Soundfields: Acoustics vs. Human Hearing

James D. Johnston
home.comcast.net/~retired_old_jj
j.d.johnston@ieee.org

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 substantially 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) 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 in the 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.)

How do our ears work?

And what can we detect in a natural soundfield?

How One Ear Works – Short Form!

(This being the short-form of what should occupy a semester's examination and discussion.)

The ear is usually broken into 3 separate parts, the outer, middle, and inner ears. The outer ear consists of the head, the pinna, and the ear canal. The middle ear consists of the eardrum, the 3 small bones, and the connection to the cochlea. Finally, the inner ear consists of the cochlea, containing the organ of corti, basilar membrane, tectoral membrane, and the associated fluids and spaces.

The Outer Ear

The outer ear provides frequency directivity via shadowing, shaping, diffraction, and the like. It is different (by enough to matter) for different individuals, but can be summarized by the "Head Related Transfer Functions" (HRTF's) or "Head Related Impulse Responses"(HRIR's) mentioned in the literature, at least on the average or for a given listener.

The HRTF's or HRIR's are ways of determining the effect on a sound coming from a given direction to a given ear.

The ear canal inserts a 1 octave or so wide resonance at about 1 to 4 kHz depending on the individual.

The Middle Ear

The middle ear carries out several functions, the most important of which, for levels and frequencies that are normally (or wisely) experienced, is matching the impedence of the air to the fluid in the cochlea.

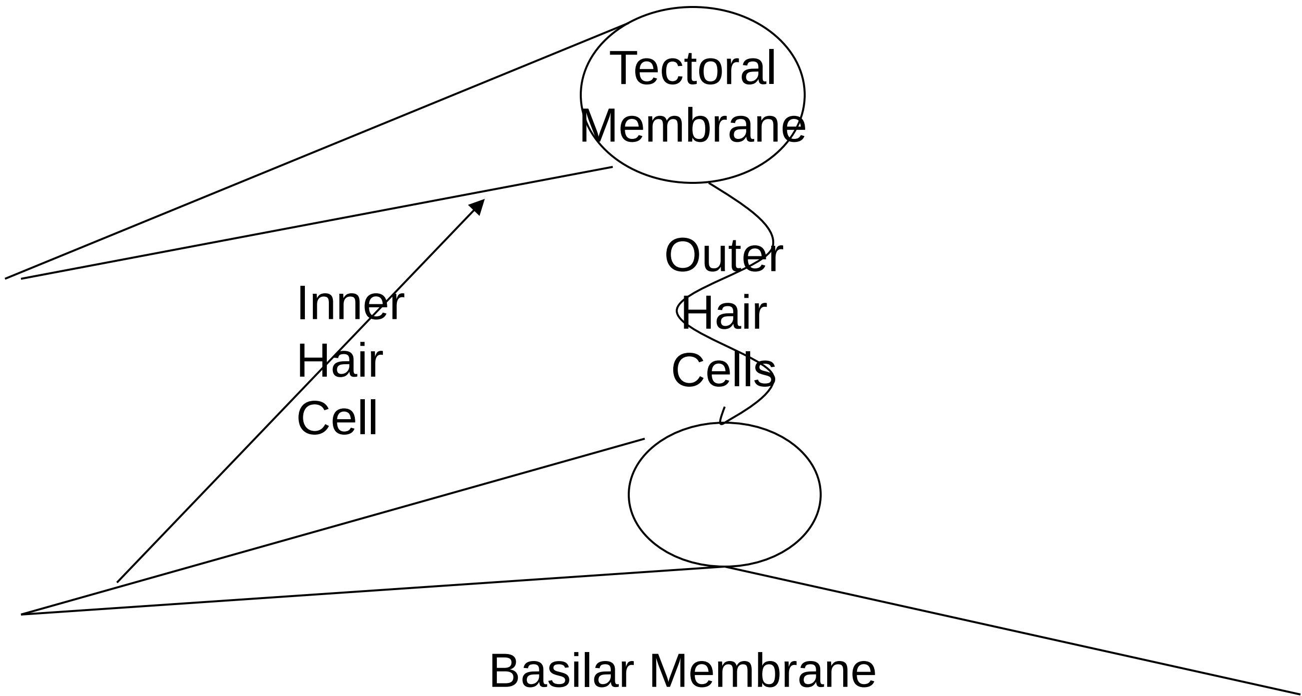
There are several other functions related to overload protection and such, which are not particularly germane under comfortable conditions.

