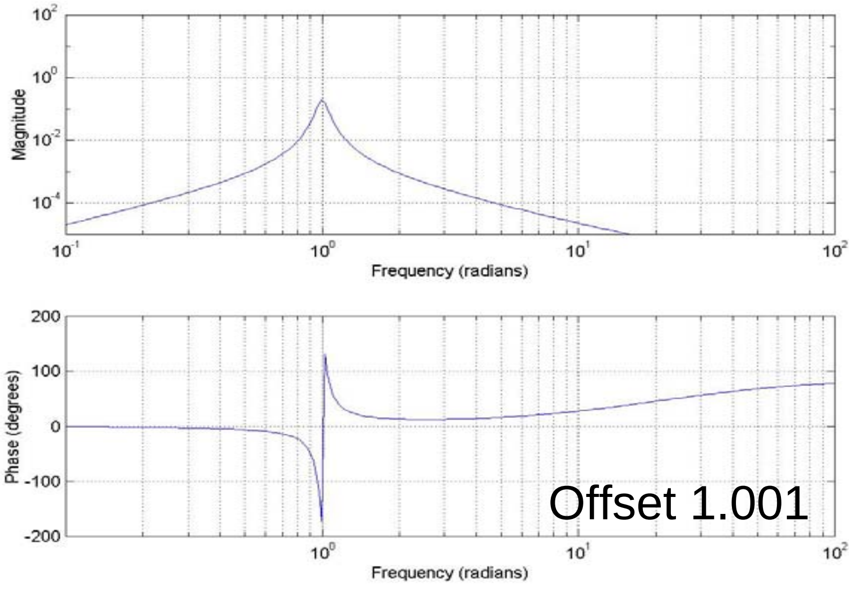
Filter split vs. Frequency Response



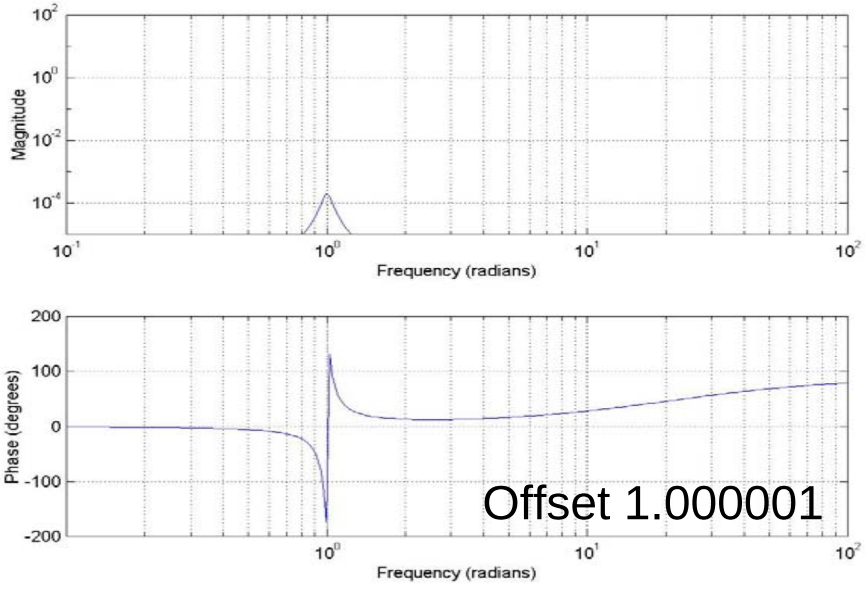
Offset 1.1



Offset 1.00001



Offset 1.001



Offset 1.000001

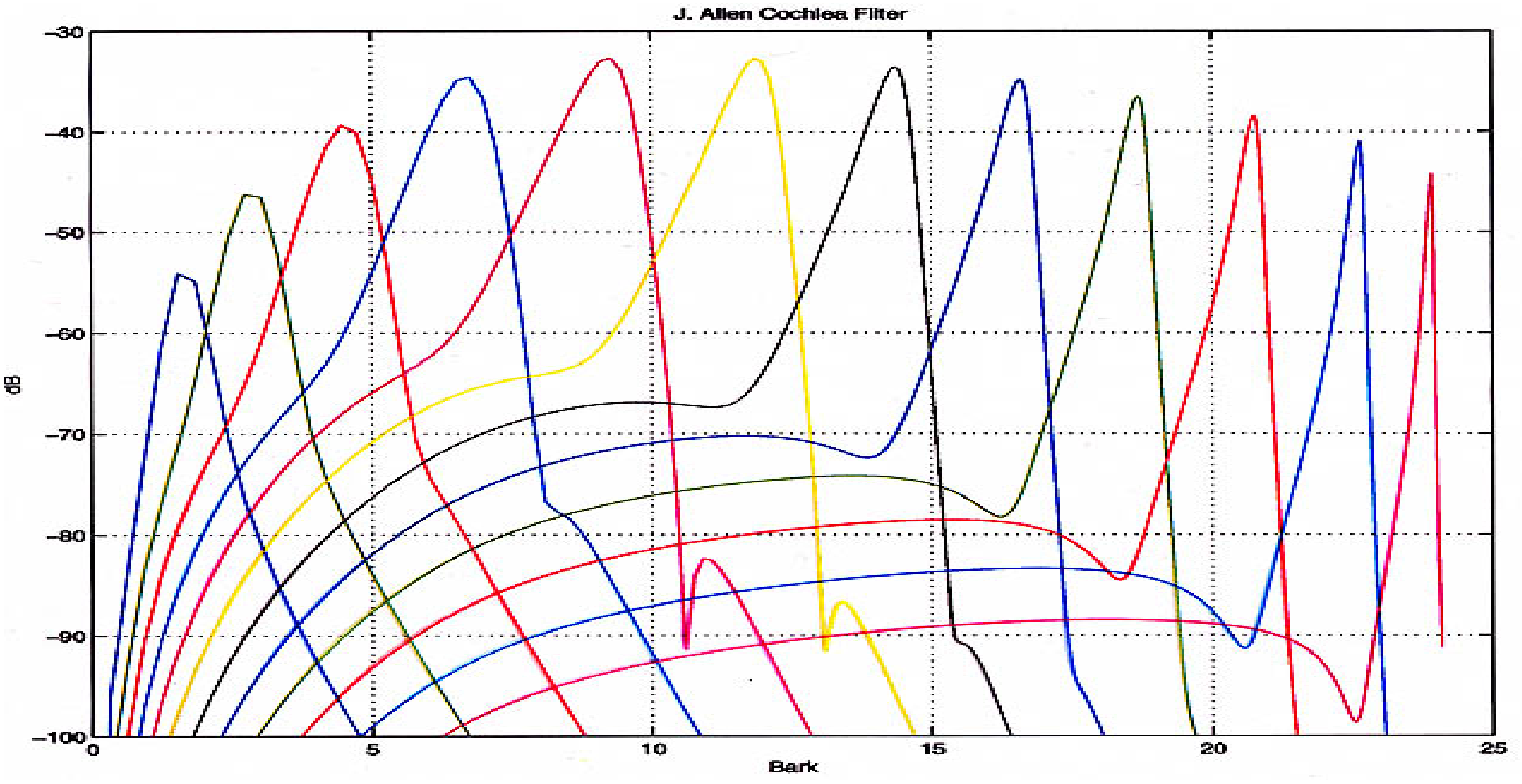
Ok, what couples these two masses?

- The outer hair cells couple the basilar and tectoral membranes.
  - At high levels, the outer hair cells are depolarized by a feedback from the inner hair cells. Depolarized cells are considerably less stiff than polarized cells.
  - At low levels, the outer hair cells remain (mostly) polarized. This maximizes the coupling, and therefore the system sensitivity.

But, but… Inner hair cells

- There are calcium channels at the base of the “hairs” (really more like villi in biochemical terms) that open and close as the “hairs” are flexed sideways.
- A very VERY small amount of flexure can therefore trigger an inner hair cell.
- The inner hair cells are the **Detectors.**
- At low frequencies, they detect the leading edge (membranes moving closer to each other) of the waveform. (Up to 500 Hz).
- At high frequencies, they detect the leading edge of the envelope. (Above 2000Hz or so).
- At frequencies between 500Hz and 2kHz or so, the detection is mixed.
- The higher the level of motion on the basilar membrane, the more often the nerve cells fire after the leading edge.

A set of filters at different points along the basilar membrane.



From Dr. J. B. Allen

Critical Bandwidths

