E. Physiology and Psychophysics of the Human Auditory System

The most sophisticated information system on Earth is arguably the human mind. While there may exist similarly complex systems, we are ignorant of them. Most of human information is analyzed in terms of symbol systems, primarily language and mathematics. Indeed some people consider that there is no information, thought or meaning without language. These people apparently don't understand music.

While music can be represented as a symbol system using MB or other notation, this static form written on a page is not "music" but merely a guide to the performer for playing the music. By rendering the notation into sounds in the real world, the performer(s) create the reality of music from the thin sketch of information contained in the notation. For many people, music is something that they can only fully appreciate if they hear it. While trained musicians may be able to create music in their head by looking at notation, this is a difficult or impossible task for most people. In many cultures of the world, written notation is not used at all. The music created by these people is often sophisticated and complex, with deep informational and emotional content.

E.1 Human Auditory System

I do not claim to be an expert in audiology and psychophysics. This section was originally planned to be only a couple of pages, but the subject area proved fascinating (and immense). One thing led to another and this piece expanded greatly. Late in the game, Dr. Dean Ayers gave his feedback as a domain expert, pointing out both the flaws of my interpretations and analyses, as well as his opinions regarding the current views of mainstream audiological science. In short, research since (Bekesy, 1960) and (Harwood & Dowling, 1986) has shown that these early researchers' theories were not entirely correct. This is no surprise in scientific research of course. Rather than attempting an exhaustive re-write to fix things, I have cleaned up the egregious errors and left some of my

reports about Bekesy, Harwood & Dowling and others as they were, since for all we know, research in the near future may show that the theories may be more correct (or less) than the current "wisdom" indicates. Every theory in science will eventually be considered erroneous. It is useful to be aware of our history, and humble about our own theories and accomplishments.

In this section we present details of how the human ear transforms sound vibrations into patterns of nerve impulses in the auditory cortex. Human beings can generally hear sound frequencies from about 20 cycles per second (Hz) to 20,000 Hz at the lower and upper limits. Many people have a restricted range of frequency perception, so this is considered a best case scenario. Below 20 Hz and above 20,000 Hz perception of audio signals is possible, but the mechanism is different from what "hearing" is usually considered to be. Frequency of sound waves as measured by laboratory instruments corresponds to the human perception of pitch or tone: low frequencies are heard as low tones, high frequencies as high tones. However, there is more complexity in the human concept of tone than merely a short list of frequencies measured in the audio input signal. First, one frequency may not always be perceived as the same tone. Loudness of the sound can change its apparent pitch under some conditions, and in some frequency ranges. Second, combinations of distinct frequencies can create the perception of frequencies which do not "officially" exist in the sound source. Church organs take advantage of this effect (and have for hundreds of years) to create extremely low notes. The length of an organ pipe determines its fundamental (lowest) frequency and very long pipes are needed for very low notes. Alternatively, several pipes of shorter length can be used in combination, and by setting up a consistent set of frequency *differences* between pipes, the extremely low tone can be generated. This low tone only exists in the human ear and mind. Its "frequency" is not physically present in the external sound field (Plomp, 2002). There are other situations where frequency and tone do not map directly onto each other. This is a vast area of research that includes psychophysics, neuroscience and applied psychology.

Figure E.1.1 shows an overview of the human auditory data collection system, although it looks like Mr. Spock's ear, so maybe it's really Vulcan. Hearing begins with sound vibrations from the air striking the ear drum. Movements of the eardrum are transferred to the inner ear by three bones in the middle ear called the hammer, anvil and stirrup. These convert the physical scale of vibrations in gaseous air to a scale suitable for the liquid environment in the inner ear whose main component is the cochlea. The cochlea is a tapered tube rolled up in a spiral. Dividing the cochlea along its length is the basilar membrane and the Organ of Corti which contains neural vibration sensors, including small hair-like cells that are frequency sensitive. Different cells sense different frequencies depending on the cell's location along the length of the cochlea. The signals from these sensors are encoded into the nerve trunk and transmitted to the audio cortex. Information features are extracted starting with the initial actions of the Organ of Corti: the amplitude of the audio signal at various frequencies, and the timing relationships between frequencies. Frequency relationships such as the detection of correlated harmonics amongst the many frequencies present in the audio signal, or phase relationships between these component waveforms, may be partially encoded by the cochlea, but these more subtle distinctions may be perceived further up the processing chain in the audio cortex or brain. Generally, audiology research has failed to show evidence of the use of phase relationships by the ear (or the brain). My opinion is that this failure is at least partly due to the research techniques used. My own perception indicates subtleties which I attribute to phase discrimination by my ear/brain. Using the metaphor of sets of specific frequencies in an audio signal is useful, but is only a mathematical model of the data present in the signal. The real world sound is a complex three dimensional system of physically coupled motion patterns of air molecules. The decomposition of this system into a specific set of frequencies and phases is a convenient approximation, but can be misleading if interpreted literally for all situations. The activity of the audio cortex is far more sophisticated than that of any currently used computerized DSP and pattern matching techniques. For

example, phase information is known to be used at low frequencies for binaural detection of directional information from the external sound field. I have not read of any similar techniques used by researchers in computer music analysis.

Figure E.1.2 shows the cochlea by itself, from several viewpoints and a cross section. The basilar membrane including Organ of Corti can be seen at several turns of the spiral, dividing the two tunnel like chambers of the cochlea.

