THE HUMAN AUDITORY SYSTEM

DAVID J M ROBINSON

In this lecture, we will examine the apparatus used by a human to hear sound. These notes include the diagrams used within the lecture, as well as a description of the material covered. Further reading is suggested at the end. You will not be examined on the contents of this lecture, but knowledge of the human auditory system will aid you in your understanding of the various audio-coding schemes that will be covered tomorrow.

WHY STUDY THE HUMAN AUDITORY SYSTEM?

In the video section of this course, you will have learnt how the properties of human vision have effected the design of television systems. For example, the chroma (colour component) information of a colour television signal is transmitted at half the resolution of the luminance (black/white intensity component) information, because our eyes have a lower spacial resolving power for chroma than for luminance. Effectively, half of the chroma information is thrown away, but our eyes fail to perceive any serious degradation in the image quality. We are reproducing the original image *"less well" than is possible*, and in doing so, saving transmission bandwidth, because humans cannot perceive the difference.

Historically, the aim of hi-fi has been to reproduce some heard event (a musical instrument being played, a newsreader speaking etc) *as accurately as possible*. The ideal would be to sit in a room, separate in time and/or location from the original event, and to hear exactly what one would have heard, had one been present at the original event. This is an impossible ideal, but we can come very close. If we close our eyes whilst listening to a good hi-fi, we may start to believe that the original event is happening right in front of us.

Whilst high resolution digital audio systems are designed to store the sound waveform with great accuracy, in some situations it is desirable to discard as much information as possible. For instance, CD quality audio consists of 1.4×10^6 bits per second. A typical modem connection to the internet might only transfer 4×10^4 bits per second. So to receive CD quality audio via a modem, we either have to wait 35 hours to receive an hour long CD, or throw away 97% of the information to receive it in real time. If we discard 97% of the digital information without any regard for how it will sound to a human listener, the result will be a very poor sounding audio signal. However, if we can examine how a human listener perceives sound, identify components that will not be audible, and only throw these away, then we can reduce the amount of information needed to represent an audio signal **without** changing the sound that is perceived by a human listener.

We cannot (yet) discard 97% of the information and retain a CD quality signal. However, in tomorrows lecture you will discover the technology by which we may discard 90% of an audio signal without changing how it sounds. In today's lecture we will examine the properties of the human auditory system that make this possible.

