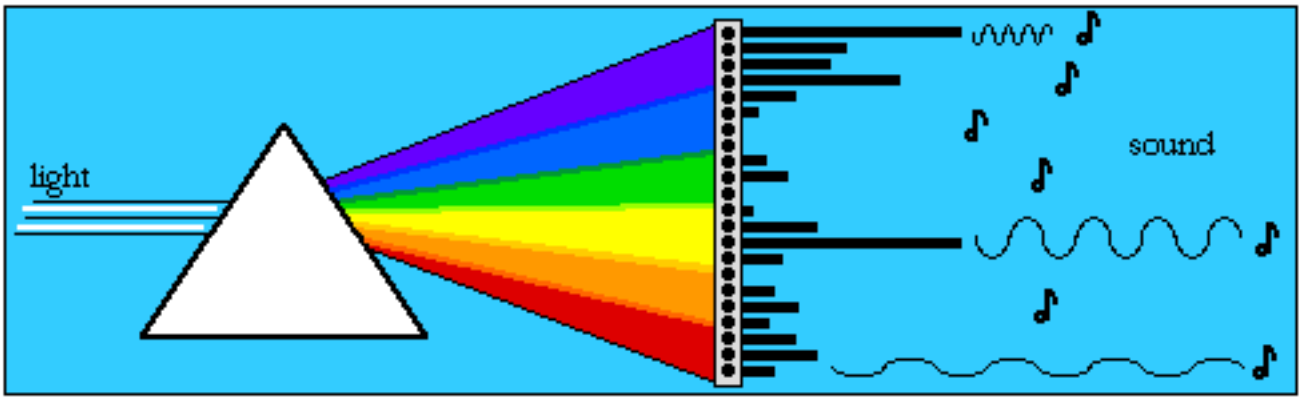
# Artificial Synesthesia via Sonification: A Wearable Augmented Sensory System

*Leonard N. Foner*
MIT Media Lab
20 Ames St, E15-305
Cambridge, MA 02139
foner@media.mit.edu
617/253-9601




## Abstract

A design for a wearable artificial sensory system is presented which uses data sonification to compensate for normal limitations in the human visual system. The system gives insight into the complete visible-light spectra from objects being seen by the user; long-term wear and consequent training might lead to identification of various visually-indistinguishable materials based on the sounds of their spectra. A detailed system design is presented, and many possible extensions to both the sonification and the sensor package are discussed.

## 1 Introduction

This paper describes the design and use of a wearable artificial sensory system that uses data sonification to compensate for normal limitations in the human visual system, and optionally to extend the sensory system into completely new senses. The system gives insight into the complete visible-light spectra from objects being seen by the user; long-term wear and consequent training might lead to identification of various visually-indistinguishable materials based on the sounds of their spectra. A detailed system design is presented, and many possible extensions to both the sonification and the sensor package are discussed.

The system is undergoing continuous improvement and redesign. Sonification is a critical part of its construction, and experiments are ongoing in determining the best way to represent the sensory information.

This paper is structured as follows:

- Section 1 (this section) describes the motivation for the project, including why wearability is an important goal. It also talks a bit about why sonification was chosen, and briefly discusses other possible senses that might be sonified in this way.
- Section 2 gives an overview of how the system works, including some of the tradeoffs necessary to make a wearable system. It then discusses the current state of sonification of the device, and the many ways to go from here.
- Section 3 briefly discusses some related work.
- Section 4 has some general conclusions.
- Finally, Appendix A presents a more detailed description of the imaging systems and their design; this medium-level description is included to give a flavor of the way the complete system works and some of the constraints that sonification must handle. This description is optional—those who care nothing about optics or about the actual hardware of the system may safely skip it.

## 1.1 A note about names

Before we begin, a note about names. This gadget is currently called the *Visor*. This term is just a temporary, stopgap name, to be changed later.[1]

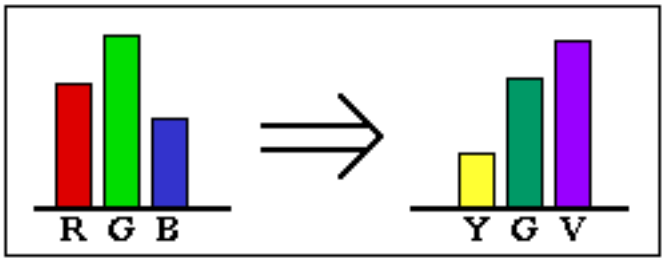
If you'd like to suggest a better name (please do!), please send mail to the author, who promises a nice dinner to whomever comes up with the name that finally sticks. If you're curious, you can also get a listing of those names already suggested.