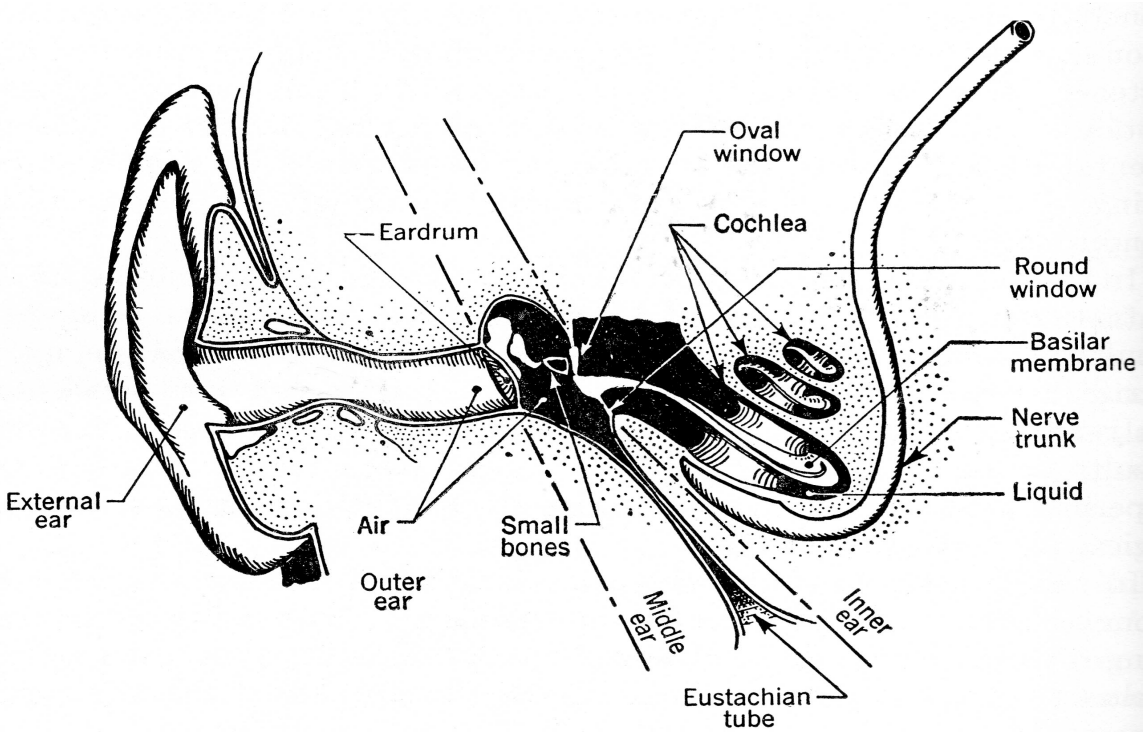

Figure E.1.1 Overview of Human Auditory Data Collection System
From (Beranek, 1954)

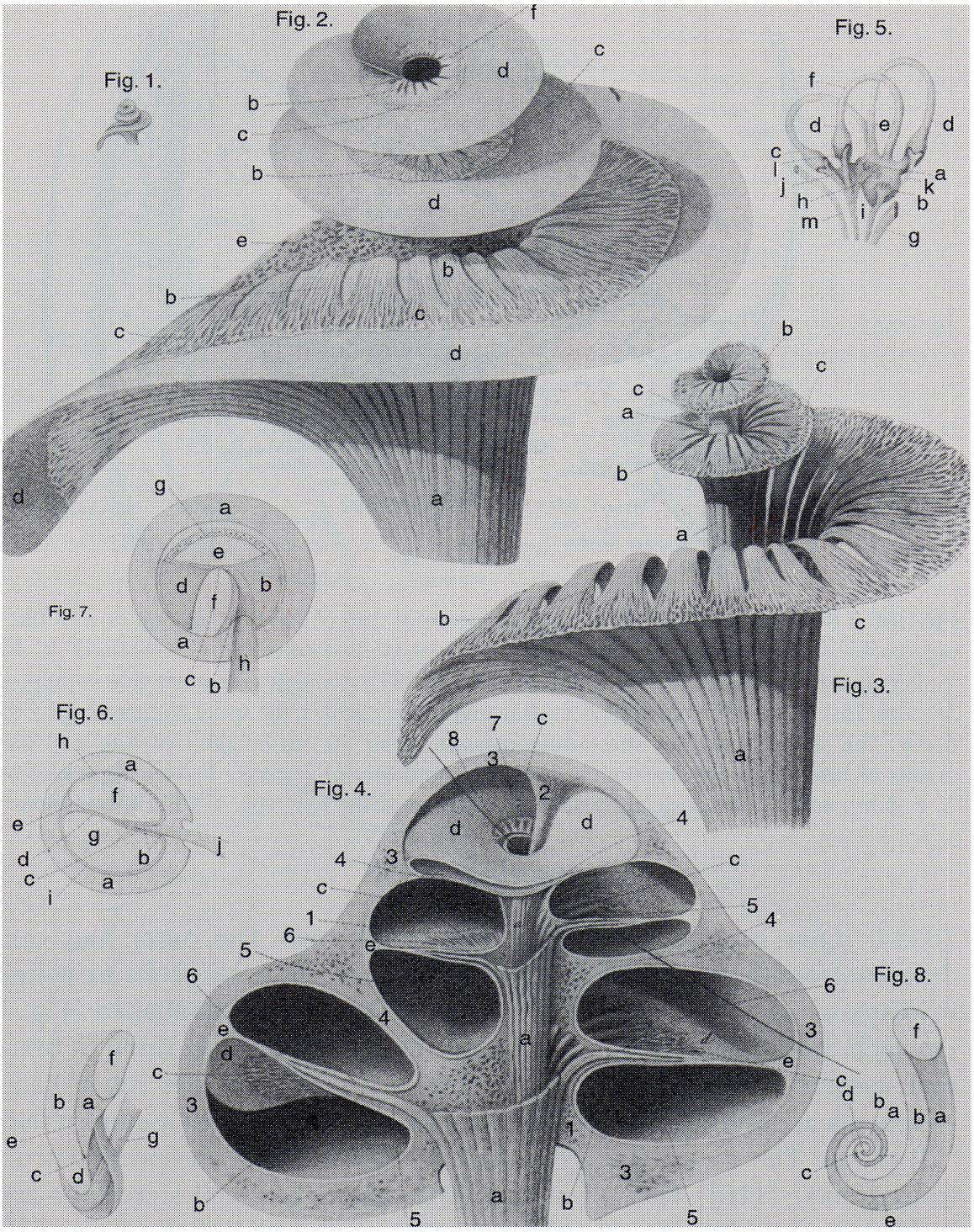

Figure E.1.2 The Human Cochlea

From (Møller, 2002). Original drawings by Brescher.

