Sound and Audio
vs.
The Ear

James D. (jj) Johnston

Retired audio scientist

Outline – This talk is in two parts, with a break in the middle

- Part 1 – How sound gets to the brain (at normal levels)
  - Sound – how it's created and propagates
  - Acoustics of rooms
  - Acoustics of the head/outer ear
  - Middle ear behavior
  - Cochlear analysis
- Part 2 – What happens after it goes down the auditory nerve?
  - Monaural processing
    - Masking
    - Pre-echo
  - Binaural processing
    - Time correlation,
    - Frequency correlation
    - Masking effects
  - Selective post processing

More on the leading edge part, first

- Remember how the outer hair cells are depolarized as the inner hair cells fire, with about a 1 millisecond delay? As mentioned before this has a couple of effects.
  - It reduces the gain of the system, thereby turning the sensitivity down about 1 millisecond after the initial sound arrives. That 1 millisecond is approximately 1 foot of travel of sound, so this suppresses some echoes.
  - It changes the tuning of the system, as well, slightly shifting the filter center, and changing the overall shape of the filter response.
  - It is responsible for some of the "pre-echo" problems in audio codecs, and in other signal processing where the impulse response is substantially longer than 1 millisecond.
    - Consider: Changing the gain before the main signal arrives can change the timbre of the sound.
    - It can also create an "echoey" sensation.

Masking

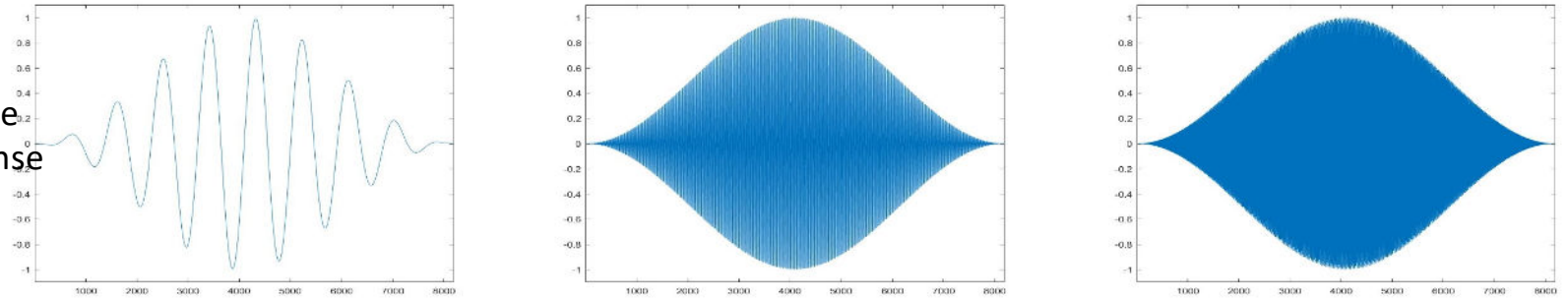
- If you consider the inner hair cell, which has typical neural refractory time of 1 millisecond, that means that the fastest rate of firing is about 1000 times/second.
  - Consider, now, shot noise. That works out directly to a signal to noise ratio of about sqrt(1000), or 30dB.
  - Again, to repeat the comment from the “loudness” section, the system of the inner and outer hair cells actively remaps this 30dB dynamic range over about a 70 to 80 dB dynamic range.
  - The filtering system, as well, ensures that frequencies far removed in the cochlea do not interact, so it is possible to have a substantial dynamic range across the audio spectrum.
- That, effectively, is what creates auditory masking. Small signals near large signals (in frequency) are simply below the noise floor of the system.
- There is another effect, based on the shape of the signal in any given ERB, that we will get to later.