The primary effect of the middle ear is to provide a 1-zero high pass function, with a matching pole at approximately 700Hz or so, depending on the individual.

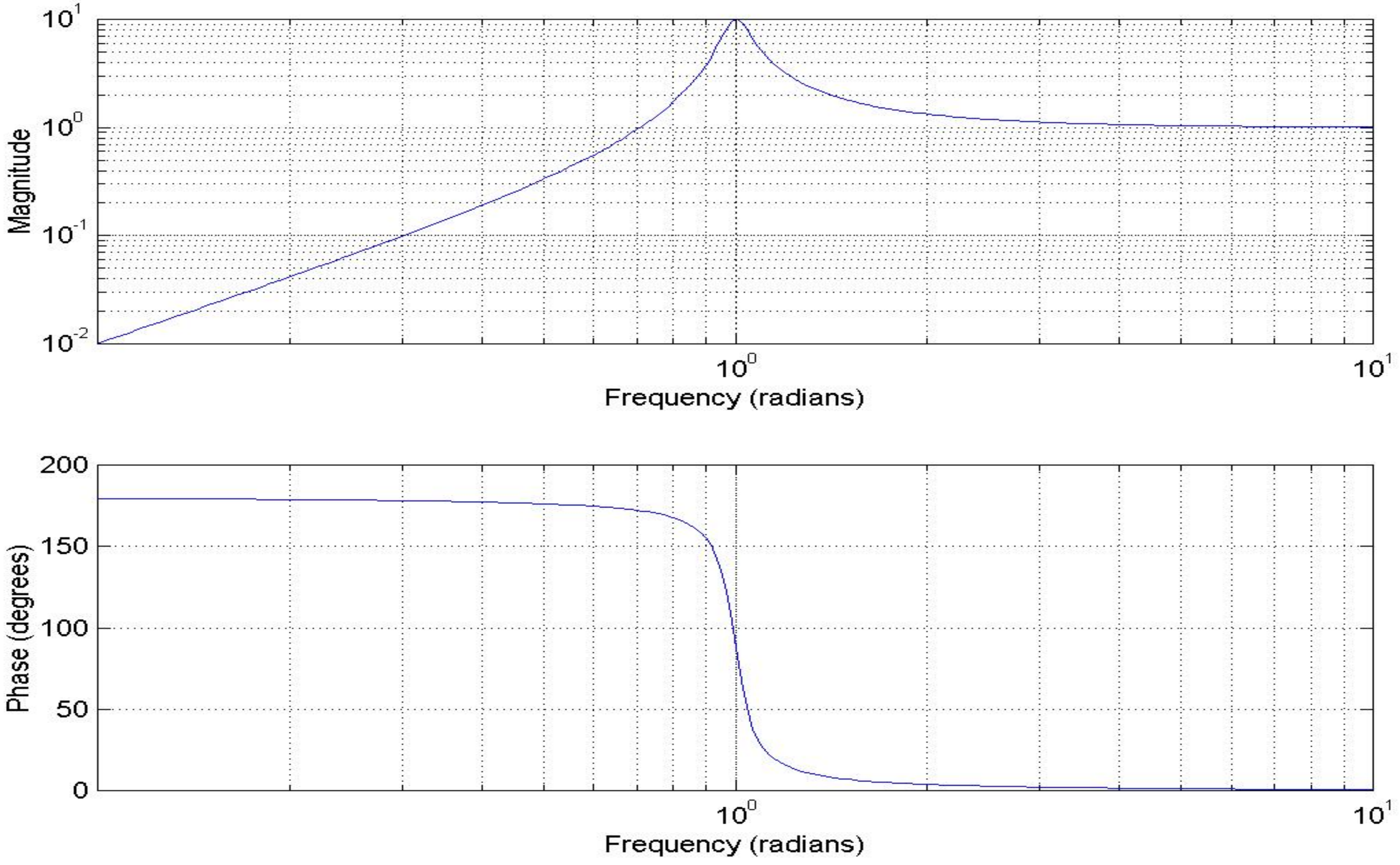
The Inner Ear

A complicated subject at best, the inner ear can be thought of as having two membranes, each a travelling wave filter, both highpass filters. Between the two membranes are two sets of **hair cells**, the **inner hair cells**, and the **outer hair cells**. The inner hair cells are primarily