- The bandwidth of a filter is referred to as the “Critical Band” or “Equivalent Rectangular Bandwidth” (ERB).
- ERB’s and Critical Bands (measured in units of “Barks”, after Barkhausen) are reported as slightly different.
- ERB’s are narrower at all frequencies.
- ERB’s are probably closer to the right bandwidths, note the narrowing of the filters on the “Bark” scale in the previous slide at high Bark’s (i.e. high frequencies).
- I will use the term “Critical Band” in this talk, by habit. None the less, I encourage the use of a decent ERB scale.
- Bear in mind that both Critical Band(widths) and ERB’s are useful, valid measures, and that you may wish to use one or the other, depending on your task.
- There is no established “ERB” scale to date, rather researchers disagree quite strongly, especially at low frequencies. It is likely that leading-edge effects as well as filter bandwidths lead to these differences. The physics suggests that the lowest critical bands or ERB’s are not as narrow as the literature suggests.

What are the main points?

- The cochlea functions as a mechanical time/frequency analyzer.
  - Center frequency changes as a function of the distance from the entrance end of the cochlea. High frequencies are closest to the entrance.
  - At higher center frequencies, the filters are roughly a constant fraction of an octave bandwidth.
  - At lower center frequencies, the filters are close to uniform bandwidth.
  - The filter bandwidth, and therefore the filter time response length varies by a factor of about 40:1.