Figure E.1.3 shows a schematic cross section of the cochlea with details including the basilar membrane and Organ of Corti. This is a close-up of one of the turns in Figure E.1.2, showing most of the upper chamber and part of the lower chamber of the cochlea.

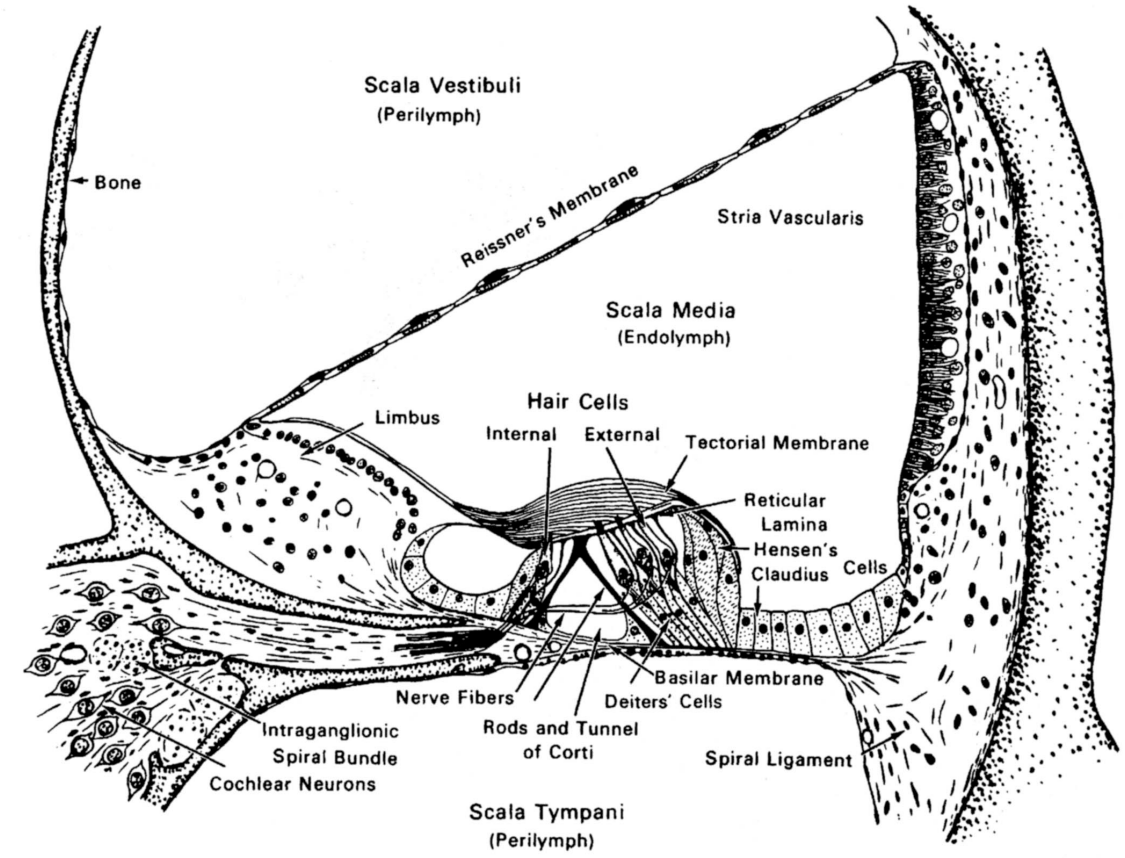

Figure E.1.3 Cross Section of Cochlea

From (Møller, 2002). Originally published by (Davis, et al., 1953) in *J. of Acoust. Soc. Am.* ***25****: 1180-1189.*

Figure E.1.4 shows an extreme close-up of a small section of the Organ of Corti and its frequency sensing hair cells, taken by a scanning electron microscope. Figure E.1.5 shows an even closer view of one of the several dozen hair tufts which are visible in Figure E.1.4. The hair tufts are colored yellow in Figure E.1.4, and orange in Figure E.1.5. These hairs move from the vibrations of the basilar membrane and the fluid in the space enclosed by the tectorial membrane, and then transmit their data to the nerve cells

colored pink in Figure E.1.4. It is known from (Bekesy, 1960) that fluid motion exists in the other chambers of the cochlea, but current theories hold that only the fluid motion in the tectorial chamber actually stimulates the hair cells.