ERB vs. the time domain

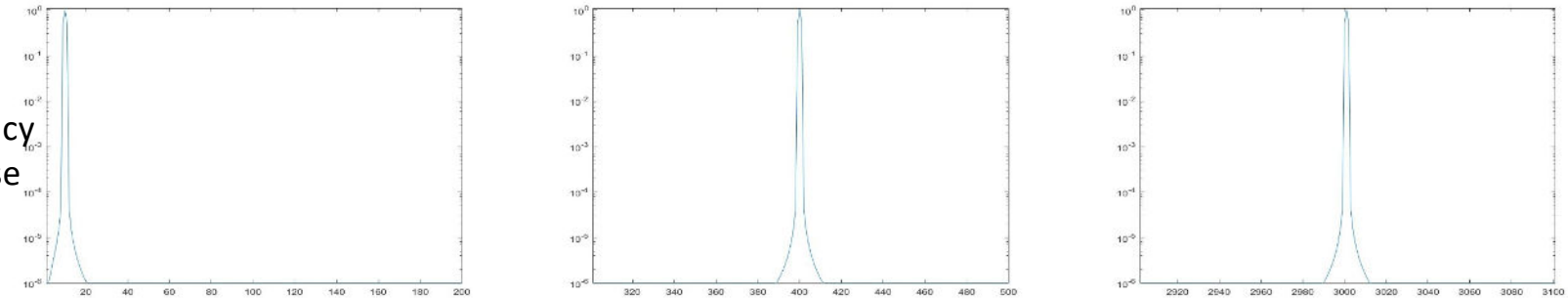
- One of the biggest mistakes made in many presentations is the use of a high-res FFT (say 8K FFT at 48kHz sampling rate) that gets about 5.9 Hz resolution, that continues at that analysis length all the way to 20kHz.
  - That kind of presentation completely obscures high frequency envelope cues.
  - It mostly suffices for frequencies under 100 or 200Hz, but, even there, issues remain.
- Remember, minimum impulse length response scales as 2/bandwidth. That is the absolute minimum possible.

Same bandwidth, 3 center frequencies

Impulse Response



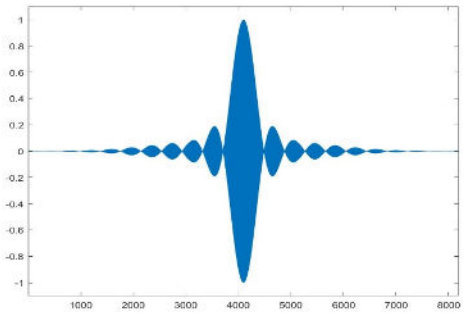
Frequency Response

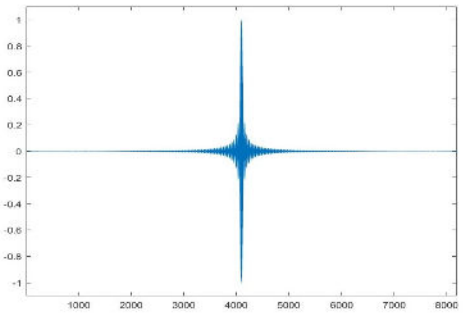


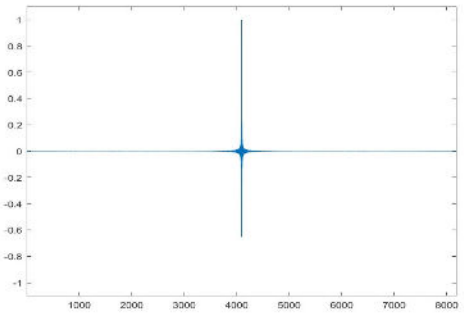
Notice the center frequency shifts, but not the bandwidth

Different Bandwidth, Same Center Frequency

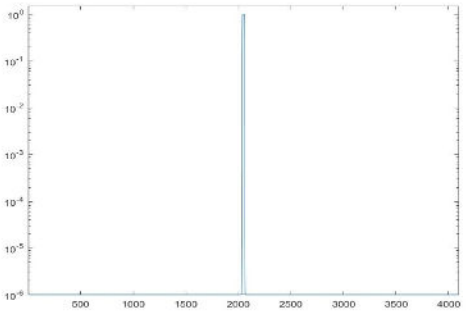
Impulse Response

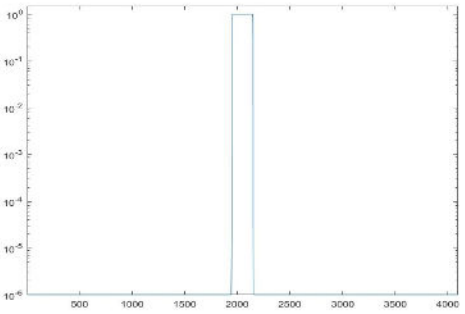


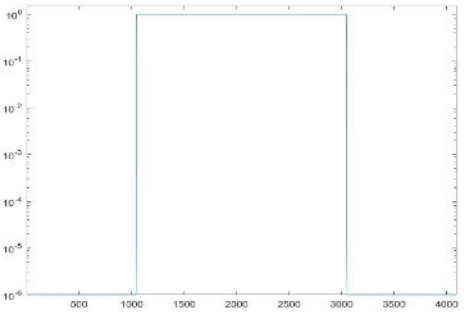




Frequency Response







What's the point?