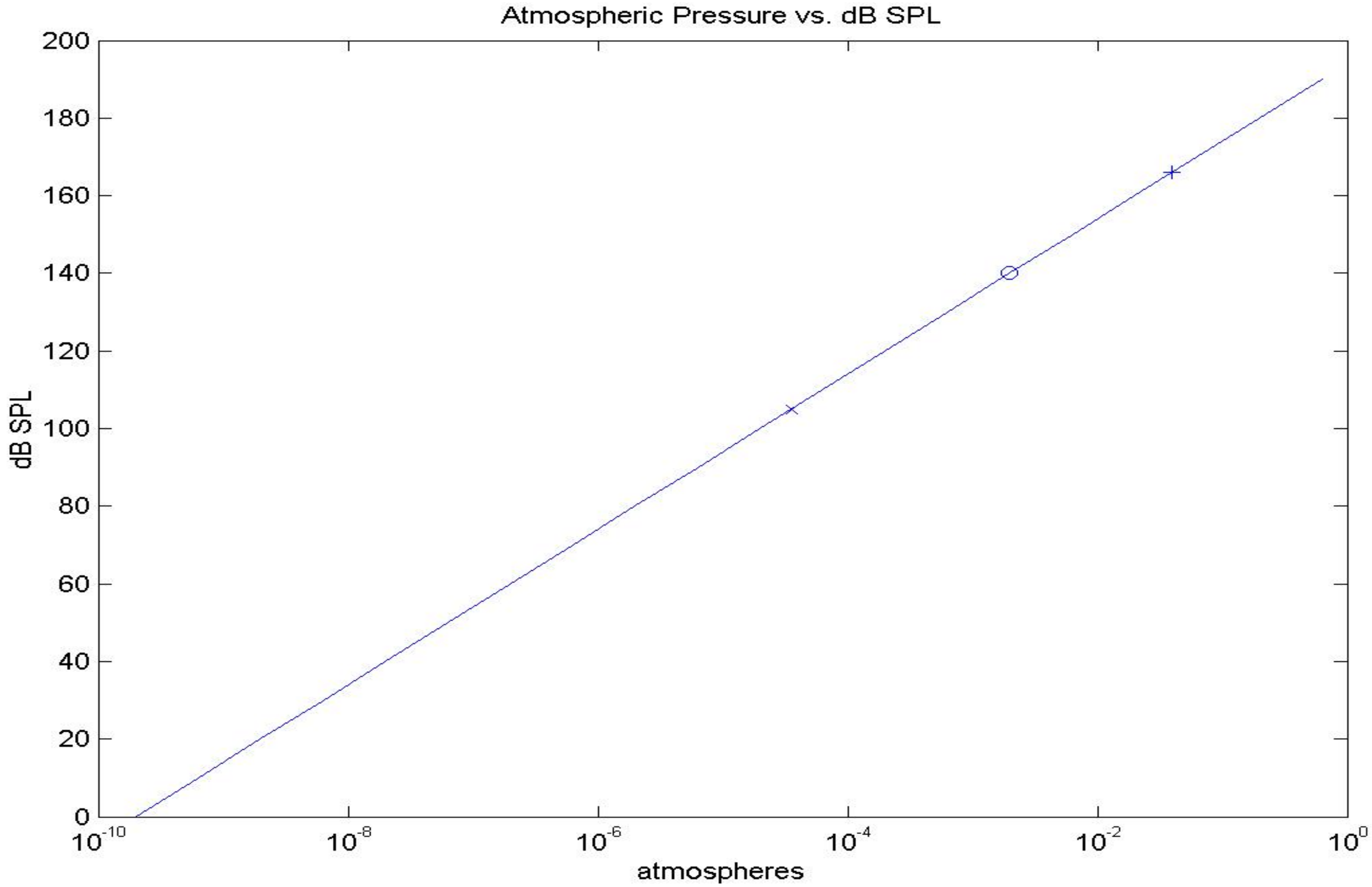
What happens as a function of Level?

- As level rises, the ear desensitizes itself by many dB.
- As level rises, the filter responses in the ear shift slightly in position vs. frequency.
- The ear, with a basic 30dB SNR (1000^.5) in the detector, can range over at least 120dB of level.

What does this mean?

- The internal experience, called *Loudness*, is a highly nonlinear function of level, spectrum, and signal timing.
- The external measurement in the atmosphere, called *Intensity*, behaves according to physics, and is somewhat close to linear.
- The moral? There’s nothing linear involved.

Some points on the SPL Scale



(x is very loud pipe organ, o is threshold of pain/damage, + is moderate barometric pressure change

Edge effects and the Eardrum

- The eardrum’s HP filter desensitizes the ear below 700Hz or so. The exact frequency varies by individual. This means that we are not deafened by the loudness of weather patterns, for instance.
- At both ends of the cochlea, edge effects lessen the compression mechanisms in the ear.
- The results of these two effects, coupled with the ear’s compression characteristics, results in the kind of response shown in the next slide.