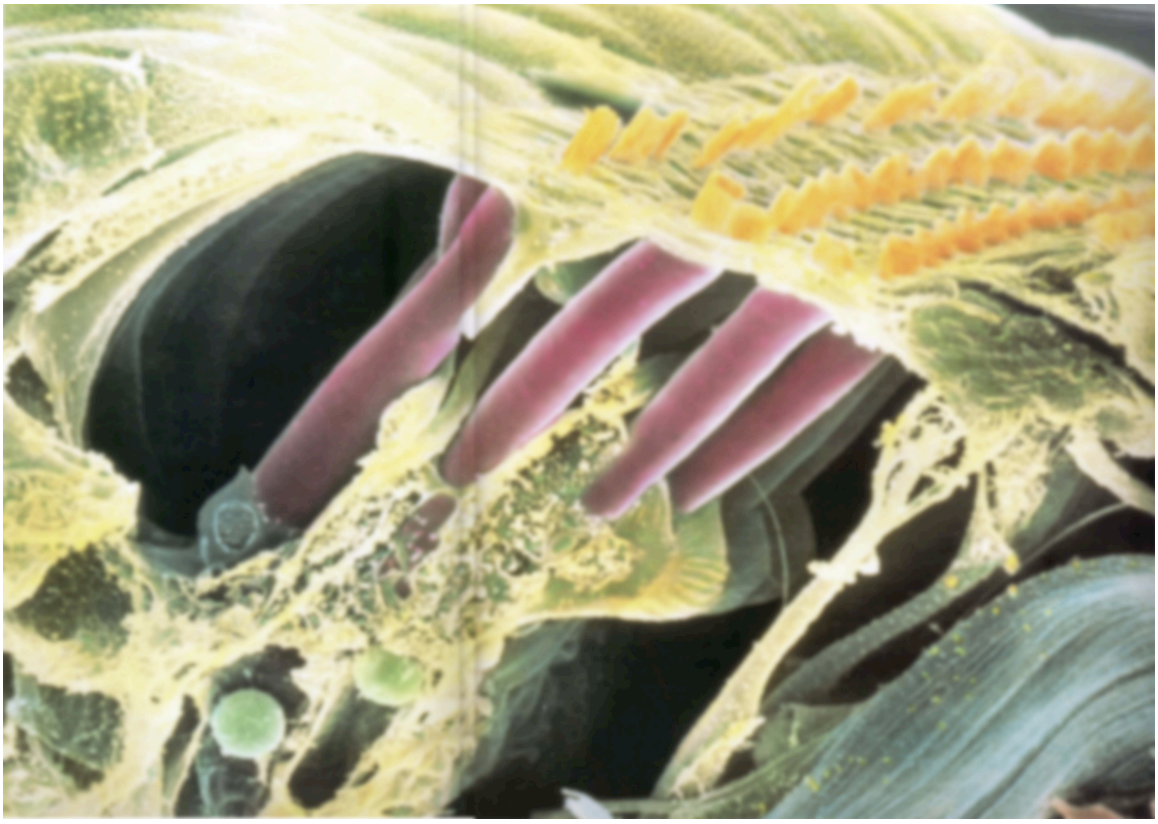

Figure E.1.4 Extreme Close-up of a Section of the Organ of Corti
From (Firefly, 2002)

These images are presented because they illustrate the complexity of the human audio data collection hardware. Whereas a computerized audio system uses two CD quality channels (44,100 samples/second, 16 bits/sample), the ear has millions of individual transducers, each collecting time based data at different locations, with patterns of time delays, phase and frequency values amongst these information channels being correlated by the neural networks in the audio cortex and brain. All of these information pathways essentially represent continuous functions in real time, while the computerized data form has very coarse granularity in both time and frequency. The action of the nervous system

is encoded as discrete nerve impulses rather than a continuous function in the mathematical sense, but the enormous number of different nerve impulses and pathways can be seen to approximate a true continuous function to a very fine granularity in both time and frequency. (Bekesy, 1960) reports detecting eddy currents in the cochlear fluids caused by sound vibrations. The hair cells in Figures E.1.4 and Figure E.1.5 respond to these fluid movements (in the tectorial chamber) as well as responding to vibrations in the basilar membrane. Neuroscience is currently charting these pathways.

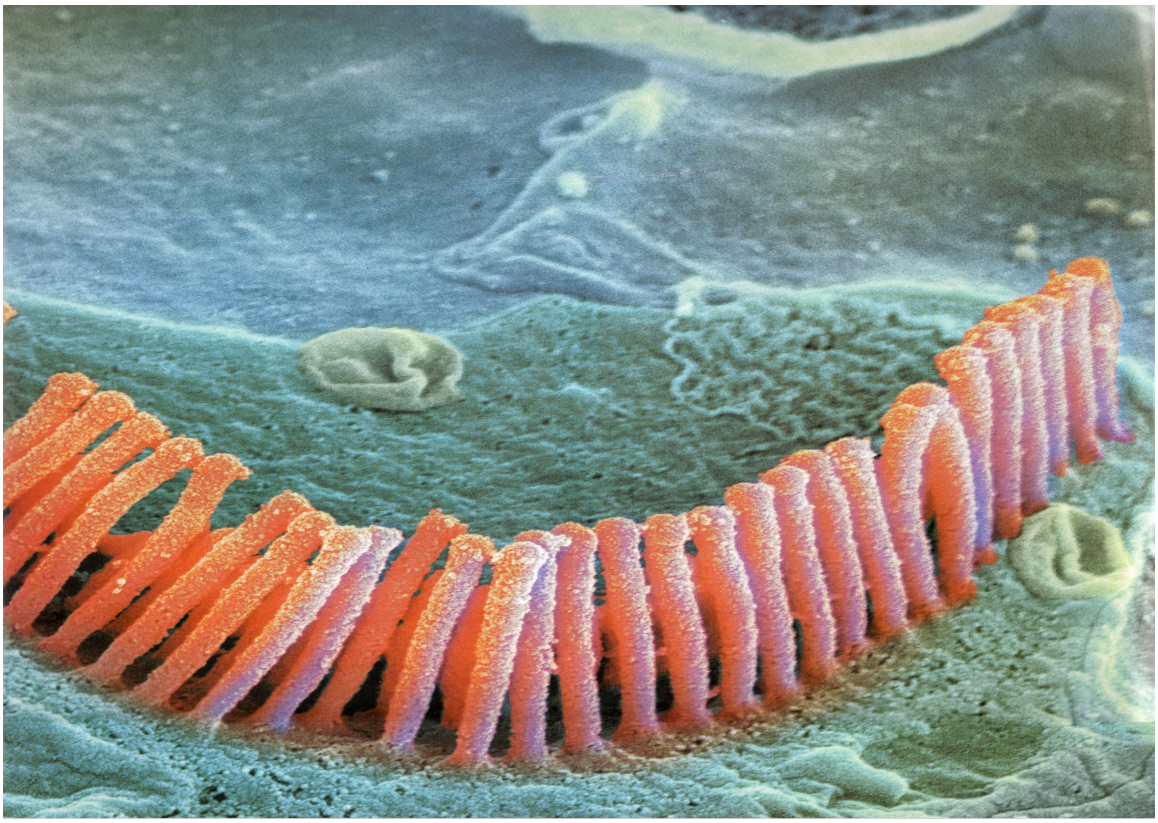

Figure E.1.5 Extreme Close-up of Frequency Sensing Cells

From (Firefly, 2002)