detectors. They fire when the movement of the two membranes are different. The outer hair cells are primarily a system that controls the exact points of the very steep low pass filters and high pass filters. The outer hair cells can polarize and depolarize, and change both their length and stiffness. This polarization is how they affect the relative tunings of the two membranes.

One point along the Membranes:



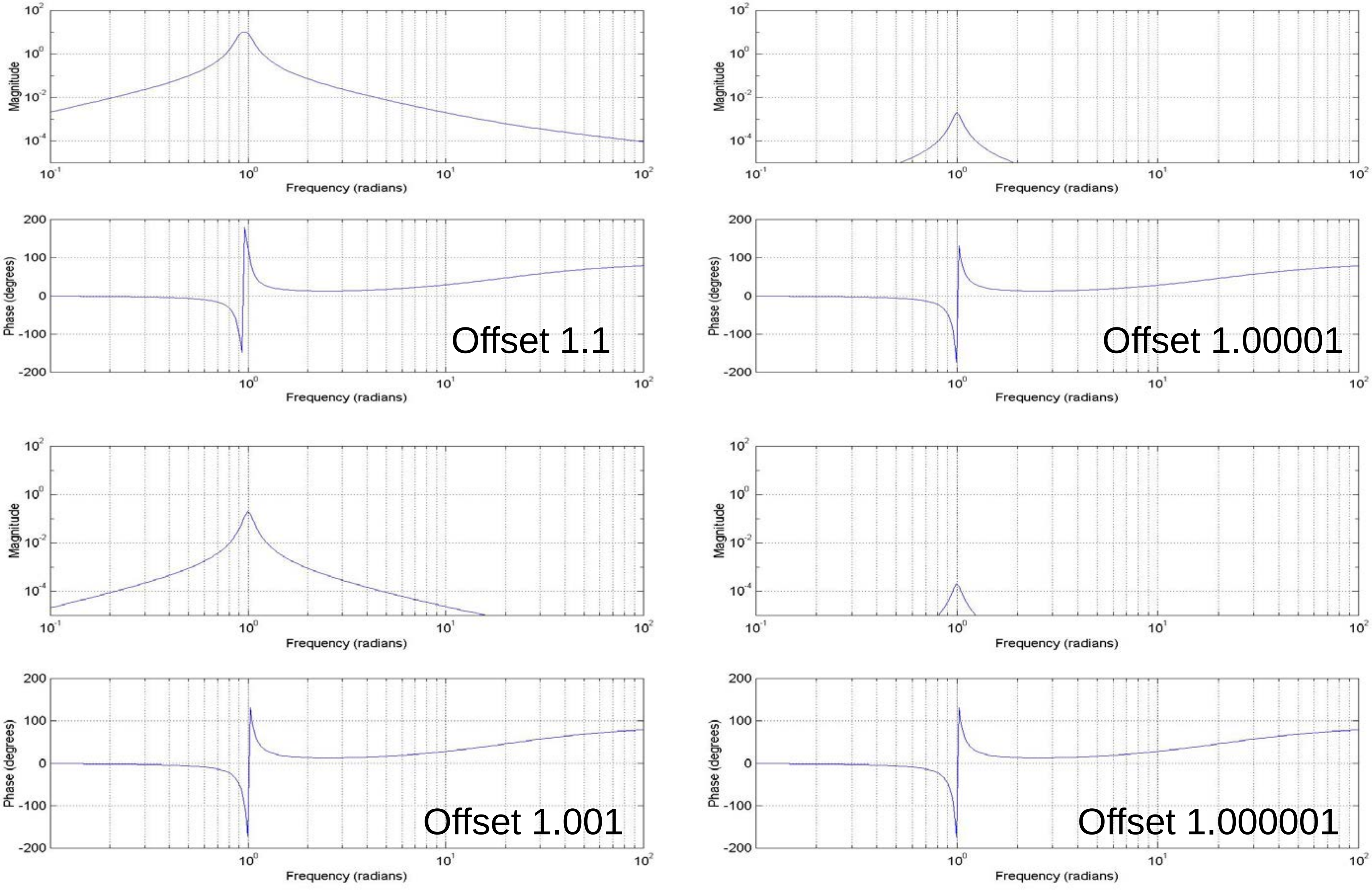
Example HP filter
(This filter is synthetic, NOT real)