Fletcher and Munson's famous "equal loudness curves".

Fig. 134.—Loudness Level Contours versus Sensation Level.

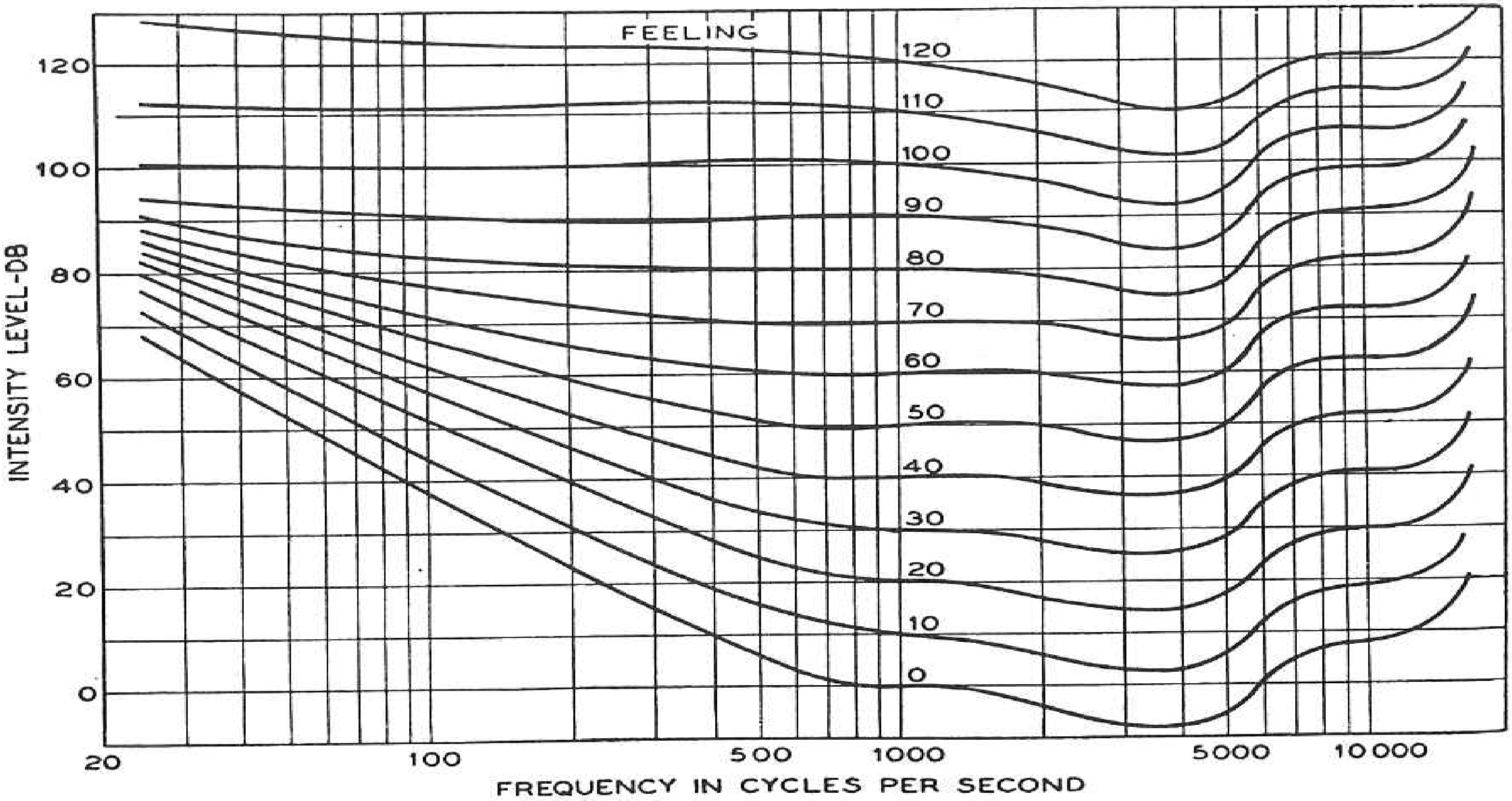


Fig. 135.—Loudness Level Contours versus Intensity Level.

The best one-picture summary of hearing in existence.

What’s quiet, and what’s loud?

- As seen from the previous graph, the ear can hear to something below 0dB SPL at the ear canal resonance.
- As we approach 120dB, the filter responses of the ear start to broaden, and precise frequency analysis becomes difficult. As we approach 120dB SPL, we also approach the level at which near-instantaneous injury to the cochlea occurs.
- Air is made of discrete molecules. As a result, the noise floor of the atmosphere at STP approximates 6dB SPL white noise in the range of 20Hz-20kHz. This noise may JUST be audible at the point of ear canal resonance. Remember that the audibility of such noise must be calculated inside of an ERB or critical band, not broadband.