## 1.2 Overcoming human visual trichromatism

The normal human visual system makes people into *trichromats:* any three wavelengths, if their amplitudes are properly chosen, can simulate any perceivable color, and hence can simulate any *other* three wavelengths. A cartoon of this concept is shown below.

---

1. If you're morbidly curious, I picked this particular stopgap because I was thinking of Geordi La Forge's VISOR from *Star Trek: The Next Generation*.




For a thorough discussion of how this can be experimentally verified using color-matching experiments, see [2]. This ambiguity in visual perception means that the visual system is easy to fool, which is fortunate, since it makes realistic color rendering possible using CRT's, photographs, or printed images.

The auditory system is not so easy to fool. Different chords sound different, and generally one chord cannot be made to sound like one composed of different fundamentals no matter how clever one gets with amplitudes. In short, the auditory system has *different* foibles; by combining audition with vision, we can perhaps get the best of both.

The device described in this paper makes it harder to fool the visual system, by mapping the colors of the environment into sound. The intent is something that can be worn comfortably for extended periods of time, which enables people to use it long enough to build up a natural mapping between sight and sound, e.g., "Oh yes, my lawn always sounds *that* way," or even, "Hey, my lawn *looks* okay, but it *sounds* funny today—maybe it's sick." For versions that have extended senses (see section 1.3 below), one might also be able to say, "Oh yeah—that car *looks* like metal, but it *sounds* like painted plastic!"

Such a gadget could hence be interesting for a number of reasons:.

- It could be a lot of fun to use.
- It has applications in, e.g., seeing through camouflage.
- It's possible that it would have scientific applications; for instance, perhaps certain plant diseases change the color balances of the leaves in a way that would change their "mapped sound" without changing their visual appearance very much.
- Any time we've come up with a different way of looking at the world, be they different timescales, media (e.g., light vs sound), or wavelengths, we've learned something. Perhaps there are some obvious but as-yet-unnoticed interesting phenomena in the world that we just don't know about because, to date, we've only seen through the eyes of trichromats.

While this gadget aims to encourage a synesthetic experience, it is not the same as natural synesthesia. Even in those cases of natural synesthesia in which sight and sound are coupled, the coupling between them is generally random—and furthermore, is fooled in exactly the same way that the visual system is fooled in general, e.g., it is visual-perception-specific, not wavelength-specific.

## 1.3 Extended senses

This device can also *extend* human sensory perception beyond synesthesia; there is no reason why it could not map infrared or ultraviolet wavelengths into sounds, too. Indeed, there are a large number of extensions to the basic idea, some of which are the subject of current experimentation:

- Extensions into the *near ultraviolet* and *near infrared*. This is extremely useful because a great wealth of chemical and compositional data is available in those wavelengths [1]. This requires appropriate modifications to the optical paths,[1] but would be relatively straightforward. Note that extension *beyond* near-UV or near-IR is more problematic—deep IR requires thermoelectric cooling of the sensor, and far-UV would require an external shortwave illuminator, since very little far-UV makes it through the atmosphere (c.f. [1], p. 69).

- Addition of a *polarimeter*, which images light based on its polarization. The daylight sky is mostly polarized, and is used by several species for navigation because of this—perhaps using a sonified polarimeter would allow people to sense the world more like a bee [10]. Further, materials under mechanical stress often display polarization effects, hence such a system could allow early detection of incipient structural failures.

- An *RF field* sensor. Using so-called “zeroth-order” sonification [5], in which the actual field strength is used to modulate sound amplitude directly, one might hear a series of clicks or even a tone in its own right from a GSM digital cellular phone, which transmits packetized data. Using more sophisticated sonification, one might be able to, for example, spatialize the perceived field received from a phased-array receiver.

- A *magnetometer*. Imagine walking by a wall and hearing the AC hum from the powerlines within.

---

1. E.g., normal glass isn’t very good outside of the normal visible spectrum, and would have to be replaced by, say, quartz; further, while most CCD’s are good in the IR anyway, few are very good in the UV.