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.

Filter split vs. Frequency Response



The exact magnitude and shape of those curves are under a great deal of discussion and examination, but it seems clear that, in fact, the polarization of the outer hair cells creates the **compression** exhibited in the difference between applied **sound pressure level** (the external power) and the internal **loudness** (the actual sensation level experienced by the listener).

There is at least 60dB (more likely 90) of compression available. Fortunately, the shape of the resulting curve does not change very much, except at the tails, between the compressed and uncompressed state until unreasonable levels are involved, leading to a set of filter functions known as the **cochlear filters**.

Critical Bands and Cochlear Filters

The overall effect of this filter structure is time/frequency analysis of a particular sort, called **critical band** (Bark Scale) or **effective rectangular bandwidth** (ERB) filter functions. Note that this is not a set of filters, but rather a continuous set of filters, with lower and higher bandwidths varying according to the center frequency.

Roughly speaking, critical bandwidths are about 100Hz up to 700Hz, and 1/3 octave thereafter. ERB's are usually a bit narrower, especially at higher frequencies.

A discussion of which is right and which should be used is well beyond the range of a one-hour seminar.

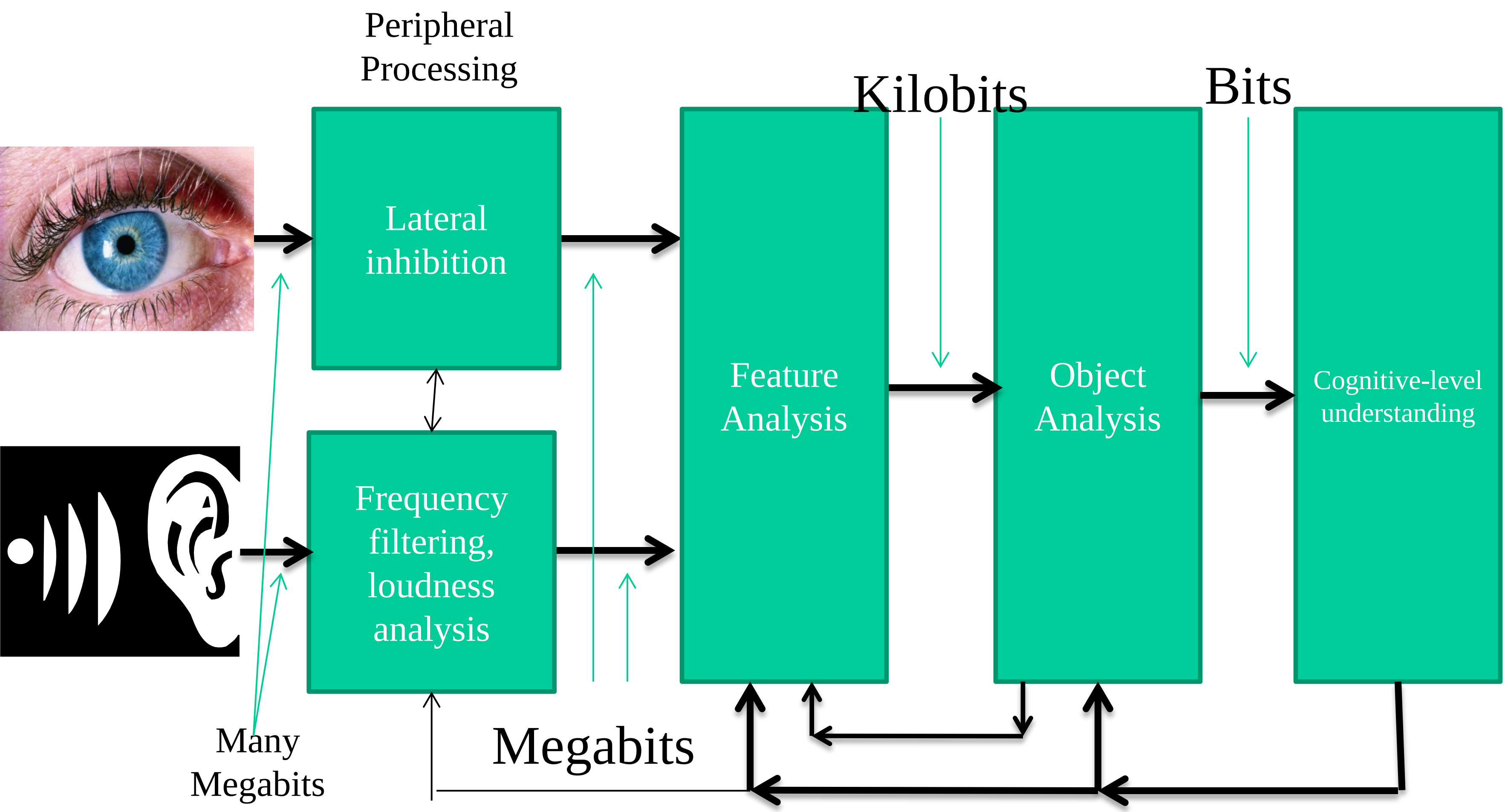
The basic points are that

- the sound arriving in an ear will be analyzed in something approximating 70-80Hz bandwidth filters at low frequencies and at something like 1/4 octave bandwidths at higher frequencies, and
- that the system will detect either the signal waveform itself (below 500Hz) or the signal envelope (above 4000 Hz), or a bit of both (in the range between 500Hz and 4000 Hz).