So, what’s “high” and what’s “low” in frequency?

- First, the ear is increasingly insensitive to low frequencies, as shown in the Fletcher curve set. This is due to both basilar membrane and eardrum effects. 20Hz is usually mentioned as the lowest frequency detected by the hearing apparatus. Low frequencies at high levels are easily perceived by skin, chest, and abdominal areas as well as the hearing apparatus.

- At higher frequencies, all of the detection ability above 15-16 kHz lies in the very first small section of the basilar membrane. While some young folks have been said to show supra-20kHz hearing ability (and this is likely true due to their smaller ear, ear canal, and lack of exposure damage), in the modern world, this first section of the basilar membrane appears to be damaged very quickly by “environmental” noise.
- At very high levels, high frequency (and ultrasonic) signals can be perceived by the skin. You probably don’t want to be exposed to that kind of level.

The results?

- For presentation (NOT capture, certainly not processing), a range of 6dB SPL (flat noise floor) to 120dB (maximum you should hear, also maximum most systems can achieve) should be sufficient. This is about 19 bits. Some signal sources are much louder, and get into levels where air itself is substantially nonlinear. In light of what can be reproduced, capture of these signals that are in the nonlinear region represent an interesting problem.
- An input signal range of 20Hz to 20kHz is probably enough, there are, however, some filter issues that will be raised later, that may affect your choice of sampling frequency.

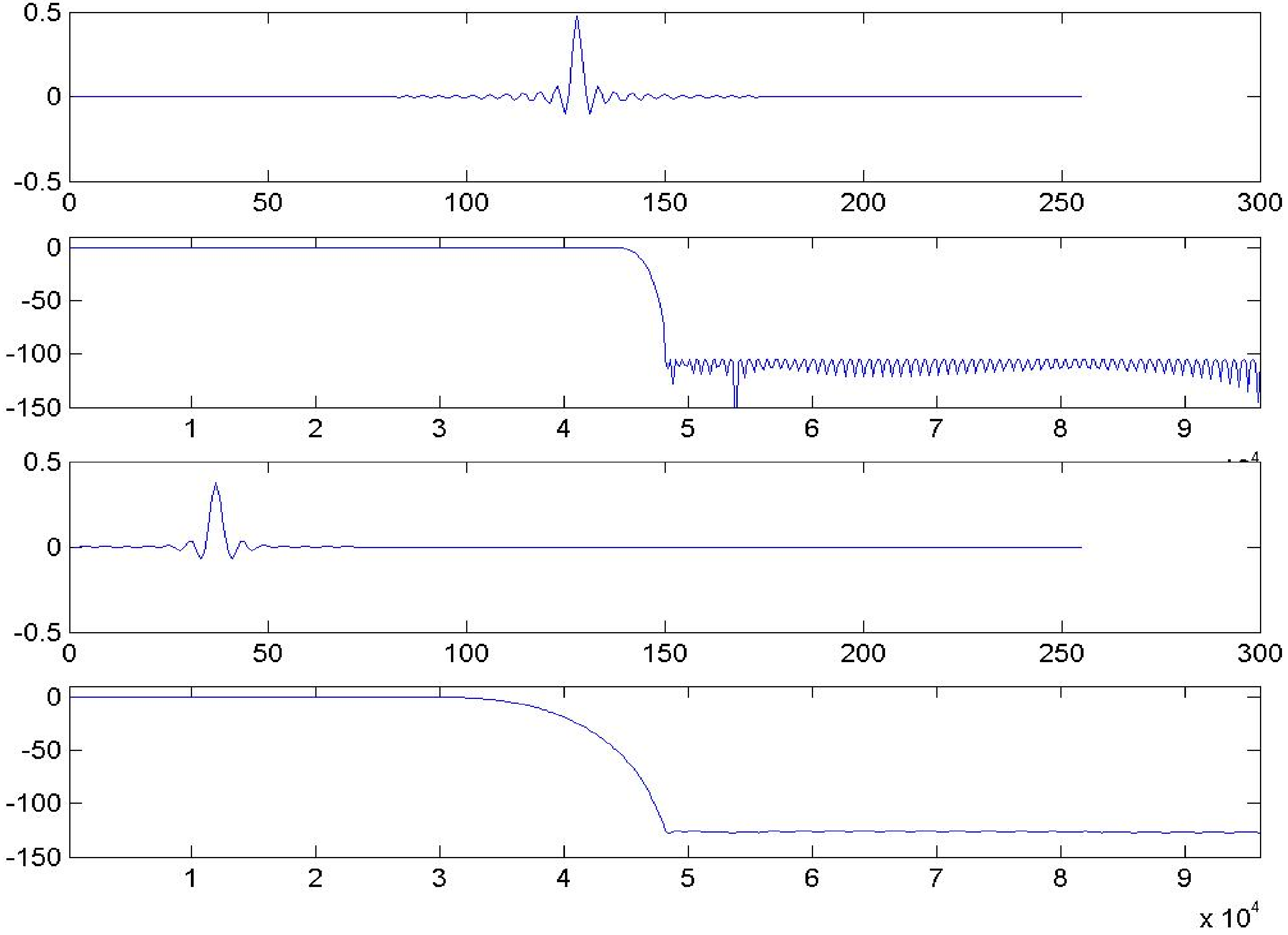
What’s this filter problem?