The transformation of audio signals into neural patterns and hence cognitive perceptions is complex and not completely understood. The main *result* of the process is the extraction of frequency and direction information from the incoming audio signal, and fusing the information into a continuous three dimensional perceptual reality. This is an

area for further research. Plomp gives a history of such scientific investigations, and cautions strongly to avoid being trapped by *a priori* thinking like reductionism (he calls it *microscopic view).* Apparently many researchers in the 150 years or so of acoustic science have believed that hearing is somehow a relatively simple process, much as early microbiologists thought that the interior world of a living cell is "formless protoplasm".

Helmholtz in the 19th century started the study of the physics of hearing, and proposed the "tuning fork", or "piano strings" model, which sees the cochlea as a fancy Fourier series analyzer that measures exact frequency components of the audio and passes them to the brain which extracts information and patterns. (Bekesy, 1960) showed that the cochlear response to frequency is more complex and subtle than merely being a row of finely spaced tuning forks as Helmholtz envisioned.

Audio vibrations enter the cochlea by the vibration of the stirrup bone on the oval window at the front of the cochlea. The signal is transmitted into the cochlea not as vibrations, but rather as a series of traveling wavefronts corresponding to the inward push on the eardrum by each incoming audio wave. These are transmitted by the bones of the inner ear (hammer, anvil, stirrup) to the oval window of the cochlea where they launch a wave disturbance in the upper chamber of the cochlea. Figure E.1.6 shows the eardrum and bones of the middle ear. The time for a wavefront to travel along the basilar membrane ranges from less than 0.1 millisecond for high frequencies to about 10 milliseconds for the lower limit of 20 Hz (Bekesy, 1960). Due to damping, the high frequencies only excite vibrations for a short distance, while low frequencies power curves peak at the far end of the cochlea.

The backward action of the incoming vibrations apparently has no effect on the frequency sensing hairs, which are only activated by the wavefront in the forward direction (Dowling & Harwood, 1986). These wavefronts are complex curves representing the instantaneous sum of many frequencies which are present in the audio signal, and the ear

extracts instantaneous frequency information from them. With each wavefront, different frequency information is induced as vibrations on the basilar membrane. The tuning of the basilar membrane creates power curves (traveling waves) from the incoming wavefronts, rather than vibrating at a particular frequency *per se.* Figures E.1.7 through E.1.11 show several aspects of the traveling wave phenomenon from (Bekesy, 1960) who mapped the frequency response of the basilar membrane. Note that the wave shapes of the traveling waves have power peaks at different locations along the length of the basilar membrane. Additionally, as the wavefront moves through the cochlear fluid, the basilar membrane shape flexes into unique shapes determined by the frequency and loudness information. The action of the basilar membrane short circuits the wavefront of particular frequencies at the location of the cochlea which is sensitive to that frequency. The energy from the wavefront is thus transmitted to the lower cochlear chamber by the basilar membrane, reducing or eliminating the energy at that frequency in the upper chamber beyond the sensitive location for that frequency. The entire collection of the various vibrational responses is probably encoded as a Gestalt as well as being decomposed by frequency (Plomp, 2002).

