Chapter 6

The Human Auditory and Visual Systems

In this chapter, some basic properties of the human auditory and visual systems are discussed. Understanding properties of human perception can help in designing algorithms and devices where either the end user is a human or even when there is a human somewhere in the loop. Many systems dealing with some aspect of multimedia fall into this category – for example television, computer monitors, CD players, telephones, digital cameras, other imaging technologies (for medicine, surveillance, etc.), printing technology, and the internet. With a human as the ultimate user of the data processed by these systems, the technology can be tailored to suit the capabilities and limitations of humans.

There is a second way in which understanding properties of human, or more generally, biological systems can help in designing engineering systems. For systems whose goals are similar to tasks carried out by biological systems, it is natural to consider "biologically-inspired" engineering designs. This is particularly true in trying to design "intelligent" systems where the tasks are extremely difficult and relatively little is understood about how best to proceed. Such areas include, speech, image, and video analysis, learning and adaptation, pattern recognition, and intelligent control. For intelligent systems, the main difficulty is usually not how to design interfaces with humans. On the contrary, the goal of such systems is often to specifically avoid human intervention. The main difficulty is simply figuring out what to do with the available data to carry out the desired task. Understanding how biological systems function can be an extremely valuable guide in these cases. In fact, one common school of thought is that the best way to make progress in "intelligent systems" is to understand and mimic the most intelligent systems known, namely humans.

†Lecture Notes for ELE201 Introduction to Electrical Signals and Systems.

‡Thanks to Richard Radke for producing the figures.

A great deal is known about the structure and function of the human auditory and visual systems In the following sections, only a very brief account of some properties most relevant to us are discussed.

6.1 The Human Ear

The human ear consists of three main components called the external (or outer) ear, the middle ear, and the inner ear. The external ear consists of the pinna (the visible flesh we informally called the ear), the outer ear canal, and the eardrum (or tympanic membrane). See Figure XX. The external ear gathers sound (variations in air pressure), carries them through the outer ear canal, and to the eardrum, which separates the outer ear from the middle ear.

The middle ear consists of three small bones call the malleus (hammer), incus (anvil), and stapes (stirrup). These convert minute vibrations of the eardrum into amplified movements of the stapes. This in turn causes motion of fluid in the inner ear.

The inner ear consists of a fluid-filled spiral structure with about 2 1/2 turns that is separated into three compartments called the scala tympani (tympanic canal), scala vestibuli (vestibular canal), and the scala media (cochlear duct). The floor of the cochlear duct is the basilar membrane that holds the organ of Corti. The organ of Corti consists of an intricate network of hair cells and supporting cells. The hair cells convert the mechanical "signal" of fluid motion into an electrical signal.

The cochlea seem to have frequency selectivity. That is, different parts of the cochlea respond to different frequencies in the original air pressure variations (which in turn correspond to different frequencies of the sound). If the cochlea was unwound, the portion connected to the middle ear is stiffer and responds to higher frequencies, while the other end is more flexible and responds to lower frequencies. Exactly how this is done is not well understood, but remarkably the human ear is effectively doing a type of Fourier transform!

6.2 Properties of the Auditory System

Various properties of the auditory system are important in the design of technology dealing with sound.

Frequency Range Humans are sensitive to frequencies in range of about 15 Hz to 20,000 Hz. Actually, these are somewhat optimistic figures and many people (especially as they age) have a range smaller than this, perhaps more like 20 Hz to 16,000 Hz. This frequency range immediately impacts design choices such as the sampling rate in digital audio. For example, we know that to avoid aliasing, we should sample at a rate faster than twice the highest frequency present in the signal. Although, many sounds may have frequencies higher than 20,000 Hz, we can first remove these frequencies without altering the perceptual

quality of the sound to humans (by filtering as discussed later in Chapter XX). Then, the highest frequency will be 20,000 Hz. In this case, sampling faster than 40,000 Hz will retain all the information necessary to fully reconstruct the filtered signal. The sampling frequency of digital audio is typically 44,100 Hz, which was chosen with human perceptual properties in mind. Oversampling (that is, choosing a frequency larger than the Nyquist rate) not only provides a buffer, but also simplifies processing in reconstructing an analog signal from samples.