Exactly what is detected is also well beyond a one hour seminar, and furthermore, a consensus is yet to emerge.

A modern view of perception

OBSERVE THE MASSIVE FEEDBACK AND THE LOSS OF DATA AT EACH STEP



Hey, we have two ears, not one!

Now that we know roughly what and how one ear can detect a signal, we can examine what can be picked up By two ears.

At low frequencies, the leading edges of the filtered signal itself.
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.

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.

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 have both uncorrelated envelopes at the two ears in the relevant frequency ranges as well as uncorrelated edges monaurally. The brain interprets such waveforms as “surrounding us” or “omnidirectional”. 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 an unnatural sensation to them when observed.

How does this relate to recording and playback?

First, a recording of a plane wave should be reproduced as a plane wave, or as something that the ear's detection mechanisms can treat as a plane wave.

Second, diffuse signals should not be recorded at one point and then reproduced as plane (or spherical) waves, because the auditory system will detect any "leading edges" in them as false directional cues, leading to the classic problem of being "sucked into" loudspeakers during things like applause, or in multichannel audio, reverberation from the rear speakers.

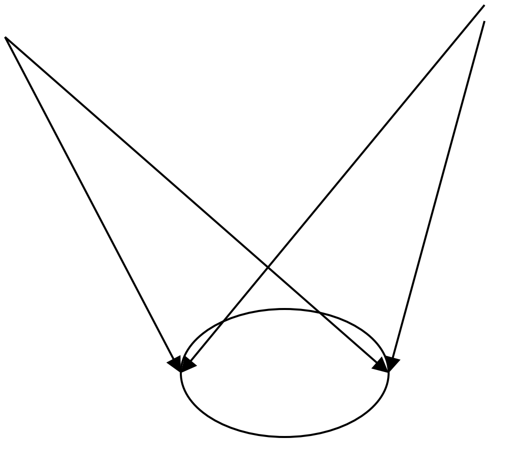
How does this relate to stereo?

Standard 2-channel stereo, using both time and delay panning or microphone capture of directions, can provide a very good front “soundstage” in the directional sense, because in the optimum listening position the stimulii from the correct speaker will reach the ear first, and carry the “leading edge” distinction.