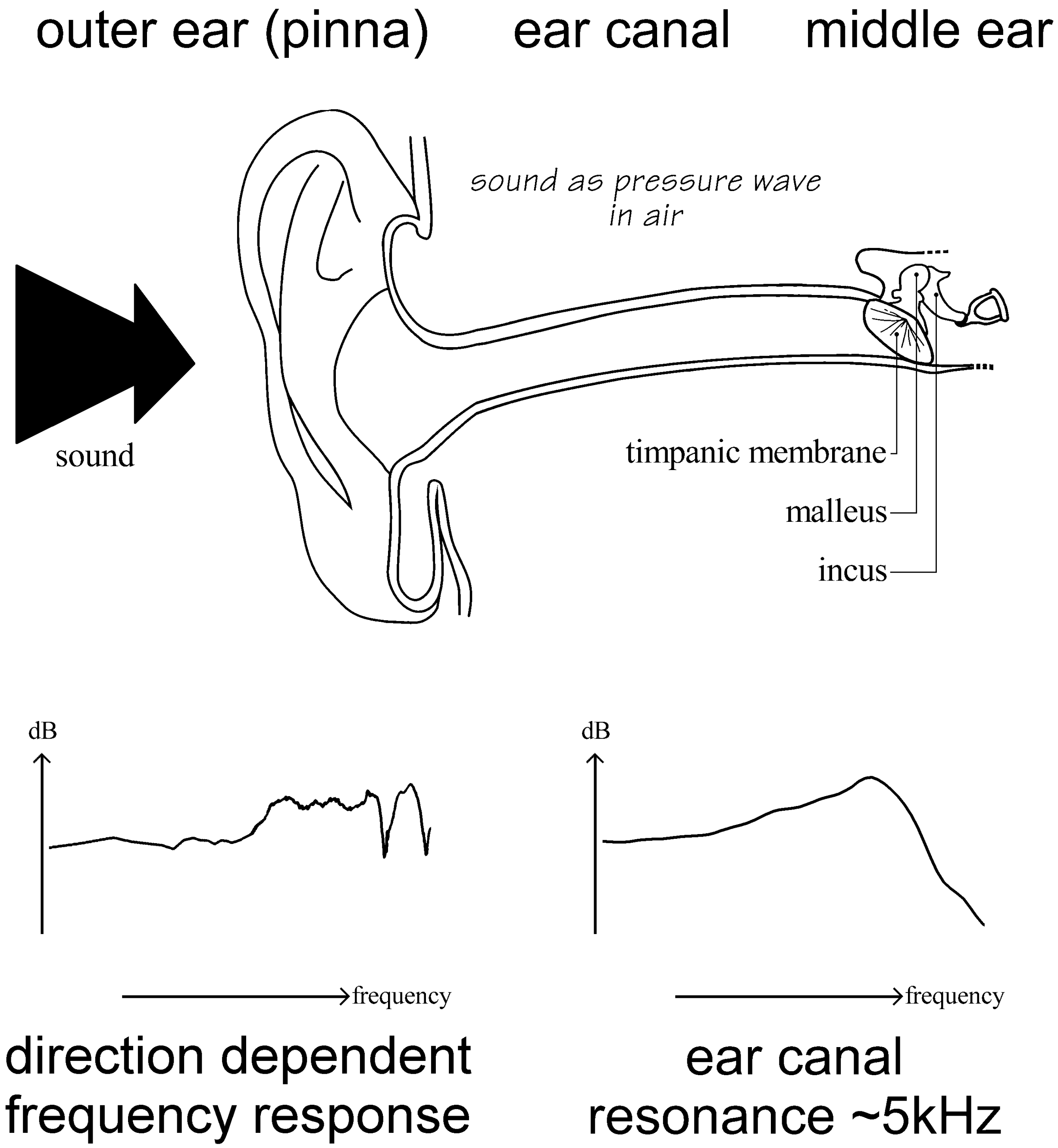
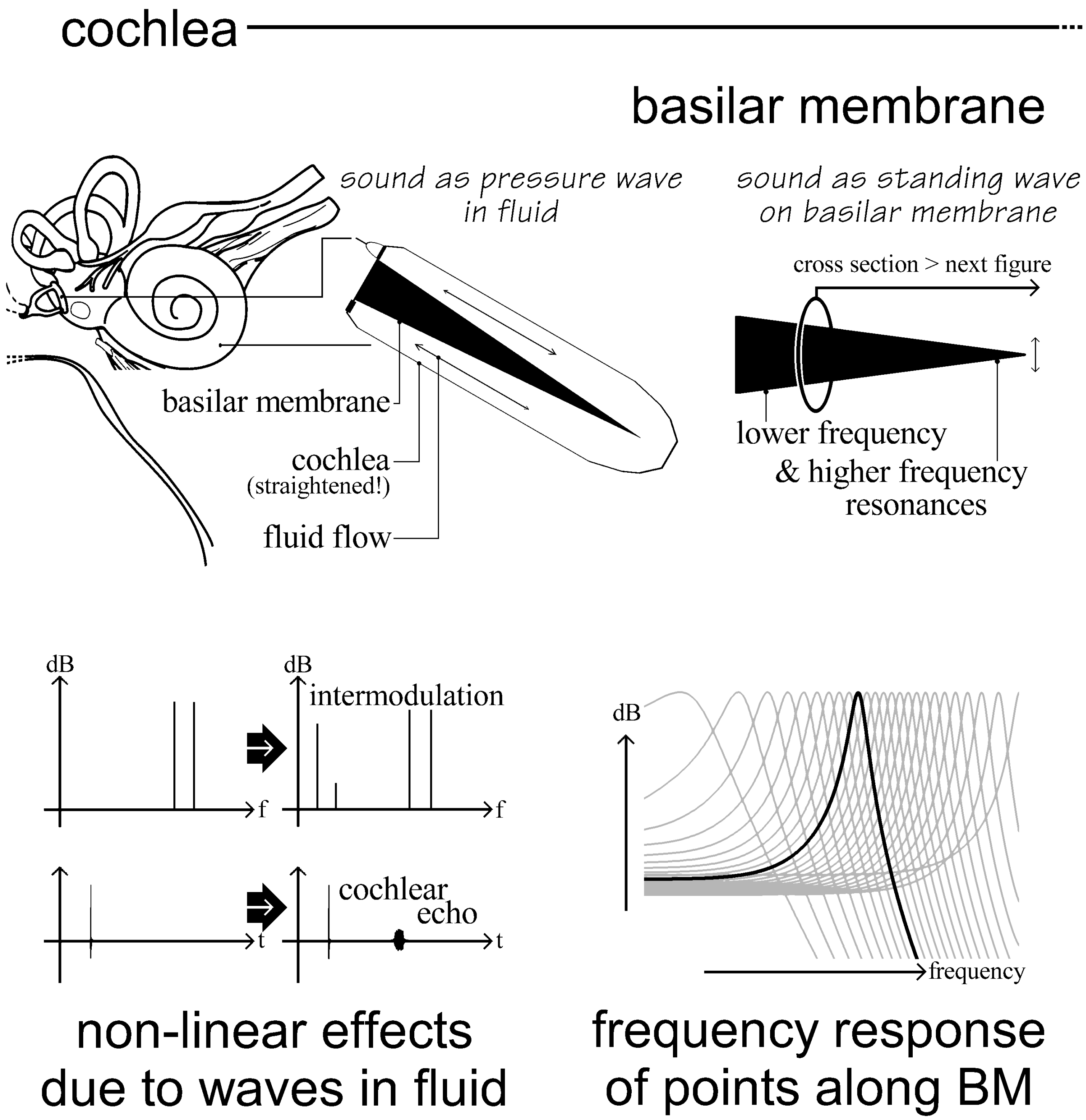

outer ear (pinna)
ear canal
middle ear
sound as pressure wave in air
sound
timpanic membrane
malleus
incus
dB
frequency
dB
frequency
direction dependent frequency response
ear canal resonance ~5kHz

Figure 1 (a) Signal path through the human auditory system

- from free field to middle ear

Upper half: physiology. Lower half: functionality

Sound waves incident from different angular positions are spectrally shaped by the pinna in a direction dependent manner. The ear canal further filters the waveform, before it passes through two small bones, and on to the cochlea.



cochlea

basilar membrane

sound as pressure wave in fluid

sound as standing wave on basilar membrane

non-linear effects due to waves in fluid

frequency response of points along BM

Figure 1 (b) Signal path through the human auditory system

- within the cochlea: the basilar membrane (BM)

Upper half: physiology. Lower half: functionality

Sound waves enter the cochlea and set the fluid within in motion. The cochlea is partially partitioned by the BM, different points of which resonate at different frequencies. Thus the BM acts as a spectrum analyser.