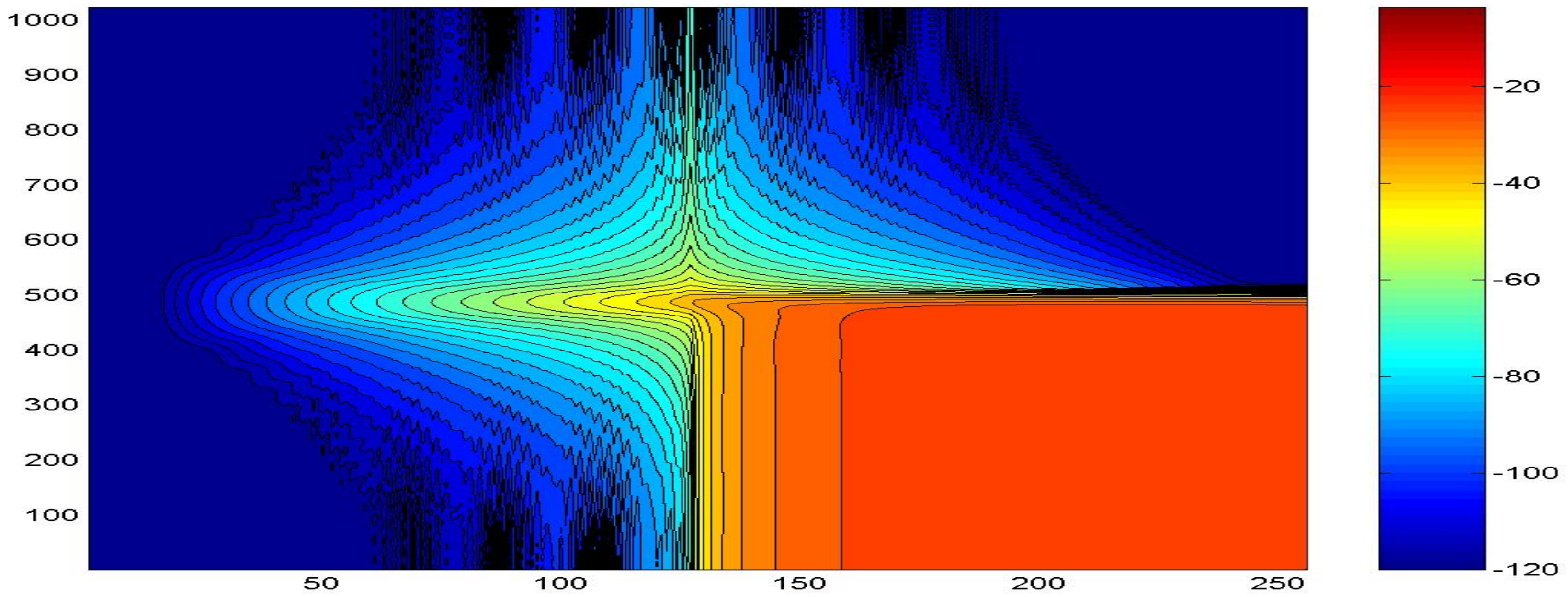
- Minimum phase filters have phase shift in the passband. This can change the envelope of the signal, and at some point, create audible effects.
- “Linear Phase” (constant delay) filters do not have this phase shift, however they have a pre-ringing. In extreme cases (some older rate convertors, audio codecs) this pre-ringing is clearly audible. Not all pre-echo is audible.