Amplitude panning has a similar effect. Perhaps we can discuss how/why it works after the talk.

You said “directional”. Why not distance and elevation, too?

For center images, the two signals from the loudspeakers conflict



very badly. The interference creates frequency shaping that is the inverse of first-arrival distance cues and that mimics the effects of positive elevation cues.

Ok, so why not equalize that out, for central images?

1) This is politely called “hard” for signals captured by actual miking, because there is no single signal that can be identified as “central”. Certainly this may be possible for more sophisticated directional miking techniques, but this is not yet a conventional method.

Even if and when it is:

2) If we do so for signals that are put in the center by some sort of artificial/mechanical/mathematical control, then the reverberation in the listening room will carry the shape of the *added equalization*, creating a misfit in timbre between the direct and reverberant fields in the listening room. This will create, depending on the listener and the subject, either an elevation error in the central phantom image, or in the location of the reverberate.

The auditory system also compares direct timbre with reverberant timbre.

What about Multichannel?

As far back as the 1930's, Fletcher, Snow, and others showed that 3 front channels provides a much better depth illusion.

Why? Because at half the angle, the time delay is smaller (moving the interference up in frequency) and the higher frequency is more highly attenuated (reducing the size of the timbre change).

There are other reasons to have at least 3 front channels.

3 Front Channels - a good start?

With 3 front channels, and using nearly coincident miking techniques, it is possible to create a good illusion of the front soundfield without having depth or elevation problems.

Aside from avoiding the problem of interference, non-coincident 3-channel miking has other good effects, specifically a wider listening area, measured by subjective test to be 6 to 8 times the listening area available in the 2-channel presentation.

BUT – You notice we're still talking *direct* sound here!

How about the Back?

Again, empirical research shows that two back channels can provide a good sense of reverberation and depth to the rear of the listener.

The same two channels can create, if they are at the side of the listener, a good sense of side to side envelopment.

You can’t have both at once, unless you have 4 more channels, of course.

So much for 5.1, then?

No, so much for 5.0

The usual 5.1 home theatre setup, meaning 5 high-frequency radiators, and one subwoofer radiating below 90Hz or so, has *additional* problems beyond the 5.0 problem.

Specifically, although one can not LOCALIZE signals below about 90 Hz, one can detect spatial effects from interaural phase differences down to about 40Hz.

The AT&T Labs "Perceptual Soundfield Reconstruction" Demo, no longer available, contained a very nice example of these effects, and how they can change "boomy bass" in the 2-radiator case into "bass spread about a room" in the 5-channel case.

(And you notice we're still talking mostly about direct signals.)

Ok, what about Indirect Signals?

Well, as mentioned long ago, indirect signals with envelopes
that have peaks or "leading edges" can create false cues
if reproduced by a direct radiator.

What does this mean?

That's a good question. What I think it means is that a
good recording and playback setup will have both direct
and indirect radiation capability in each loudspeaker,
and that the two will be driven with appropriate signals.

Alternatively, binaural processing can provide the same effect in headphones,
but there are "gotchas" there.

Ok. Where do we go now?

Modern recording methods are just now to the point where we can in fact capture a large number of channels in an accurate, clean fashion, and either transmit them or process them for transmission as fewer channels.

This should, with associated hardware, processing, and authoring capabilities, give us the ability to provide appropriate direct and indirect signals to the listener, providing a much more immersive experience, and one more like (for a live recording) the original venue.

What Can Be Done?

Except for a few parts of this space*, recording methods have not been systematically or carefully examined, and the results of many promising miking and simulation methods are simply unknown.

Effectively, the field is just now becoming feasible.

* For example, Gerzon's Soundfield Microphone, various widely separated miking methods, and AT&T Labs (Johnston/Lam AES Preprint) Perceptual Soundfield Reconstruction

What else?

Object oriented audio (stems plus processing) offers the opportunity to “do as well as you can” without having to constrain the playback setup when the audio is recorded and/or mixed.

Some new speaker technology can clearly provide a superior verisimilitude, either for recorded signals in a performance venue, or for signals processed after the fact to simulate a given environment.

Finally, personalized headphone virtualization can handle the portable reproduction problem