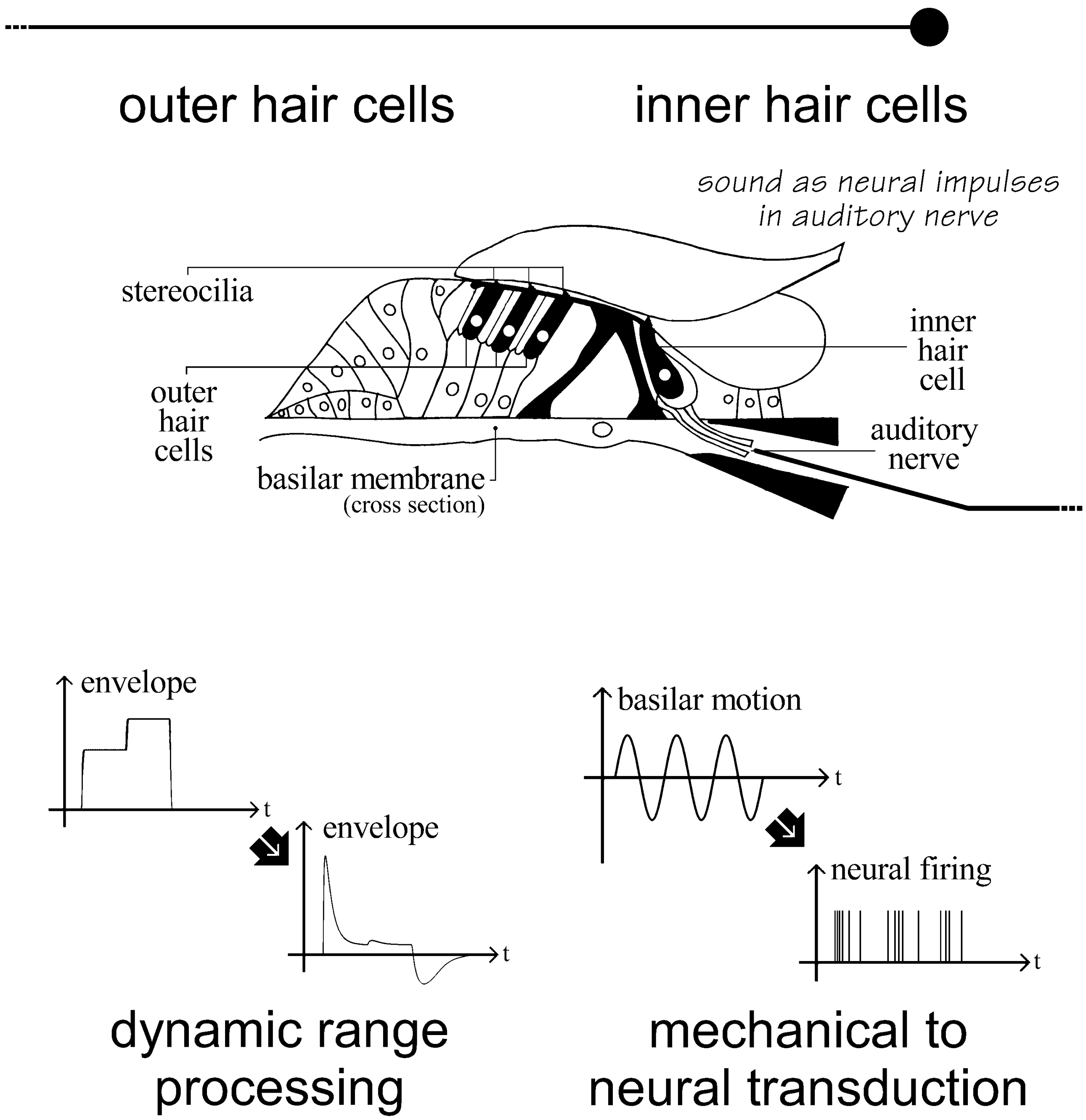


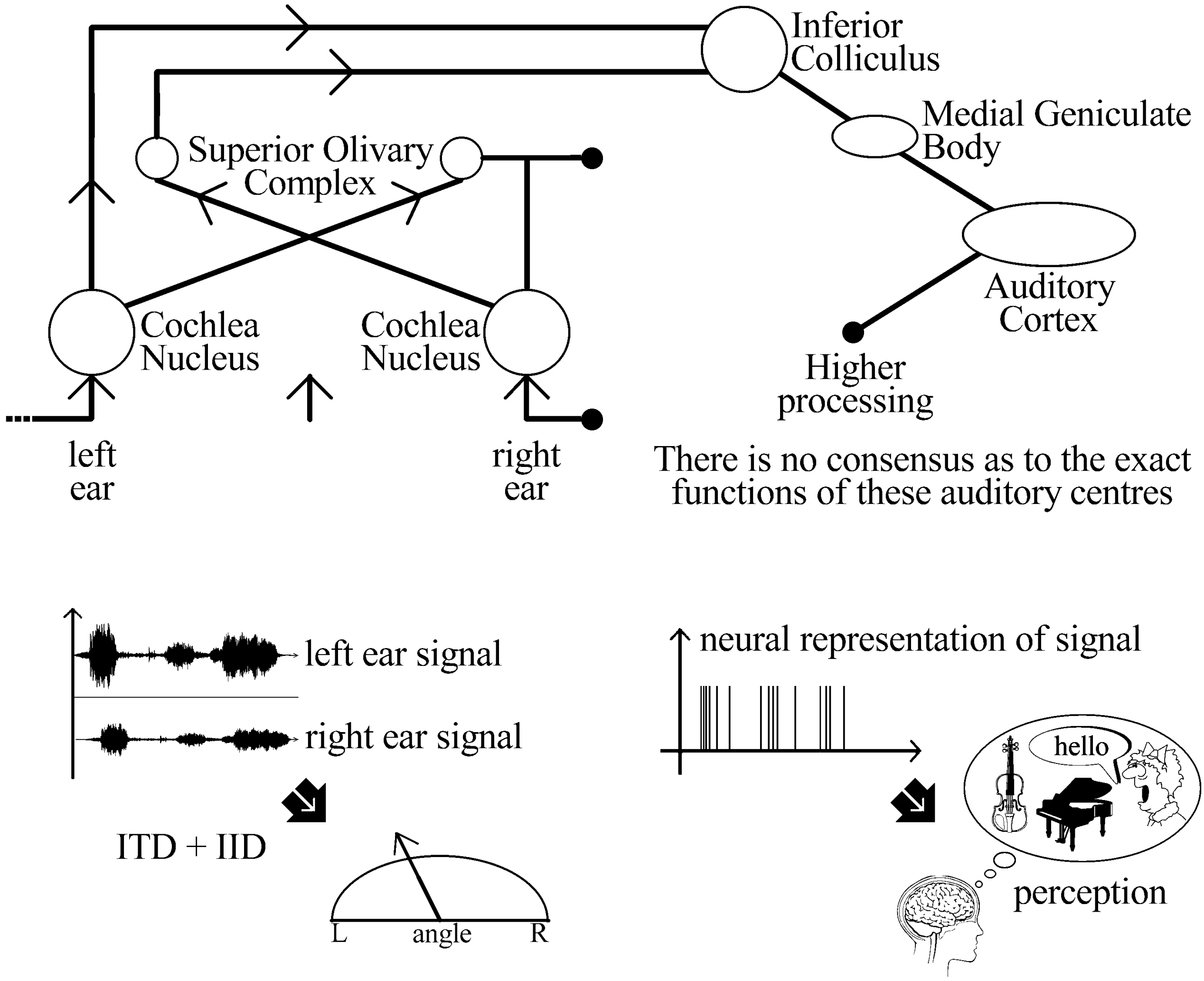
Figure 1 (c) Signal path through the human auditory system