- For a given ERB, you must also use an appropriate bandwidth and impulse response length.
  - This was difficult a while ago. It ate FLOPS
  - Now FLOPS are cheap. So DO THE RIGHT THING.

- You may recall, many years ago, spectrograms used things like octave bands, or ½ octave bands, or maybe even bark-scale bandwidths.
  - They had the right idea.
  - The ability to observe when a part of the signal happens is pretty much key to everything past this point.
- The ear has substantial detail in frequency at low frequencies.
- Conversely, it has substantial detail in TIME at high frequencies.
- NEVER FORGET THAT!

Now this talk must bifurcate

- First, I'm going to talk about monaural hearing.
  - But remember, binaural hearing can resolve things, sometimes at least, that aren't observable in monaural hearing.
  - Monaural hearing is mostly about masking and absolute thresholds.

- Then I'm going to talk about binaural hearing
  - For instance, the infamous "Suzanne Vega" problem, also known as "Binaural Masking Level Depression"
  - Here, the time domain really becomes important in very many ways.

Some other things to remember:

- Everything is filtered by the cochlea, so the properties of both time and frequency are very important.
- At any point, the CNS can look across time, OR across ERB's, or both.
- Yes, this get complicated very fast.

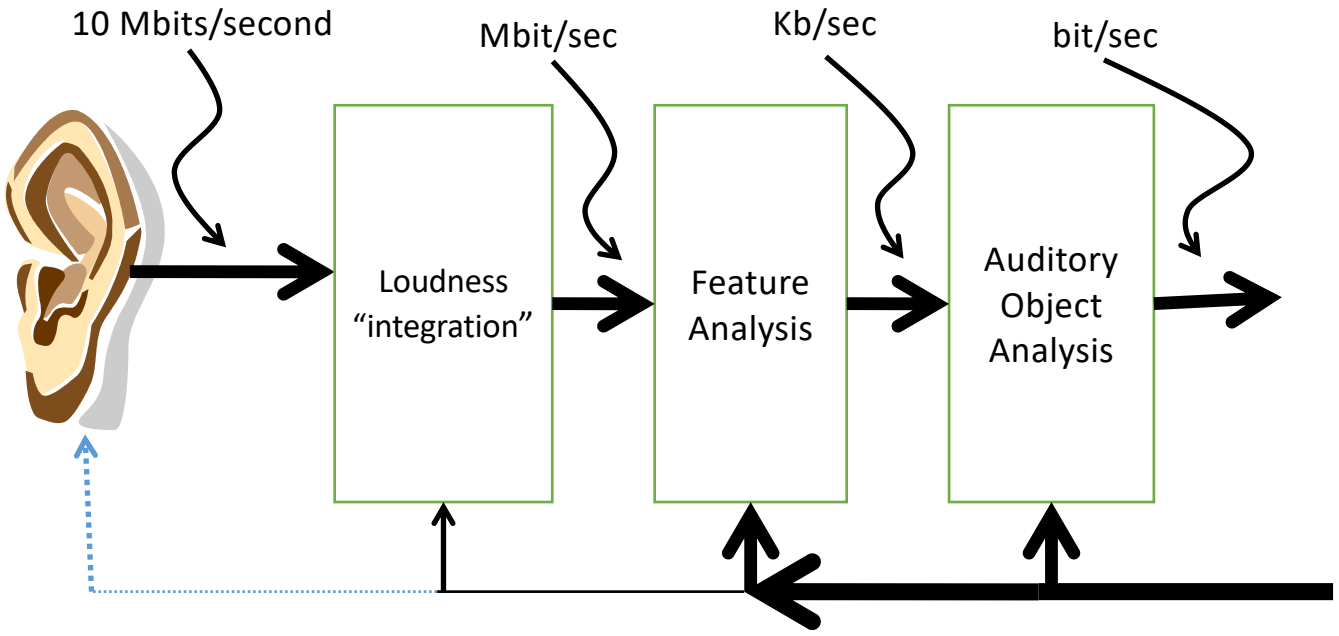
- The CNS, of course, past the most "outer" parts of the hearing system, is extremely plastic, and can be changed by expectation, experience, or intentional cognitive effort.
  - Yes, that makes things a lot more "interesting," to say the least.