Frequency Discrimination This refers to the ability to distinguish sounds at two different frequencies. The discrimination ability of humans is about 3 or 4 Hz for sounds with frequencies between 15 Hz and 2,000 Hz. Above 2,000 Hz, the change in frequency needed for discrimination is about 0.3 % of the frequency of the sound. So for a sounds near 10,000 Hz, the frequencies would have to differ by about $(0.003 \times 10,000) = 30$ Hz to be distinguishable. This gives about 600 distinguishable frequencies in the range of 15 Hz to 2,000 Hz, and about 720 distinguishable frequencies between 2,000 Hz and 16,000 Hz. Thus, humans can distinguish over 1,300 frequencies.

Sensitivity This refers to the weakest signal that is just audible at a given frequency. A standard way to measure sensitivity is with respect to a intensity of a reference signal. Humans are most sensitive to sounds in the frequency range between 1,000 Hz and 3,000 Hz. If we let P_r denote the intensity of the weakest signal in this range, then we can measure the sensitivity at other frequencies with respect to P_r. Because the dynamic range is so huge, it is standard to compare intensities to P_r on a logarithmic scale. That is, we define the intensity of P_t in *decibels* (or dB) as

$$\text{intensity of } P_t \text{ in dB } = 20 \log_{10} \frac{P_t}{P_r}$$

By definition, P_r is at 0 dB, and physically this corresponds to about 0.2×10^{-9} of atmospheric pressure. (Atmospheric pressure is about 14.7 pounds per square inch.) We can plot the just audible intensity as a function of frequency as shown in Figure XX. Every 20 dB corresponds to a factor of 10 in signal amplitude. From the figure we can see that the sensitivity of human hearing varies by several orders of magnitude as a function of frequency.

Amplitude Discrimination This refers to the ability to distinguish changes in amplitude. Regardless of the level, humans can generally detect about a 3 dB change in intensity. However, since dB measures the log of the ratio of intensity with respect to the reference intensity, we need about the same *proportional* change in intensity to be able to just detect the change (rather than the same additive change in intensity). Specifically, for a given intensity,

the just-noticeable new intensity satisfies

$$20 \log_{10} \frac{\text{just noticeable new intensity}}{\text{intensity}} = 3 \text{ dB}$$

Therefore,

$$\log_{10} \frac{\text{just noticeable new intensity}}{\text{intensity}} = 0.15$$

so that

$$\text{just} - \text{noticeable new intensity} \approx 1.4 \times \text{intensity}$$

Equivalently, we can write

$$\frac{\text{just noticeable } \textit{change} \text{ in intensity}}{\text{intensity}} = 0.4$$

This type of relationship (that the just-noticeable change divided by intensity being constant) is quite common in perceptual phenomena and is often referred to as Weber's Law or the Weber-Fechner Law.

Amplitude Range This refers to the loudest sound we can hear without pain compared with the weakest detectable sound (which has intensity P_r). The amplitude range is about 120 dB. This means that the loudest sound (without pain) is about 10^6 times the intensity of P_r, which is a huge dynamic range.

6.3 The Human Eye

Light enters through the lens and falls on the retina. Receptors in the retina are sensitive to light of wavelengths approximately 350-750 nm, and can operate over roughly 10 orders of magnitude in illumination, which is a remarkable dynamic range.

There are two types of photoreceptors in the human retina: rods and cones. The rods are long and thin and are primarily responsible for *scotopic vision*, which refers to visual capabilities in low illumination levels.

The cones are shorter and thicker than the rods and are responsible for *photopic vision*, vision in high illumination levels. The cones are also responsible for color vision, which is discussed further in Section 6.6.

The rods and cones are connected to various types of other cells which join to form the optic nerve that connects to higher brain centers.

6.4 Luminance Versus Brightness

Clearly the "brightness" perceived at a point is related to the amount of light falling on the receptors corresponding to that point. However, it turns out the *perceived* brightness at a point is also a function of the "brightness" of neighboring points. Thus a distinction needs to be made between our *perception*

of brightness, which is a subjective quantity, and the actual energy incident on the receptors at a point and the resulting response of the receptors, which can be precisely defined.