- within the cochlea: the hair cells

Upper half: physiology [1]. Lower half: functionality

The motion of the BM causes the firing of the inner hair cells that are distributed its length. The outer hair cells act to tune the resonant properties of the BM due to signals fed back from the brain. The signals from the inner hair cells pass along the auditory nerve.

neural signal processing



Inferior Colliculus

Medial Geniculate Body

Superior Olivary Complex

Auditory Cortex

Cochlea Nucleus

Cochlea Nucleus

Higher processing

left ear

right ear

There is no consensus as to the exact functions of these auditory centres

left ear signal

right ear signal

ITD + IID

L angle R

neural representation of signal

hello

perception

interaural difference gives source angle

attention, analysis and perception

Figure 1 (d) Signal path through the human auditory system

- neural signal processing

Upper half: physiology. Lower half: functionality

The cochlea nucleus acts to sharpen the features of the incoming sound, while the superior olivary complex is responsible for our perception of sound location. The function of other neural centres higher up the human auditory system is debated, but they lead to our perception and understanding of the audio signal as speech, music, noise, or any other event.

A WALK THROUGH THE HUMAN AUDITORY SYSTEM

Figure 1 (split across four pages) shows the main components of the human auditory system. The upper illustrations represent the physiology – the actual physical components that are present and identifiable within the human anatomy. The lower graphs indicate the functionality of each section. All frequency domain plots show amplitude in dB against log frequency. All time domain plots are linear on both scales.

Referring to Figure 1, the function of each section is as follows.

- **The Pinna**

Incoming sounds are funnelled into the ear canal by the pinna (also referred to as the concha). This is the flap of cartilage and skin found on each side of the head, which is the only external part of the ear.

Until the late 19th century, the pinna was thought to be a simple funnel that directed sound into the ear. We now know that the pinna is responsible for our ability to locate sounds in 3-D space. This ability is worth examining, as fooling it forms the basis of all commercial systems which aim to create a virtual auditory environment (e.g. Dolby virtual surround, Lake theatre-phones, and any binaural [2] or transaural [3] synthesis).

Humans can locate sounds in 3 dimensions. This is evident from the fact that sounds coming from outside our field of vision can readily be located. There are six mechanisms, or cues, by which we locate sounds. We will use a simple example to explain these. Imagine there is someone stood speaking in front of us, slightly to our right. Firstly, as the speaker is nearer to our right ear than to our left, the sound will reach our right ear first, yielding a delay between our two ears. This is called the inter-aural time (or phase) difference, or **ITD**. Secondly, due to the level of sound reducing as the square of the distance from the source, and because the speaker is slightly nearer our right ear, the sound reaching the right ear will be louder than that reaching the left. This is called the inter-aural intensity (or level) difference, or **IID**. Third in order of importance is the action of the pinna, which we'll discuss in the next paragraph. Fourthly, as we move our heads, **all** the cues change dynamically, depending on the exact location of the speaker. This gives us a more accurate bearing on their location, and itself acts as a fourth localisation cue. The fifth and sixth cues are weaker. Fifthly, in a reverberant environment (e.g. any normal room), the loudness of the direct sound from the speaker compared to the level of reverberation will give an indication of the distance we are from the speaker. Finally, if the speaker is a very long distance away, the

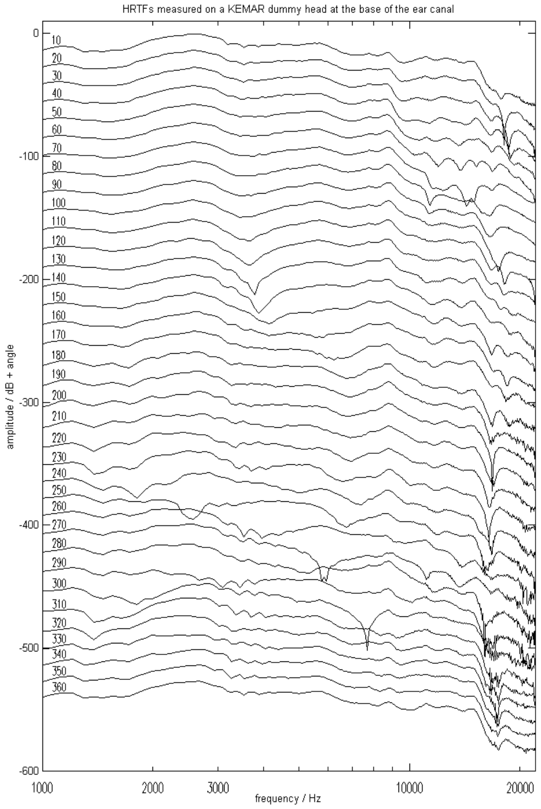


Figure 2

Head related transfer functions of a dummy head.

large volume of air through which the sound must pass to reach our ears will attenuate high frequencies, yielding another distance cue.