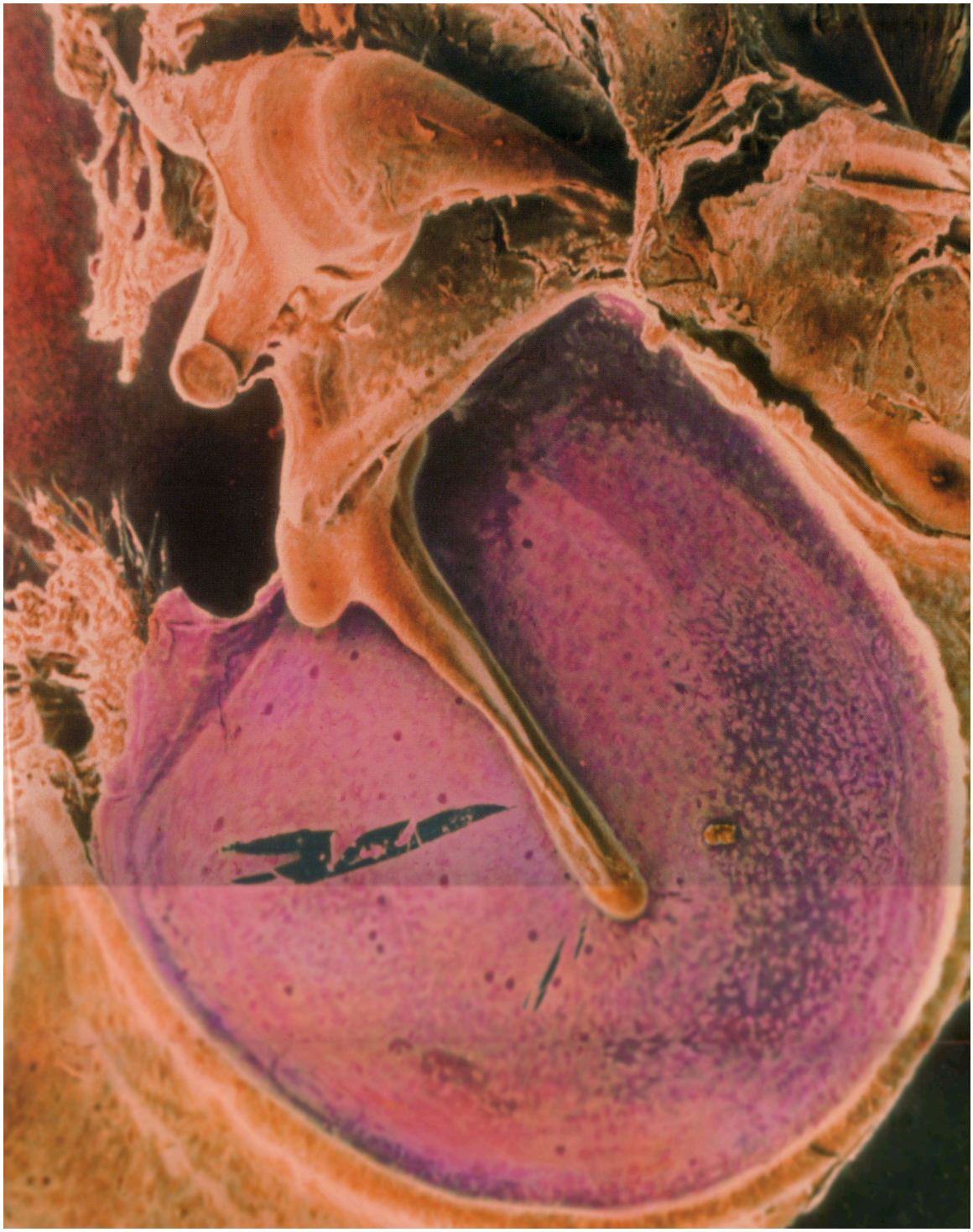

Figure E.1.6 Eardrum and Middle Ear Bones

From (Firefly, 2002)

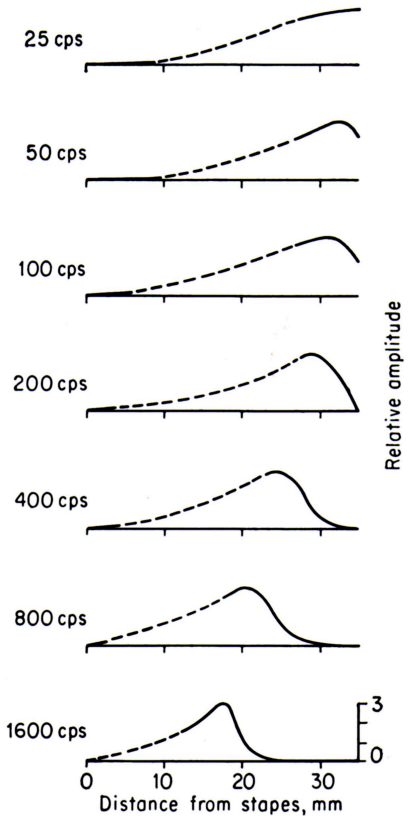

Figure E.1.7 Power vs Distance Curves in Cochlea for Several Frequencies

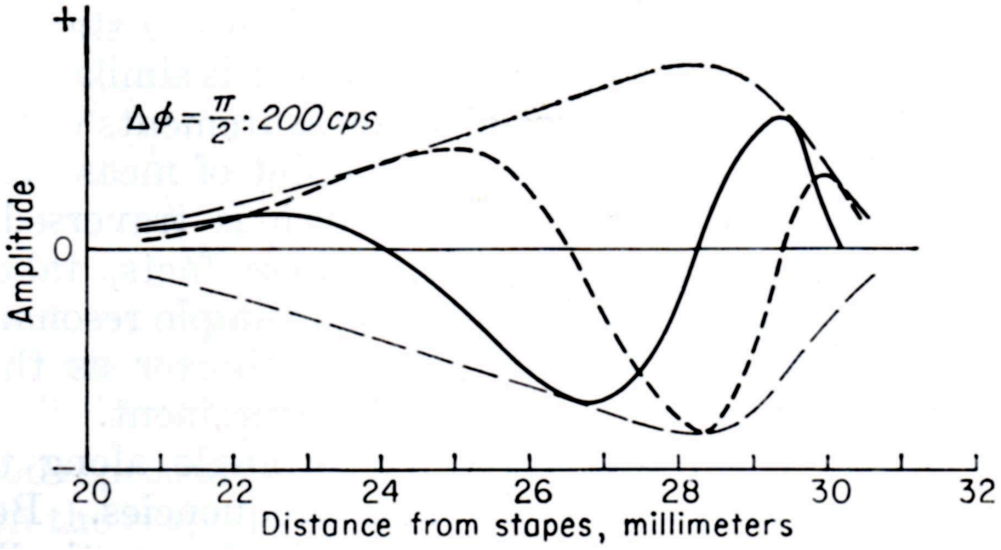

Figure E.1.8 Traveling Wave for 200 Hz at Several Moments in Time

From (Bekesy, 1960)