First, Monaural masking

- As you may recall, the inner hair cell, the detector, has an SNR of about 30dB, which means that you're not going to do much better than that at a given frequency.
- However, the system tuning and loudness adaptation also change gain.

- Time, I think, for a picture.

Monaural, very approximate schematic for masking.



Here, the "ear" actually represents the cochlea.
We're well past the head and pinna at this point.

The Loudness Integration

- As previously mentioned, as a signal arrives, the gain of the cochlea drops with the depolarization of the outer hair cells.
- However, that does not mean that loudness goes down. There is a kind of integration-like effect that also sums, so to speak, across time.
- There's two parts to the output from the "integration".
  - First there's "first onset" either of the waveform itself at low frequencies (under 500Hz or so) , or the signal envelope above 2kHz or so.
    - Between 500Hz to 2kHz or so, it's some of one, some of the other, and there is some conflict between the two kinds of cues.
  - Second, there's "what's the total neuron firing rate?" That is what establishes the loudness.
- Yes, this process also applies to the binaural side of things, more on that later.

Simultaneous Masking

- Simply put, the question of masking is:
  - Can the difference in signal make it past the loudness integration?
  - The 30dB SNR more or less establishes a maximum sensitivity, but
  - The waveform phase lock at low frequencies can get involved too, but of course that is also “ jittered “ by the noise floor.
  - So, for a perfect regular signal (with a flat envelope, i.e. a tone or tone-like signal) 30dB inside of an ERB works out very well with the actual subjective testing.

  - But wait, what if the signal does not have a flat envelope?
    - You ***HAD*** to ask, but fortunately, there is quite some literature on the question, although some of it not directly useful for estimating masking thresholds.

More on Simultaneous masking

- There are a number of results, depending on the signal structure.
  - Tone masking tone -30dB
  - Tone masking noise -30dB
  - Noise masking tone -6.5dB (formerly 5.5 but I have my doubts)
  - Noise masking noise -3.5dB
- All of that results from both the irregularity of the waveform attack at low frequencies, or the amount of variation in the envelope at higher frequencies.
- As we will see later, some of those lower numbers can be pushed very, very lower in the binaural case. But, that's only "sometimes".

Time considerations in masking

- There is, to a great extent, absolutely NO pre-echo kind of masking. A bit of a leading edge is all it can take to make an audible pre-echo, or start the cochlear compression mechanism.
  - Once you're very close, of course, it becomes a question of spectrum and bandwidth of the interfering noise.
  - Inside of 1 millisecond, it's less likely to 'dull' the sound of the attack, because the gain reduction hasn't started completely.
- Post masking exists, for several reasons:
  1) The length of the cochlear filter stores the immediately previous energy, as it must do in order to properly filter.
  2) The integration has a time constant inside of which it's roughly collectively adding things. That integration provides a post-masking period. The post-masking effect is therefore much larger than pre-echo masking.
- Remember, this is all for each "line", which is a center frequency of an ERB.

A note about monaural perception

- Binaural hearing is how it appears the auditory system developed, with coordination between ears, enormous “crossfeed” between ears (yes, making hearing damage different in two ears a problem), to the extent that I’m almost tempted to say that ‘monaural’ is simply handled as ‘two ears with one central source’.
- There are some issues with ‘speech in a single ear’ that do seem to be unique. I simply haven’t the data to comment.

So, monaural hearing, briefly summarized

1. Time/frequency loudness analysis with emphasis on leading parts of the short-term signal.
2. Conversion of partial loudnesses into features