If we keep our heads stationary**, all these cues give us no way of knowing whether the speaker is in front of, or behind us**, or above or below us. Here we come to the action of the pinna. The pinna filters the incoming sound wave in a directionally dependent manner. Figure 2 shows the frequency response of the sound reaching the ear canal (via the pinna), from sources at various angles around a listener. These responses are called Head Related Transfer Functions (HRTFs) [4]. However, we don't perceive this pinna filtering as changing the spectrum of the sound – our auditory system decodes the spectrum shape into spatial information. Hence we don't consciously notice that a sound coming from 45^o to our right has a notch in it's frequency response at 10kHz – instead, we perceive a sound with a notch at 10kHz as being 45^o to our right. In fact we don't need two ears to localise sounds at all – blocking one ear proves that even without the ITD and IID information, the spectral cues due to the pinna are sufficient to allow us to localise a sound source in 3-D space [5].

- **The ear canal**

This is the resonant cavity between the outer and middle ear (see Figure 1). It has a resonance at around 3-5kHz, hence it attenuates higher and lower frequencies. This response contributes to our increased sensitivity to speech (mid-frequency) sounds. Figure 3 (from [6]) shows the level of sound at each frequency which we *perceive* as being the same level. Examination of these curves shows the importance of the ear canal resonance, clearly visible as an increased sensitivity to sounds in the 3-5kHz frequency region.

The **timpanic membrane** (ear drum), **malleus** and **incus** transmit the sound pressure wave from the ear canal into the cochlea.

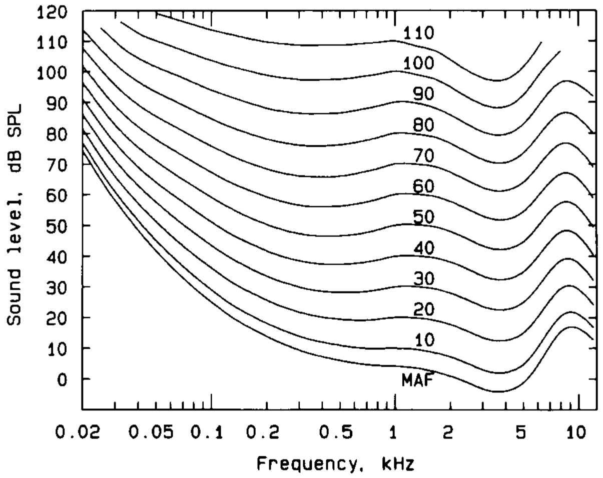


Figure 3

Equal loudness curves and Minimum audible field.

- **The Cochlea**

The fluid-filled cochlea is a coil within the ear, partially protected by bone. The fluid vibrates with the incoming sound wave. The cochlea is semi-partitioned along its length by a thin flap called the basilar membrane.

- **The Basilar Membrane**

The basilar membrane (BM) vibrates with the incoming sound, and acts as a spectrum analyser, spatially decomposing the signal into frequency components. The BM is not of uniform thickness – it tapers from one end to the other. The resonant frequency of a point on the BM varies with thickness, thus sounds of different frequency cause different parts of the BM to vibrate. This yields two interesting consequences.

- **Logarithmic pitch perception**

The spacing of frequency resonances along the BM is not linear with frequency. For instance, if we take the point on the BM that resonates at 2kHz, and measure the distance to the point which resonates at 4kHz, we may find that it is 1cm. If we now move a further 1cm along the BM, we might expect to reach the 6kHz resonant point. In fact we find a resonance of around 8kHz. The resonant frequencies of various points along the BM are shown in figure 4, from [7]. The scale that relates the resonant frequency to position on the BM is called the Bark scale, or critical band scale [8]. In engineering terms, it approximates to a log scale.

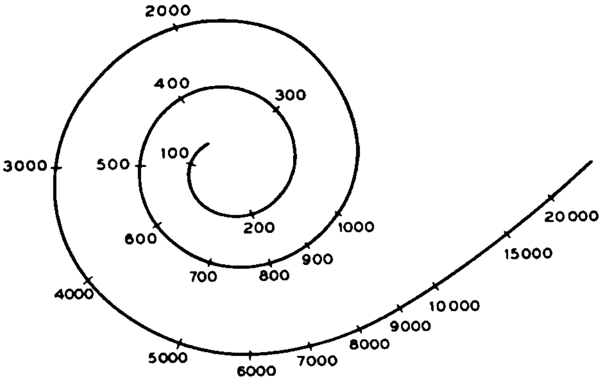


Figure 4

Position of resonant frequencies along the coiled length of the basilar membrane.

This matches our perception of pitch, which is logarithmic. Musical instruments are also tuned logarithmically. A demonstration of linearly spaced and logarithmically spaced tones clearly shows that we hear pitch in a logarithmic dimension. This shows how poorly a linear scale approximates to what we actually perceive, and how much closer we come to human perception simply by adopting a logarithmic scale.

- ***Spectral masking***

Figure 5(a) shows the response at a particular point on the BM [9]. This graph shows, for a range of input frequencies (x axis), the amplitude of vibration of one particular point on the BM (y axis). Each point on the BM has a similar response, shifted up or down in frequency (shown in grey). If we now stimulate the BM with a pure tone, the response is the excitation pattern show in Figure 5(b). This graph shows, for each point on the BM (x axis), the amplitude of vibration of each point (y axis) due to a pure tone. The excitation pattern is the reversal of the frequency response. Mathematically, we have convolved the frequency response (Figure 5(a)) with a delta function (the pure tone), hence the reversal. To show this empirically, examine the responses of the resonances either side of the tone (grey plots on Figure 5(a)). The height at which they intersect the pure tone corresponds to the amount the tone will excite that point on the BM. Note that all points *above* the tone are excited by it somewhat, but points significantly *below* it are