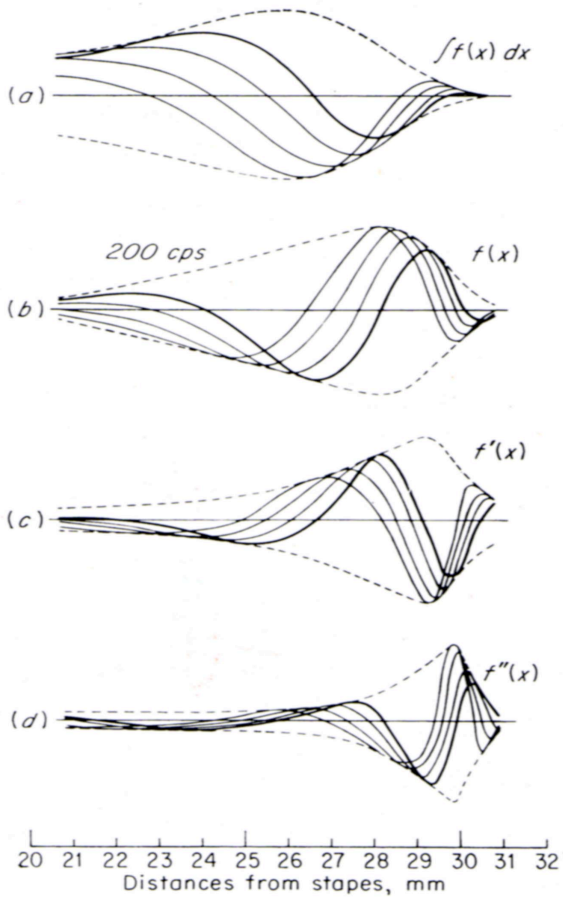

Figure E.1.9 Waveform, First, Second Derivatives, and Integral for 200 Hz

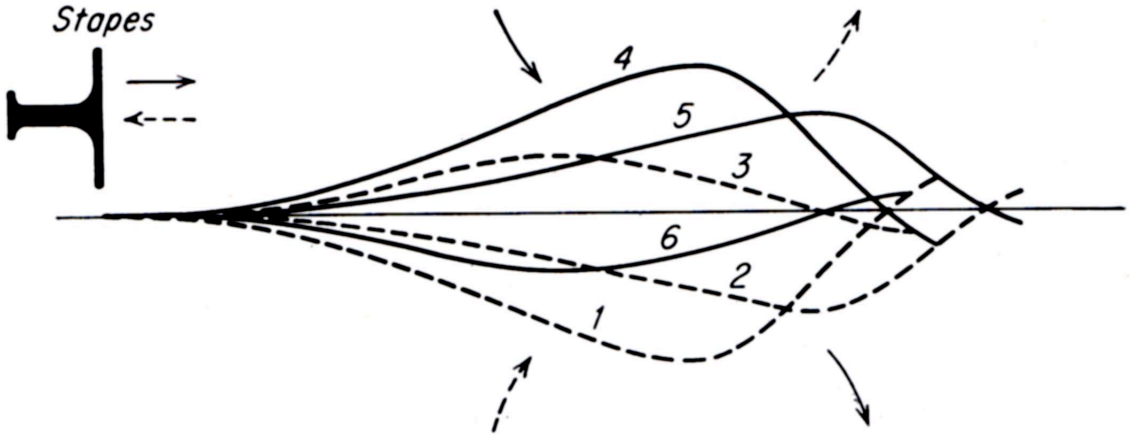

Figure E.1.10 Traveling Wave, Showing Generation of Eddy Currents

From (Bekesy, 1960)

At low frequencies, below about 100 Hz, the entire basilar membrane vibrates as a whole. Above about 200 Hz, the frequency sensors mentioned above are affected by the incoming vibrations. From approximately 200 to 2000 Hz there are two mechanisms for extracting information from the audio signal: the frequency sensors, and detection of beat patterns in the incoming waveform. The beat patterns are caused by constructive and destructive interference between waves present in the external three dimensional space, much as waves on the surface of a pond have peaks and troughs as they intersect each other. The results of both the beat and frequency detection mechanisms are interpreted as tonal information at higher cognitive levels in the audio cortex. The incoming beats directly trigger nerve impulses, while the frequency sensors respond to wave shapes of the incoming vibrations. Nerve cells have a limit to how fast they can fire, and so above approximately 2000 Hz (0.5 milliseconds), only the frequency sensors extract information. The beat pattern extraction mechanism operates down to extremely low frequencies, well below the 20 Hz "limit" of human hearing. At very low frequencies, the beat patterns are perceived as individual events rather than tones (Dowling & Harwood, 1986). NB: the beat phenomenon as reported by Dowling & Harwood may be erroneous. Events with fast onsets such as percussion sounds have time scales for the onsets in the frequency range of the beat mechanism. Thus there may be some recognition of such events in the front end of audio processing, as well as later in the cortex where neural processing stages extract timing information from the changing input stream. This phenomenon is an important consideration for the Ile Aye caixa experience described in Appendix A.

Each frequency sensing hair is broadly tuned, responding to a range of frequencies. Due to the traveling wave effect, the power distribution along the cochlea takes on different waveforms depending on the frequencies in the signal as described by (Bekesy,1960). The spatial variations of the power distribution is used for distinguishing different frequencies. The amplitude increases as the wavefront moves into the area of the cochlea that is tuned to the frequencies corresponding to the incoming wave shape. After

the wavefront passes through the tuned section, the amplitude begins to decrease. The detection of the peak is enhanced by lateral inhibition from nearby frequency sensors whose signal strength is less than the peak. The neurons in the audio cortex subtract this nearby data from the power distribution waveform, sharpening the peak relative to the broad waveform. This is one of the reasons we hear precise tones rather than a smeared combination of frequencies. (Dowling & Harwood, 1986). NB: Dowling & Harwood may be incorrect about their lateral inhibition report. While lateral inhibition is proven to have a significant role in visual perception, research in audiology has not produced clear evidence that lateral inhibition operates in hearing perception.

There is substantial evidence that the ear produces sounds by itself, called *oto-acoustic emissions.* These sounds have been detected by sensitive microphones placed in the outer ear. It is not entirely clear what role these sound emissions play in hearing, but theories that they assist in frequency discrimination are the most prevalent[3].

Mathematical models of the basilar membrane vibrations are commonly used in neuroscience research. Figure E.1.11 is an example which shows patterns of a normal frequency response, and one with a damaged basilar membrane. We include this because it provides two insights. First, it is a clear intuitive example of how loss of proper vibrational response of the cochlea contributes to hearing impairment. Second, it actively shows how the physiology of the ear helps to generate forms of information from the input sound vibrations, due to the nonlinear resonant action of the tissues and fluids. Artificial neural networks rely heavily on nonlinear response for teasing out subtleties in complex entanglements of signals.[4] DSP uses linear processing for the most part, and consequently is more limited in the availability of information than a nonlinear system.

---

[3] Dr. Dean Ayers, Southern Oregon University, Department of Physics

[4] Dr. Charles Jorgensen, Senior Research Scientist. Neuro-Engineering Lab, NASA Ames Research Center. Mountain View CA USA.