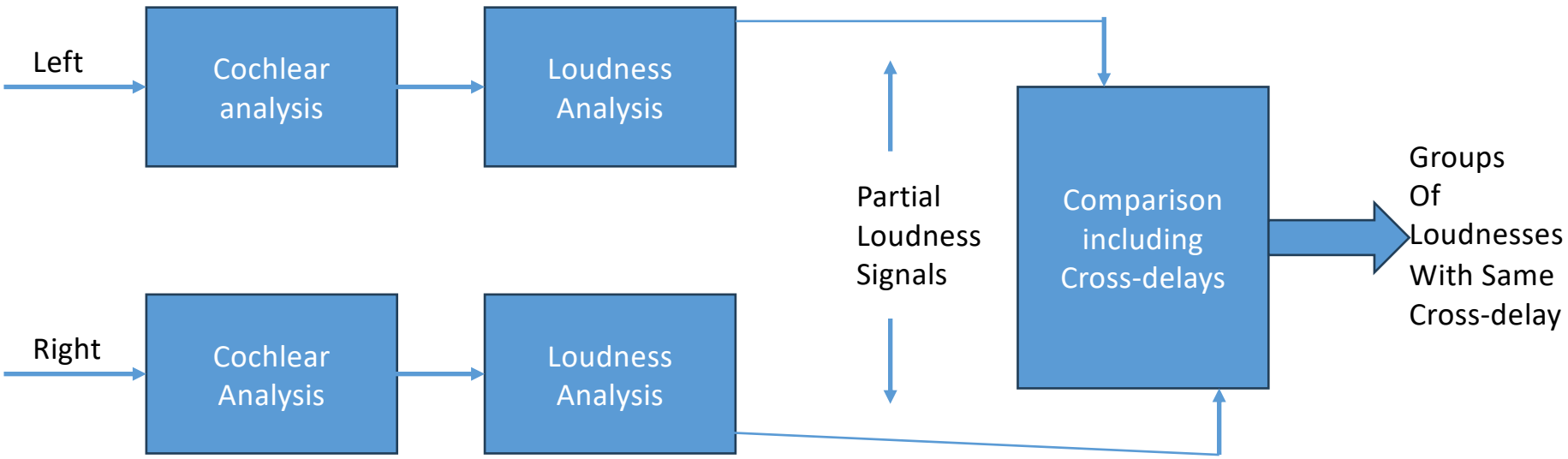
   WHICH ARE GUIDED BY DOWNSTREAM PROCESSES INCLUDING EXPECTATION, LEARNING, AND EXPERIENCE, as well as other sensory inputs.
3. Analysis into concepts/ideas/language

   Which are, obviously, subject to all of the above, and more.