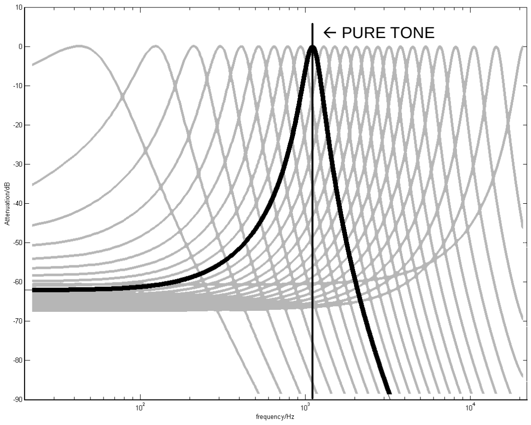


Figure 5 (a)

Frequency response of points on Basilar Membrane

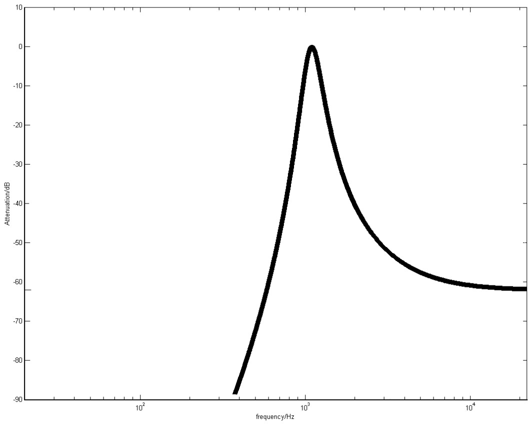


Figure 5 (b)

Excitation of Basilar membrane due to a 1kHz pure tone not, which corresponds to the shape of the excitation curve.

Thus a single pure tone excites a wide region of the BM. This is unsurprising; the BM is a continuous membrane, so it would be impossible for one point on it to move, whilst an adjacent point was stationary, as that would cause it to rip to tear. So even a pure tone (which is spectrally very narrow) must excite a finite region. This region covers the resonant peaks of several frequencies either side of the pure tone.

Consider what would happen if we applied a second pure tone, at the same time as the first, of a lower amplitude and a slightly higher frequency, such that the vibration that it *should* cause on the BM is already being cause by the original tone. If this new tone were played in isolation, the BM would vibrate, and we would hear it. However, the BM is already vibrating, and the presence of the new tone will not cause any increase in vibration. This means that we will *not be able to hear* the new tone. We call this effect spectral (or frequency) masking. No matter how hard or carefully we listen, the transducing apparatus in our ear does not have the capability to pass information about the second tone to our brain, and we can never hear it in the presence of the first (louder) tone [9-10].

This can be demonstrated by generating a series of tones ascending in pitch, and then adding a louder tone, at an intermediate pitch, which persists through the entire sequence. A spectrogram of such a sequence is shown in Figure 6. As the pitch of the stepped tones passes through the louder tone, the stepped tones become inaudible, demonstrating the spectral masking due to the louder tone. Another feature is that the masking extends further above the tone than below it, due to the upwardly extended excitation pattern on the BM, as shown in Figure 5(b) – this is usually referred to as the upwards spread of masking.

Spectral masking is a very important concept in the design of audio codecs – if we are looking for redundant information to discard, or somewhere to hide some noise or distortion, then we have just found a very large spectral space – hidden just under the loudest spectral components of every audio signal.

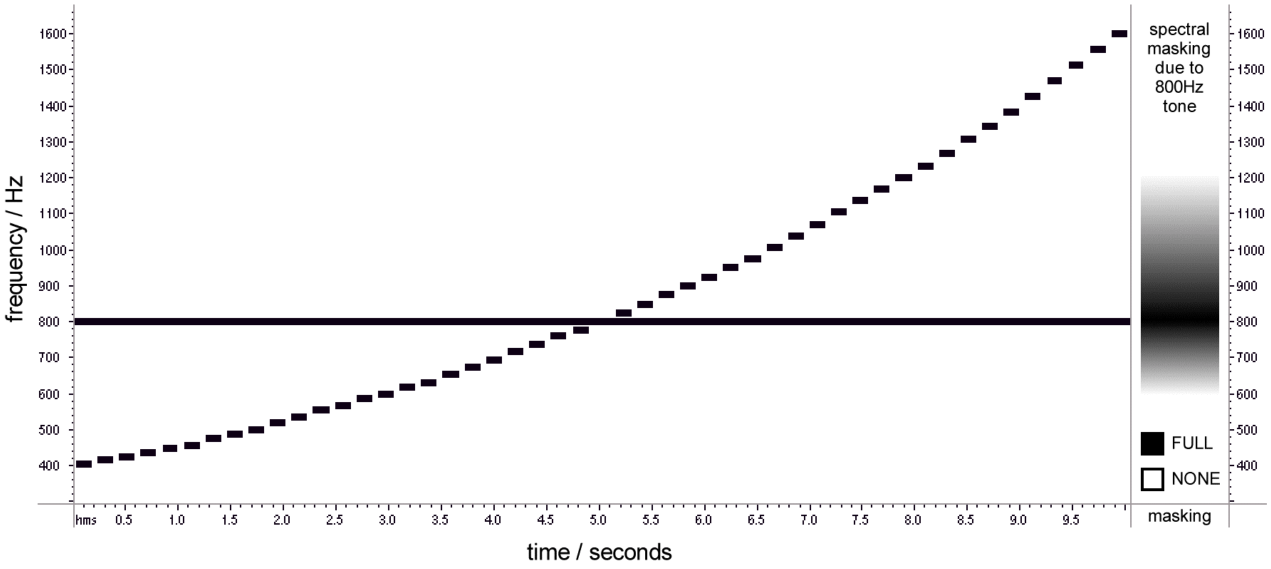


Figure 6

Left: Spectrogram of a demonstration of spectral masking. Right: Indication of masking due to 800Hz tone.

- **The Inner Hair cells**

We now return to our journey through the human auditory system. The basilar membrane is moving in response to the incoming sound wave, but how do we detect this? The answer is the thousands of tiny hair cells, running along the length of the BM. There are many different types of cells on the BM, but the two most important are the inner and outer hair cells, labelled in Figure 1(c).