Although the retina is spherical and there are not receptors at every point on the retina, for simplicity a point on the retina will be denoted by (x, y) and we will talk about the energy falling on a small retinal patch at (x, y). The energy which falls on a retinal patch at a given point is normally light which has been reflected off of some object patch in the visual scene. This energy is a function of a number of quantities such as the amount of light falling on the object patch, the reflectivity and geometry of the object patch, the amount of light collected by the optical system (namely the eye), etc. The energy falling on a retinal patch at (x, y) is called the *irradiance* and will be denoted $E(x, y)$. The irradiance evokes a response in the receptors in the retinal patch, and the response is called the *luminance* at (x, y), denoted $I(x, y)$. The receptors respond differently to light at different wavelengths. The response can be be characterized by a function of wavelength $V(\lambda)$ which is called the *luminous efficiency function*. The luminance at (x, y) is given by

$$I(x, y) = \int_0^\infty E(x, y, \lambda) V(\lambda) d\lambda$$

where $E(x, y, \lambda)$ is now the irradiance at wavelength λ.

From the expression above, we see that the luminance at a point is independent of the luminance at surrounding points. On the other hand, there are two standard phenomena which show that the perceived brightness (or apparent brightness, or simply brightness) at a point depends not only on the luminance at that point, but also on the luminance at neighboring points. These phenomena are called Mach bands and simultaneous contrast, and they illustrate nicely the difference between luminance and brightness.

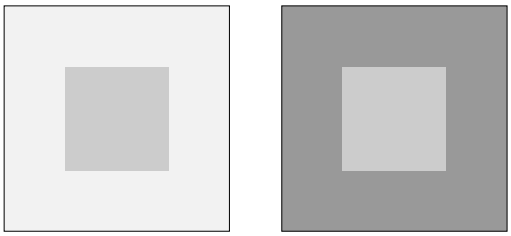

Figure 6.1: Simultaneous contrast.

Figure 6.1 illustrates the simultaneous contrast phenomenon. The two inner squares have the same luminance but the one on the right appears brighter. This illustrates the fact that our perception of brightness depends on luminance *contrasts* as well as absolute luminance values. Let I_0 be the luminance of object, and let $I_s = I_0 + \Delta I$ be the luminance of surround which results in a

just-noticeable difference in intensity. Weber's Law states that

$$\frac{|\Delta I|}{I_0} = constant$$

For just noticeable differences

$$\frac{\Delta I}{I} \approx d(\log I) = constant$$

Thus, the change in intensity required for a just-noticeable change in contrast depends logarithmically on the ambient intensity level.

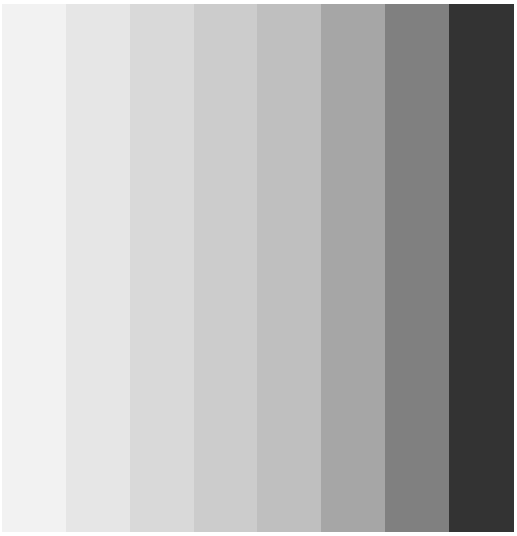

Figure 6.2: Mach bands.

Figure 6.2 illustrates the Mach bands phenomenon. Each band has constant luminance, but one perceives overshoots and undershoots in brightness in the neighborhood of the band boundaries. This phenomenon can be explained by the "on-center off-surround" or "lateral inhibition" of visual system.

6.5 Spatial and Temporal Properties

Acuity refers to the ability of the eye to detect fine spatial detail. The response of the visual system to sinusoids has been extensively studied. The ability to resolve spatial detail depends not only on the actual physical dimensions of the object being viewed, but also on the distance of the object to the observer. The important quantity is the angle over which the object extends as measured from the eye — i.e. the physical dimension of the detail per degree. That the response of the human visual system to sinusoidal patterns is greatest for approximately 3-10 cycles/degree.

Bloch's Law Light flashes of different duration but same energy are indistinguishable for durations less than approximately 30 ms for moderate light levels. This duration is somewhat larger when eye is adapted to low light levels.

Critical Fusion Frequency A light flashing at a rate faster than a certain rate called the Critical Fusion Frequency (CCF) is indistinguishable from a steady light of same average intensity. The CCF is generally around 50-60 Hz. This has implications on refresh rates required in image display devices, such as televisions or computer monitors, to prevent flickering.