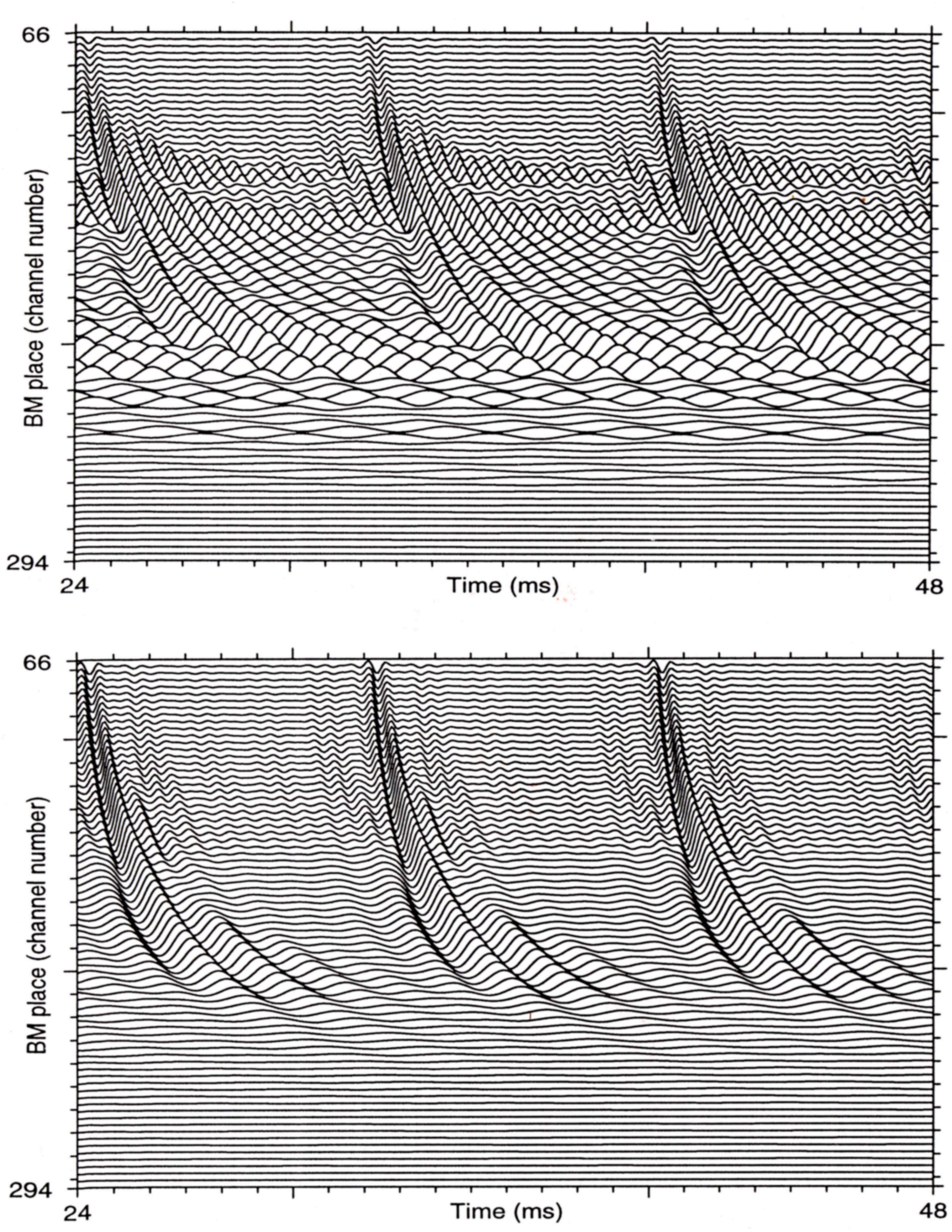

Figure E.1.11 Mathematical Model of Vibrations in the Basilar Membrane

From (Giguere & Smoorenburg, 1999) in (Dau et al., 1999)

Other mechanisms for transforming data into information features will no doubt be discovered by neuroscience in the coming years. We have scratched the surface with our readings in (Dau et al. 1999). We believe that percussion events are very useful in this type of research because the onset of events is very short, typically from 1 millisecond up to about 40 milliseconds. These quick events are easier to track in the audio cortex using EEG than are more complex sounds, such as speech or melodic instruments. Percussion events are more complex than the audio signals used in psychology research which are typically simple sine waves or square pulses. The complexity of the percussion sounds can be used to study the pattern recognition pathways in the audio cortex.

E.2 Psychological Studies of Human Perception

(Fraisse, 1982), (Deutsch, 1992) and others have done perceptual studies that identify certain time ranges as "natural" for human imitative tapping and rhythm. These are mostly in the tempo range of standard music pieces. If tempo increases or decreases beyond the natural range, most people shift to the next higher or lower synchronized pattern whose timing fits in the natural range.

This background information is relevant to the story in the appendix about learning the caixa batida from Ile Aye. We also reference human timing perceptual issues in our descriptions of creating seamless rhythmic loops. Our experience leads us to the conclusion that some of the standard psychological models of reaction time and human response time in general are inadequate. All of the studies we have read have tested subjects using isolated sequences of events. We have found that temporal context of patterns of events is an important part of a perceptual mechanism that is temporally more fine grained that the standard models of human time perception.

E.3 Human Emotions and the Meaning of Music

The connection between music and emotions is widely, perhaps universally, recognized. "Music hath charms to tame the savage beast" is an old folk saying, and music

has been used for therapeutic medical purposes. Recently, researchers have applied modern methods such as electro-encephalogram (EEG) to monitor a subject's physiological responses during music therapy (Fox, 2005). This is strong support for our contention that music, health and emotional well being are closely related. It also supports our purpose to facilitate the learning, teaching and playing of music by using technical approaches that help people understand non symbolic information and other subtleties which are fundamental to music.