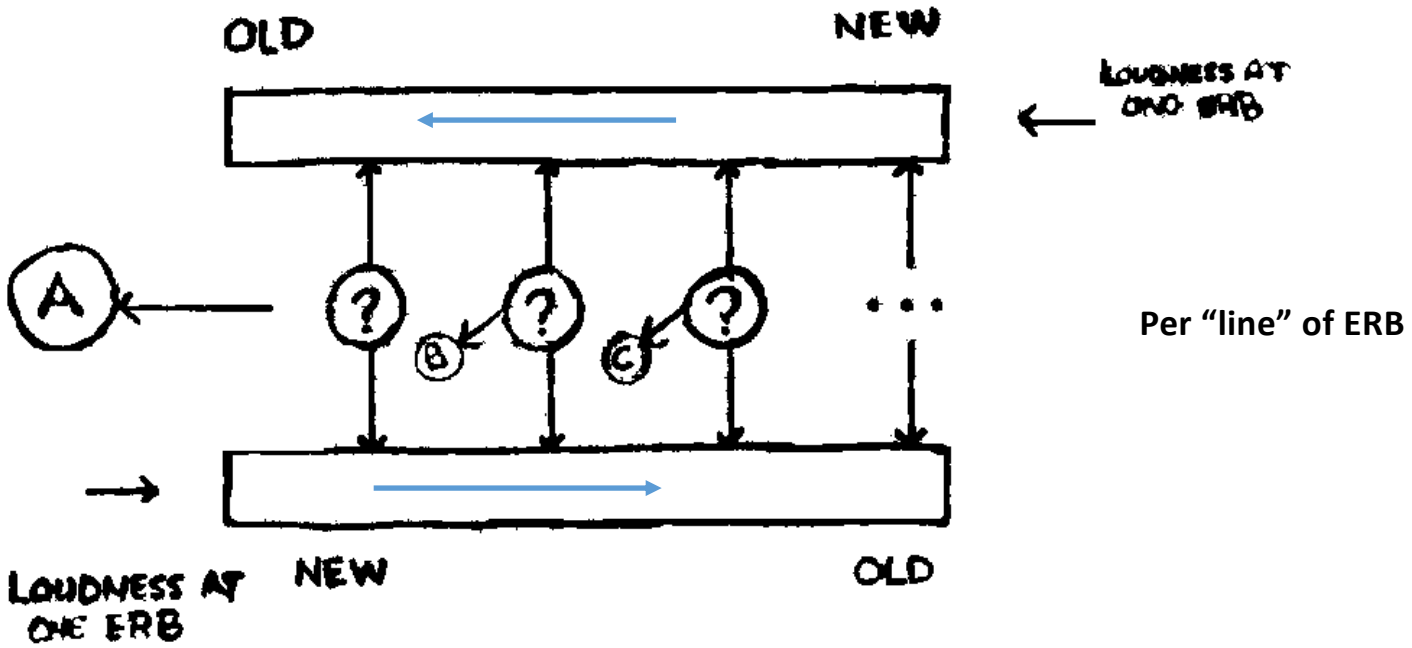
On to Binaural hearing

- Note, please. I am leaving out many details, including most of those understood by understanding of anatomy and the like. There are entire books and many medical papers discussing this kind of thing.
- I am sticking to what the RESULTS of the anatomy and neural function appear to do.
  - As always in this kind of subject, everything is subject to 'look what we discovered yesterday'.
  - The CNS is ***plastic*** and learns. I'll mention some of that in a few places, but always remember "this might change with experience, expectation, and so on."

An inexact, approximate graphic flow.



The comparison works something like this:



The "?" indicates a process that tests in some fashion for a match of some sort at different time delays.

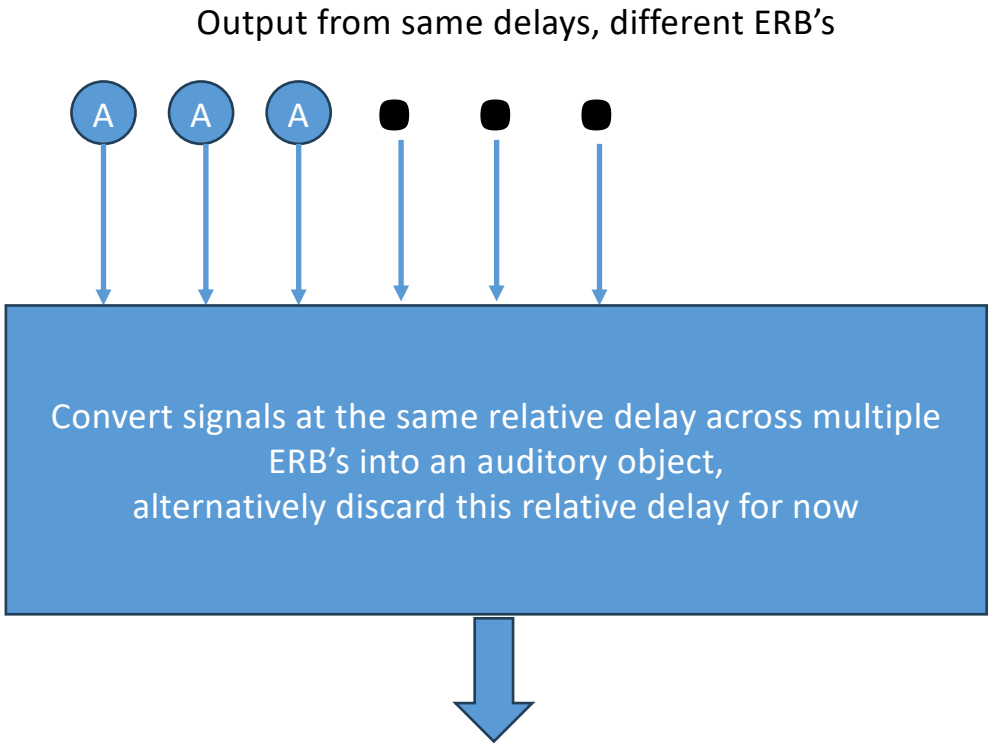
- The partial loudness as a function of time is a very “wide” vector, having many “lines”, each of which has, effectively, two values, a “first arrival” envelope, and a longer-term “rate” measurement.
  - These two pieces of information are relayed by both pulse position and pulse rate.
  - Remember, there is a lot of overlap inside of one ERB.

- The input to the mysterious box labeled “comparison …” is the partial loudness information, as neural impulses. Remember it is both a function of frequency and time.

What happens in that mysterious box?

- It appears that there is some mechanism for detecting coincidences between left and right, not only for ‘same delay to each ear’ but also for a range of delays between something like ½ millisecond lead, to the same amount of lag, for either channel, as compared to the other channel.
- It seems evident that somewhere in all of that, there can be several “objects”, at different relative delays, and that these are somehow sorted out into multiple “auditory objects” that are perceived in different orientation relative to the head.
  - Don’t ask me about the mechanism. I don’t know. I’m pretty sure nobody else does, at least very well.