6.6 Color Perception and Representation

Most of the work done in image processing deals only with gray-level images. However, color is becoming increasingly important in image processing and it is worthwhile to know the basics of color for several reasons. The development of new technology in sensors, display, and printing is making the use of color more widespread. In some applications color can be a powerful cue for various image analysis tasks such as edge detection and image segmentation. Another important area is the use of *pseudo-color* for the enhancement of gray scale images. Gray levels can be mapped to different colors in order to bring out or emphasize particular features.

Monochromatic light refers to light of a single wavelength. *Chromatic* or *colored* light refers to light which has energy distributed at various wavelengths. To completely specify a chromatic light source, one needs to specify the energy distribution as a function of wavelength, which is referred to as the *spectral characteristics* or the *spectral distribution* of the light.

However, we know that we can get "all colors" by combining three primary colors — red (R), green (G), and blue (B). If the primary colors are modeled as monochromatic light sources at three different wavelengths (corresponding to R, G, B), then mixing colors would still only allow light with energy at these three wavelengths. How is it possible to generate all colors by combining different levels of R, G, B since we certainly can't generate all spectral distributions this way? The reason is that there are only three different types of cones which are responsible for color perception.

The three cones have different absorption spectra $S_1(\lambda), S_2(\lambda), S_3(\lambda)$ which are shown in Figure 6.3. The absorption spectrum $S_3(\lambda)$ peaks in the blue region of the spectrum and so corresponds to B, $S_2(\lambda)$ corresponds to G, and $S_1(\lambda)$ corresponds to R (although its absorption spectrum actually peaks in the yellow-green region of the spectrum).

Given any spectral distribution of light, $C(\lambda)$, the color sensation produced depends only on the spectral responses of each the three types of cones. The response of the i-th receptor ($i = 1, 2, 3$) can be written as $\alpha_i(C) = \int S_i(\lambda)\, C(\lambda)\, d\lambda$. Any light sources that give the same responses α_i for each of the three receptors will be perceived to be identical, even though the actual spectral distributions of the light sources might be different.

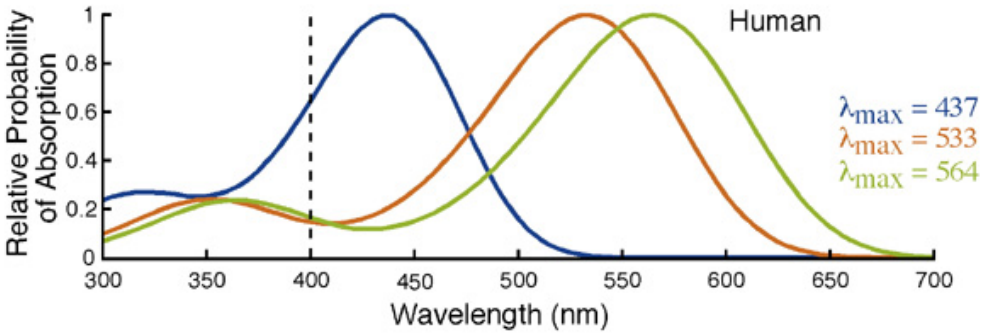

Figure 6.3: Absorption spectra of human cones.

Color Matching and Reproduction Given three light sources, how can we match or reproduce arbitrary colors? Let $P_1(\lambda), P_2(\lambda), P_3(\lambda)$ be three sources called primary sources. We want to match a color $C(\lambda)$ using a combination of these three sources. That is we want to find constants β_k such that

$$\sum_{k=1}^{3} \beta_k P_k(\lambda)$$

is perceived to be the same color as $C(\lambda)$. From our discussion above, we only require that the responses of each of the three receptors be the same. Therefore, we want

$$\begin{aligned} \alpha_i(C) &= \int \left[\sum_{k=1}^{3} \beta_k P_k(\lambda) \right] S_i(\lambda) d\lambda \\ &= \sum_{k=1}^{3} \beta_k \int S_i(\lambda) P_k(\lambda) d\lambda) \\ &= \sum_{k=1}^{3} \beta_k a_{i,k} \end{aligned}$$

where $a_{i,k}$ is the response of the i-th receptor type to the k-th primary light source. Recall that $\alpha_i(C) = \int S_i(\lambda) C(\lambda) d\lambda$ is the response of the i-th receptor to the source $C(\lambda)$.