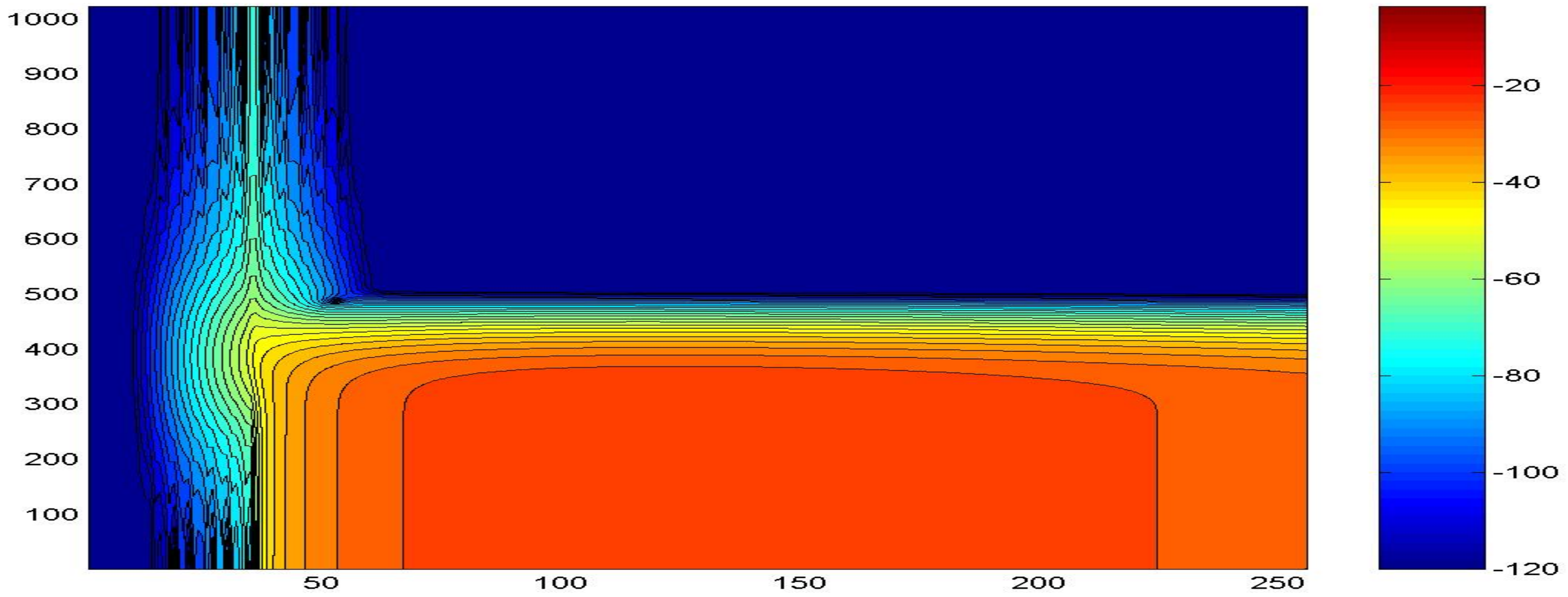
Remember the Filters in the ear?

- Your ear is a time analyzer. It will analyze the filter taps, etc, AS THEY ARRIVE, it doesn’t wait for the whole filter to arrive.
- If the filter has substantial energy that leads the main peak, this may be able to affect the auditory system.
  - In Codecs this is a known, classic problem, and one that is hard to solve.
  - In some older rate convertors, the pre-echo was quite audible.
- The point? We have to consider how the ear will analyze any anti-aliasing filter. Two examples follow.

Two examples:




1000
900
800
700
600
500
400
300
200
100
50
100
150
200
250
-20
-40
-60
-80
-100
-120


1000
900
800
700
600
500
400
300
200
100
50
100
150
200
250
-20
-40
-60
-80
-100
-120

When does this become audible?

- We frankly don’t know. I’ve watched hours and hours of argument over the issue of “IF” it can be audible, even though extreme systems do already show such effects.
- It’s hard to test. To be absolutely sure, we’d need a system with a gradual filter running at at least 2x the desired sample rate.
- To the best of my knowledge, nobody has run the sensitivity test to determine how much is audible, and at what frequency.
- It’s probably true that nothing is audible for a 96kHz converter cutting off between 40 and 48kHz, I would expect so, but we don’t know that yet.
- A sampling rate in the range of 50kHz makes the main filter lobe the same width as the shortest cochlear filter main lobe.

What would a 32 bit Uniform PCM Capture Imply?

- If we set the noise level of the capture to the level of atmospheric noise, that starts at 6dB SPL, give or take.
- Adding 192 dB, that gives us a “level” of 198dB SPL.
- 194dB is the theoretical level of a 1 Atmosphere RMS sine wave. This can not exist.
- 198dB, even if possible in a linear sense, which it is not, would represent 4dB over 1.414 atmospheres overpressure, or 2.24 atmospheres of overpressure. It would apply about ¾ tons to a 1 square foot window.

What do we detect?

- At low frequencies, it seems clear that the ear detects the waveform, and especially the cochlear-filtered “leading edges” very well. (Up to 500Hz)
- At high frequencies, the ear appears to detect the leading edges of the signal envelope (Above 2000Hz to 4000Hz)
- In the middle, the two mechanisms conflict.

Undoubtedly the reality is more complicated.

Binaural Effects in Hearing

Hey, we have two ears, not one.

Well, so what can one ear detect?

That would be, at low frequencies, the leading edges of the filtered signal itself, and at high frequencies, the leading edges of the signal envelopes.

The shapes of the waveform/envelope can also be compared, to some extent, as well as their onsets.

Interaural Time Differences

- The ear can detect the delay in arrival between the ears quite well. This translates directly into a “cone of confusion” that represents approximately one angle of arrival around the axis connecting the two ears.