The inner hair cells transduce the movement of the BM into neural impulses, which are carried onto the brain via the auditory nerve. The inner hair cells only fire when the BM moves upwards, so the signal is effectively half wave rectified. Each cell needs a certain time to recover between firings, and the firing of any individual cell is pseudo-random, modulated by the movement of the BM. However, in combination, signals from large groups of cells can give an accurate indication as to the motion of the BM.

At lower frequencies the firing of the inner hair cells may phase lock to the incomming signal, and the phase of the signal is preserved and transmitted along the auditory nerve. However, above approximately 1.5kHz, the hair cells do not lock onto individual cycles, and only the amplitude envelope is transmitted. This can be useful when designing an audio codec, as high-frequency phase information, and in particular phase differences, may be discarded without affecting what is heard.

- ***Temporal masking***

The response of a group of inner hair cells to a short and long tone burst is shown in Figure 7 [11-13]. (This response also incorporates the AGC mechanism described in the next section). Three important features are apparent. Firstly, the hair cells are most sensitive to the onset of sounds. Secondly, the hair cells take a finite time to recover their full sensitivity after the end of a sound. Thirdly, the magnitude of this recovery depends on the duration of the sound. This gives rise to a process known as temporal masking. If a sound (of a similar frequency) occurs during the period of recovery, the hair cells may be unable to register it, hence it may be inaudible.

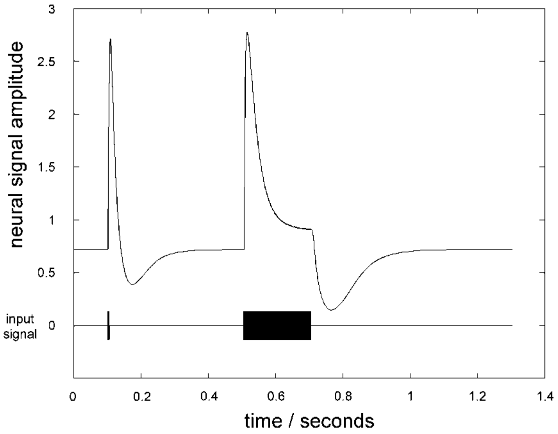


Figure 7

Overall response of hair cells to 2 1kHz tone bursts (5ms then 200ms)

This has been demonstrated experimentally. A burst of white noise was followed by a short tone. Usually this tone would be audible above a level of 20dB. The graph in Figure 8 [15] shows the threshold level of just being able to hear this tone at various times after the burst of noise. Note that the time scale on the graph is logarithmic. The two plots are for 200ms and 5ms bursts of noise. Sounds below the level indicated are inaudible.

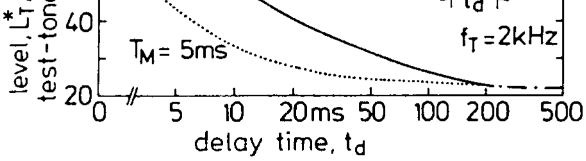


Figure 8

Temporal masking threshold due to a 200ms (solid line) and 5ms (dashed line) burst of white noise. Threshold shown for 5ms burst of 2kHz tone.

This too is very important in the design of audio codecs, as there is an inaudible temporal space hidden just after any abruptly ceasing components an audio signal.

- ***Absolute Threshold***

The inner hair cells fire at random, even in the absence of any incoming sound. In silence, the blood flowing around the regions of the inner ear becomes audible. These two factors combine to set an absolute minimum threshold of hearing. The shape of the threshold of audibility (Figure 9) is set by the resonance of the ear canal, the fall off in mechanical sensitivity of the BM at the frequency extremes, and the death of hair cells as humans age. However, the absolute minimum value, due to random hair cell activity, is 10^6 times less than the loudest sound we can hear.

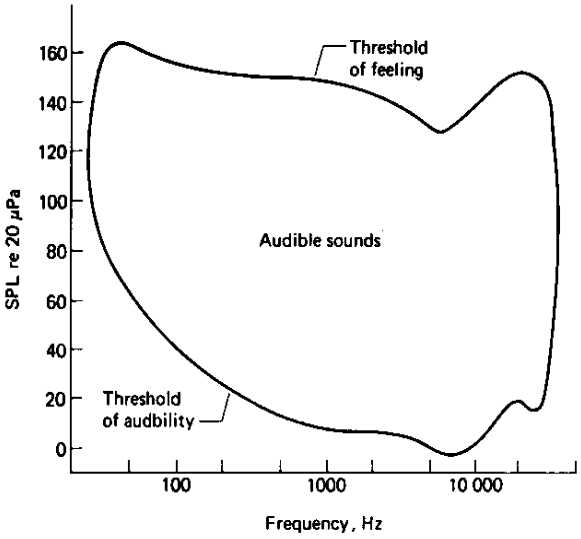


Figure 9

Minimum audible field for a range of frequencies.

- **The outer hair cells**

The outer hair cells are also distributed along the length of the BM. They react to feedback from the brainstem, altering their length to change the resonant properties of the BM. This causes the frequency response of the BM to be amplitude dependent.

Together, the inner and outer hair cells interact in an active feedback system, which increases the movement of the BM for quiet sounds, and suppresses it for loud ones. In effect, we have a sophisticated automatic gain control mechanism [16] in our ears[1]. This is linked to the temporal masking process described above, but in terms of amplitude sensitivity, it causes us to be logarithmic in our perception of loudness. This should come as no surprise to audio engineers – we usually measure levels and responses in decibels (dB) which is a logarithmic scale (20*log$_{10}$amplitude) – it is the AGC function of our ears which causes us to hear this way.

The net result so far is to take an audio signal, which has a relatively wide-bandwidth, and large dynamic range, and to encode it for transmission along nerves which each offer a much narrower bandwidth, and limited dynamic range.