Followed by something resembling this:



Note, it is possible that both steps are in parallel somehow! I'm not sure which.

- This can, for instance, create "ghost frequencies" by beating between two ears
- It can provide a "time delay location" for sources
- It can provide a way for the "cocktail party effect" to work, and can also, by interacting with the output of the visual processes, concentrate on particular parts of a spectrum at a particular relative delay.

- Nobody, I am fairly sure, can say exactly how this works.

Wait, there's more!

- Do you remember where I mentioned acoustic dispersion, way back at the start?
  - Well, that dispersion means that both monaurally and binaurally, sometimes higher frequencies (in particular) do not quite "line up" in first onset time, even inside one ERB.
  - The effects of temperature gradient, air movement, etc., all increase with distance.
  - Yes, that dispersion, if it's something similar to what actual acoustics does, can create a sense of distance. If it's something quite unnatural, it may just sound strange. There are many more ways to "sound strange" than there are to "sound natural".
  - In binaural situations, when the first onsets do not line up between L and R, that becomes an immersive sensation. Sometimes this sounds distant. If your dispersion isn't pretty much 'acoustic-like', on the other hand, strange sensations can ensue.

- Finally, “distance” can move into “immersive” as the time differences across ERBs and between ears increases.
- Even then, a first arrival that is not extremely diffuse (considering interaural diffusion) can be localized, even with extreme diffusion. This is part of how you can sense direction far away from a source in a diffuse field.
  - Direct path, even diffused, GETS THERE FIRST. Remember, that gets more heavily weighted than the rest of the signal thanks to cochlear dynamics.

More on directional sensation

- A level shift between ears can also create a directional sensation, however (first, let's stick to headphones) that will work best if you do NOT have a signal with substantial time cues.
  - Of course, there's a problem when a signal (say glockenspiel) has a sharp attack (brief high frequency content, subject to time delay analysis) followed by a ring note (providing only interaural level differences). As a result, you may hear the attack dead center, but the ring panpotted off to some other position.

- A level shift, combined with 2 speakers, can create an effect that somewhat mimics the interaural time delay, at least at low frequencies, and create the "virtual image" through interaction between HRTFs.
  - Again, of course, there's a problem when a signal (say glockenspiel) has a sharp attack (brief high frequency content, subject to time delay analysis) followed by a ring note (providing only interaural level differences). As a result, you may hear the attack dead center, but the ring panpotted off to some other position.

- That's the problem with pan pots. Level plus delay panning also helps with image shift off axis, but that is hard to quantify without more experiment.

Finally, about those low frequencies.

- It is a reasonably factual statement to say that localizing a pure tone under approximately 90Hz seems just about impossible without head movement and translation about a space.
- HOWEVER, immersive sensations, usually described as “width” or sometimes as “size”, can arise with time delay between the signal to the two ears, much like arises in a wide, reflective venue.

- Yes, this addresses the issue for subwoofers. You should stick to “low frequency radiators” below about 40Hz, if at all.

Finally, about longer time delays

- For this I mean 30 milliseconds and on up.
- This can be echoes, which are mostly specular (i.e. less diffuse) delayed signal, which is ok if that's what you wanted.
- This can be reverberation (highly diffused signal).

- This time delay can mimic what happens in a real hall that's very live, with very diffuse reverberation.
  - Your brain can, via the TIME DELAY separate out the reverberation from the direct signal, such that you are not "swimming in reverb" even though the reverberation is substantially stronger in energy.
  - I've only seen this work well in 5 channel systems with 2 back speakers, and even better in 7 channel systems with both side and back speakers, and appropriate delays to each.

Even LONGER time delays

- Specular reflections (not very diffused) at ¼ to ¾ second can create severe difficulty for many speakers and singers, as the feedback from the vocal tract to the ear conflicts destructively with the hall feedback.
  - This is a reason that monitor setups and acoustic treatment can make or break a singer or speaker.
  - Yes, some people can “get past” this. Many, perhaps most, can not.

- Finally (pet peeve warning here), reflections from the 50-ish millisecond up that are very strong and very coherent can completely destroy the ability to make out speech at the audience position. This is more than just the comb filtering that results, it also creates issues within the hearing system.

There’s lots more to say

SOME
OTHER
DAY!

Questions?
Comments?
Ibuprofen?