Because we all live with our HRTF’s, that change slowly while we are growing, and very little afterwards (hair has some effect at high frequencies), the brain learns them, and we take them into account without any conscious effort.

In short, we know the delay at different frequencies between the two ears, the relative attenuation, and other such information as part of learning to hear binaurally.

This means that we can do several things.

Effects of HRTF/HRIR’s

- They allow us to disambiguate along the “cone of confusion”, because the interaural difference in frequency content will be different from different positions on the cone of confusion. Note, this mechanism is not perfect, and can be fooled, sometimes easily.

Detection of Direct (planar) Waves

Since we know the time delay at a given frequency, as well as the relative attentuations to the two ears as a function of frequency and direction, we can place a direct plane wave sound in direction by understanding at a very low level the time delays and amplitude attenuations at different frequencies for wideband signals with quick onsets. For such signals, having an idea of what the source "sounds like" is very helpful, as the spectral shaping at either ear can disambiguate the time-delay information, and provide more easily accessed elevation and front/back information.

Listening to Diffuse Soundfields

Diffuse soundfields will not have correlated envelopes at the two ears in the relevant frequency ranges. The brain interprets such waveforms as "surrounding us" or "omnidirectional". Note that these waveforms inside a critical bandwidth may or may not have leading edges. Note that leading edges can sometimes line up to provide false or confusing directional cues.

A perceptually indirect signal (a term applying only above 2kHz or so) will have a flat envelope, and thereby provide no information to correlate. In other words, the flat envelopes will in fact 'correlate' but the auditory system has no features to lock onto to determine either direction or diffusion. Such signals are often *ignored* in the most complex stages of hearing, but will have the same effect of "surround" or "omnidirectional' when focused upon.

Interaural Level Difference

- ILD can also, especially at high frequencies, create a sense of direction, due to our innate understanding of HRTF’s.
- ILD can sometimes also provide contradictory cues, when time cues do not match level cues.

So, what do we get in terms of spatial cues?

- Interaural Time Differences
- Interaural Level Differences
- Timbre of diffuse reverberation
- Differences between direct and reverberant timbre
- Direct/reverberant ratio

- Note that specular reflections (and long-period reflections, i.e. echoes) can act like secondary sound sources.

Some notes on subwoofers

- Various people have reported, sometimes anecdotally, that above 40Hz (and below 90Hz), although one can not localize a sound source, differences in interaural phase can create a sensation of space.
- This suggests that for accurate perception of space, 2 or 3 subwoofers may be necessary.
- This also, as in many other places in audio, creates a situation where what one might consider the “optimum” solution (maximum bass flatness) does not in fact convey the perceptual optimum.

A summary:

- From approximately 0dB SPL to about 120dB SPL, the ear has well-behaved frequency distinguishing characteristics.
- The 30dB of the transducer (inner hair cell) is spread approximately over this 120 dB.
- At low and high frequencies, the compression is less at low levels, and the thresholds elevated.

THE EAR IS A TIME/FREQUENCY ANALYZER

- This creates problems:
  - Systems that introduce nonlinearities (new frequencies create new perceptions, usually a bad thing) can sound strange.
  - The time aspects of this analysis, as well as the frequency aspects, must absolutely be considered.
  - The analysis is very lossy, but in a fashion not usually described by LMS algorithms.
    - Predicting what is heard, and what is lost, is not a simple matter.
    - The interaction between this lossiness and LMS makes things like THD, IMD, and SNR “mostly useless”.

- The range of 20Hz to 20kHz is a reasonable choice for almost all human subjects.
  - The filtering methods used must be carefully examined.
  - Children and youngsters may be an exception to this observation.
  - The atmosphere itself isn’t that good at transmitting high levels (it’s nonlinear and dispersive) or high frequencies.

- The level range of 6dB SPL to 120 dB SPL is probably enough for reproduction.
  - It is very unwise to (repeatedly) expose the listener to anything louder
  - The nonlinearity of air becomes an interesting and difficult to accommodate factor above this level.
  - How to compress things like rimshots into this range is also an interesting question.
  - There’s no point in capturing sound with a noise floor that will be presented as lower than the air noise at the eardrum.

- That doesn’t mean 20 bits are enough, though.
  - 20 bits are probably sufficient for safe reproduction at home, or in most any venue.
  - Given the noise in the modern world, 16 bits is probably sufficient in most places.
  - 24 bits is not an unreasonable dynamic range for capture, since the level of presentation to the listener may be substantially changed during production.
  - Signal processing algorithms may require many more than 16 or 24 bits in order to provide the requisite 20 bits or 16 bits to the ear in the final result.

A Final Warning

- The actions and interactions regarding the inner and outer hair cells are a subject of ongoing dispute. The mechanism presented here is only one such proposed mechanism. The predictions of this mechanism provide a good measure of the performance of the ear. That does not mean that the mechanism is right.
- You don’t regrow inner or outer hair cells.

TAKE CARE OF YOUR EARS!