---

[1] This mechanism is widely hypothesised, but unproven. It is difficult to measure directly as it only opperates in live, undamaged cochlea. All other phenomena have been verified in experiements using probed mamalian cochlea.

The function of each individual stage of the subsequent neural processing is less well understood (see Figure 1(d)):

- **The Cochlea Nucleus** is thought to sharpen features of the (now highly compressed) signal.
- **The superior Olivary Complex** is responsible for lateralisation of sound sources. It probably measures the inter-aural time delay via a series of delays and co-incidence detectors, as suggested in [17].

Little is known of the following stages of neural processing, other than their existence, and the fact that they give rise to our human "understanding" of the sounds around us as speech, music, and noise.

CONCLUSION

We have followed the path of a sound wave, from outside a listeners head, past the pinna and ear canal, into the cochlea, onto the basilar membrane, through the inner hair cells, along the auditory nerve, and into the higher processing centres.

The most critical factor to engineers is that any information lost due to the transduction process within the cochlea is not available to the brain – the cochlea is effectively a lossy coder. The vast majority of what we *cannot* hear is attributable to this transduction process.

So, if we take a "perfect" audio signal, and add some noise, can we say that the amount of noise we've added is a good indication of how bad the signal will sound? NO! If all the noise is hidden in spectral or temporal regions that we simply cannot hear, then the signal may still sound "perfect". Hence we *must* consider the human auditory system when assessing or processing audio signals. It's called "matching signals to the final receiver" [18].

Adapted from:

D. J. M. Robinson & M. J. Hawksford, "Time-Domain Auditory Model for the Assessment of High-Quality Coded Audio," preprint 5017, presented at the 107th convention of the Audio Engineering Society in New York, Sept 1999.

REFERENCES

[1] G. K. Yates, “Cochlea Structure and Function,” in B. C. J. Moore, Ed., *Hearing* (Academic Press, San Diego, California, 1995), pp. 41-74

[2] D. J. M. Robinson and R. G. Greenwood, “A Binaural simulation which renders out of head localisation with low cost digital signal processing of Head Related Transfer Functions and pseudo reverberation,” preprint 4723, presented at the 104th Convention of the Audio Engineering Society in Amsterdam, May 1998.

[3] K. C. K. Foo, M. O. J. Hawksford & M. P. Hollier, “Three-Dimensional Sound Localization with Multiple Loudspeakers using a Pair-Wise Association Paradigm and Embedded HRTFs,” preprint 4745, presented at the 104th convention of the Audio Engineering Society in Amsterdam, May 1998.

[4] H. Moller, M. F. Sorensen, D. Hammershoi, and C. B. Jension, “Head-Related Transfer-functions of Human-Subjects,” *J. Acoust. Soc. Am.*, vol 43, no. 5, pp.300-321 (1995).

[5] D. W. Battaeu, "The role of the pinna in human localization", *Proceedings of the Royal Society of London*, Series B, Vol. 168, pp. 158-180, (1967).

[6] C. J. Plack & R. P. Carlyon, “Loudness Perception and Intensity Coding,” in B. C. J. Moore, Ed., *Hearing* (Academic Press, San Diego, California, 1995), pp. 123-160.

[7] E. C. Carterette, “Some Historical Notes on Research in Hearing,” in E. C. Carterette & M. P. Friedman, Eds., *Handbook of Perception, Volume IV, Hearing* (Academic Press, New York, 1978), pp. 3-34.

[8] H. Traunmüller, “Analytical expressions for the tonotopic sensory scale,” *J. Acoust. Soc. Am.*, vol. 88, no. 1, pp. 97-100 (1990 July).

[9] T. Irino and D Patterson, “A time-domain, level-dependent auditory filter: The gammachrip,” *J. Acoust. Soc. Am.*, vol. 101, no. 1, pp. 412-419 (1997 Jan.).

[10] B. C. J. Moore, “Frequency Analysis and Masking,” in B. C. J. Moore, Ed., *Hearing* (Academic Press, San Diego, California, 1995), pp. 161-205.

[11] B. C. J. Moore, J. I. Alcántara, and T. Dau, “Masking patterns for sinusoidal and narrow-band noise maskers,” *J. Acoust. Soc. Am.*, vol. 104, no. 2.1, pp. 1023-1038 (1998 Aug.).

[12] R. Meddis, “Simulation of mechanical to neural transduction in the auditory receptor,” *J. Acoust. Soc. Am.*, vol. 79, no. 3, pp. 702-711 (1986 Mar.).

[13] R. Meddis, “Simulation of auditory-neural transduction: Further studies,” *J. Acoust. Soc. Am.*, vol. 83, no. 3, pp. 1056-1063 (1988 Mar.).

[14] R. Meddis, M. J. Hewitt, and T. M. Shackleton, “Implementation details of a computation model of the inner hair-cell/auditory-nerve synapse,” *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1813-1816 (1990 Apr.).

[15] E. Zwicker, “Dependence of post-masking on masker duration and its relation to temporal effects in loudness,” *J. Acoust. Soc. Am.*, vol. 75, no. 1, pp. 219-223 (1984 Jan.).

[16] T. Dau, D. Püschel, A. Kohlrausch, “A quantitative model of the “effective” signal processing in the auditory system. I. Model structure,” *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3615-3622, (1996).

[17] L. A. Jefress, “A place theory of sound localization,” *Journal of Comparative and Physiological Psychology*, vol. 41, pp. 35-39, (1948).

[18] E. Zwicker & U. T. Zwicker, "Audio Engineering and Psychoacoustics: Matching Signals to the Final Receiver, the Human Auditory System", *Journal of the Audio Engineering Society*, Vol. 39, No. 3, pp. 115-126, (1991).

Background reading:

B. C . J. Moore, *An Introduction of the Psychology of Hearing*, 4th ed. Academic Press, New York, 1997.

E. C. Carterette & M. P. Friedman, *Handbook of Perception, Volume IV, Hearing* Academic Press, New York, 1978.