In matrix form, we have

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} \alpha_1(C) \\ \alpha_2(C) \\ \alpha_3(C) \end{bmatrix}$$

Thus, $\bar{\beta} = A^{-1}\bar{\alpha}(C)$ where A^{-1} is the inverse of the matrix $A = [a_{i,k}]$.

Formally this gives us the solution to our original question. To reproduce the color $C(\lambda)$ we use the k-th primary source with intensity β_k. However, note that

for some colors $C(\lambda)$ the formal solution may result in negative β_k. In this case, it means that with the sources P_1, P_2, P_3 we cannot physically generate colors with negative β_k's. It would be nice to find three primary sources for which *all* colors are producible — i.e., result in positive β_k. However, there is no known set of such primary sources. Two natural attempts at selecting primaries which will allow the reproduction of all colors are (1) let $P_i(\lambda) = S_i(\lambda)$ or (2) Define $P_i's$ by requiring the $\beta's$ to equal $\alpha_i(C)$ In the first case, it can be shown that for some colors at least one of the β_k will be negative. In the second case, it can be shown that the $P_i(\lambda)$ are not physically realizable since at least one of the $P_i(\lambda)$ will be negative for some λ. One can attempt to implicitly define a set of primaries such that the β_k are always positive, but again the resulting primary sources will not be physically realizable. Thus, one has to settle for a set primary sources which allow reproducing a sufficiently rich set of colors. The set of colors producible by a set of primaries is called the *gamut* of the primary sources.

The β_k can be thought of as a representation for a color $C(\lambda)$ in terms of primaries $P_1(\lambda), P_2(\lambda), P_3(\lambda)$. One common standard set of primary sources is the C.I.E. (International Commitee on Color Staandards) standard primary sources in which the $P_i(\lambda)$ consist of monochromatic sources corresponding to R, G, and B. Specifically, the sources are $P_i(\lambda) = \delta(\lambda - \lambda_i)$ where λ_i=700 nm (R), λ_2=546.1 nm (G), and λ_3=435.8 nm (B).

Often, rather than using the β_k directly to represent colors, the primary sources are calibrated against a reference white light source $W(\lambda)$. In this case, a color $C(\lambda)$ can be represented by the normalized values

$$T_k(C) = \frac{\beta_k}{w_k}$$

where w_k is the amount of the k-th primary needed to match the reference white light source $W(\lambda)$. These normalized values are called the *tristimulus values* of the color $C(\lambda)$.

Chromaticity Diagram Another alternative representation that is commonly used are the *chromaticities* of a color which are defined by

$$t_k = \frac{T_k}{T_1 + T_2 + T_3}$$

However, note that $t_1 + t_2 + t_3 = 1$, so that only two of the chromaticities are independent. To get a complete representation of the color, two of the chromaticites need to be augmented with another parameter. This third parameter is the *luminance* of the color which is defined as

$$Y = \int \left(\sum_{k=1}^{3} \beta_k P_k(\lambda) \right) V(\lambda) d\lambda = \int \left(\sum_{k=1}^{3} w_k T_k P_k(\lambda) \right) V(\lambda)$$

Thus the color is represented by (t_1, t_2, Y). The luminance Y is a measure of the total response of the visual system to the color or more informally the "bright-

ness" of the color, while the first two components t_1, t_2 are the *chrominance components*, which specify the specific shade of the color.

Color Coordinate Systems The preceeding discussion on color representations was based on a particular (fixed) set of primaries $P_1(\lambda), P_2(\lambda), P_3(\lambda)$. As mentioned, one common choice for the primaries is the C.I.E. R,G,B standard primary system. One could choose a different set of primaries and obtain a different color representation or *color coordinate system*. A number of different color coordinate systems have been used for various reasons. Some common ones are R_n, G_n, B_n N.T.S.C. (National Television Systems Committee) receiver primary system, Y, I, Q N.T.S.C. transmission system, and Y, U, V. All of these are simply linear transformations of the C.I.E. R,G,B system. Another common system is IHS (for Intensity, Hue, Saturation) which is *not* a linear transformation of R,G,B. Details on these and other color coordinate systems can be found in various books on image processing (e.g., by Jain, Pratt, etc.).