## 1.4 Why sonification?

The basic idea of the Visor is to increase the perceived resolution of the visual spectrum—in other words, to make *trichromatic* humans into *polychromats*. (Extended ideas, as mentioned in the previous section, include mapping nonhuman senses as well.) The approach taken here is to take a foveal-sized region (about 1 degree of arc) of visual scene and sonify it. This is hardly the only way to remap some part of the visual scene's spectrum to make it more explicit to the user, though it seems the best and most natural way examined so far. Before examining the rationale for this approach, it may be helpful to examine some other, rejected approaches for doing the same thing.

### *1.4.1 Non-sonification ways to remap the visual spectrum*

*1.4.1.1 Passive prismatic mapping.* Take a prism and put it in front of the user's eyes. This has the advantage of being the cheapest, most obvious, and most straightforward approach. However, it has the serious disadvantage that it sacrifices spatial resolution for spectral resolution: any polychromatic light sources in the image are smeared out along the prism's separation dimension. This makes this approach infeasible for everyday wear, even if we ignored the other psychophysical artifacts which accompany it. For example, everything will seem to be off a few degrees in some direction because of the prism itself; while people are known to be able to adjust to this in a few days [and to take a few days to un-adjust after removing the prism], someone who only wore this part-time (a virtual necessity) would be constantly adjusting in one direction or another. While use of a diffraction grating instead of a prism would help this effect, it would tend to lose even more spatial information from the scene.

As a final disadvantage, strong emission peaks will tend to obscure weaker ones despite the smearing, especially if the weaker ones are in the blues (try looking at sunlight through a prism for an example).

*1.4.1.2 Chirped optical bandpass filter.* Imagine that the user looked through something like a double-pass monochromator, which acts as a very narrow bandpass filter. Imagine that we could continuously chirp the filter wavelength over a span of a few seconds, so that the user first saw only those elements of the scene that radiated at 700 nm, then 600, then 500, etc, slowly sliding down from red to blue and then back again.[1] Unfortunately, this approach sacrifices temporal resolution for spectral

1. As near as I can tell, this is effectively—to his nervous system, anyway—what Geordi La Forge's VISOR does in *Star Trek: The Next Generation*, with the proviso that it appears to do so over a much wider EM bandwidth and probably into low-energy subspace bands as well—for examples, see in particular "Heart of Glory" and, to a lesser degree, "The Mind's Eye." But as stated in section 1.1, I'm unhappy with this name, and would like another.

resolution: if the image is not static, it is likely that the user will miss some part of it while the monochromator is tuned to some other wavelength. Furthermore, it destroys the mixing effect of normal color vision, hence turning *all* spectra into peculiar artifacts. This makes it hard to know what things "really" looked like in the first place.

Since just about any scheme for increasing visual spectral resolution tends to run afoul of either reduced spatial or temporal resolution, methods employing synesthesia seem to be more promising. The major disadvantage of synesthetic approaches is that they tie up a different sensory system; however, the auditory system is quite good at discrimination of multiple input sources, and also at attending to only one of several if the task demands it. Hence, applying extra input to the auditory system, as long as it is not so loud as to drown out the outside world, seems both natural (the user is accustomed to multiple simultaneous audio inputs, and generally has little trouble discriminating them) and safe (he or she can fail to attend to the "extra" input if something more pressing demands attention).

The dimensions of the area imaged are important; we want an area that is likely to contain only one object or color in it. Since the fovea is really the only region of the eye where color vision is very important anyway, this seems a reasonable design criteria. A region that is extremely small (for example, the spot produced by a good laser, which is about a milliradian in divergence—about 1/20 degree or 3 arcmin) may present problems because there may not enough light coming back from such a small region to be imaged well with convenient sensors; further, such a small spot size means that the sensor is likely to cross many color boundaries, picking up a lot of fine detail in objects, which may make the acoustic signature unnecessarily confusing. A foveal-sized spot is about right in averaging out such effects.

Similarly, *where* the imaged area is in the user's visual field is important. It seems that the most natural location should be where the user's own fovea is currently imaging; after all, that's the part of the scene that the user has the best color information for at the moment, and is likely to be the part of the scene being attended to as well. Unfortunately, acquiring this information without being intrustive is the major source of some difficult design compromises; see section 2.2.1.

# 2 How it works

## 2.1 Visor implementation

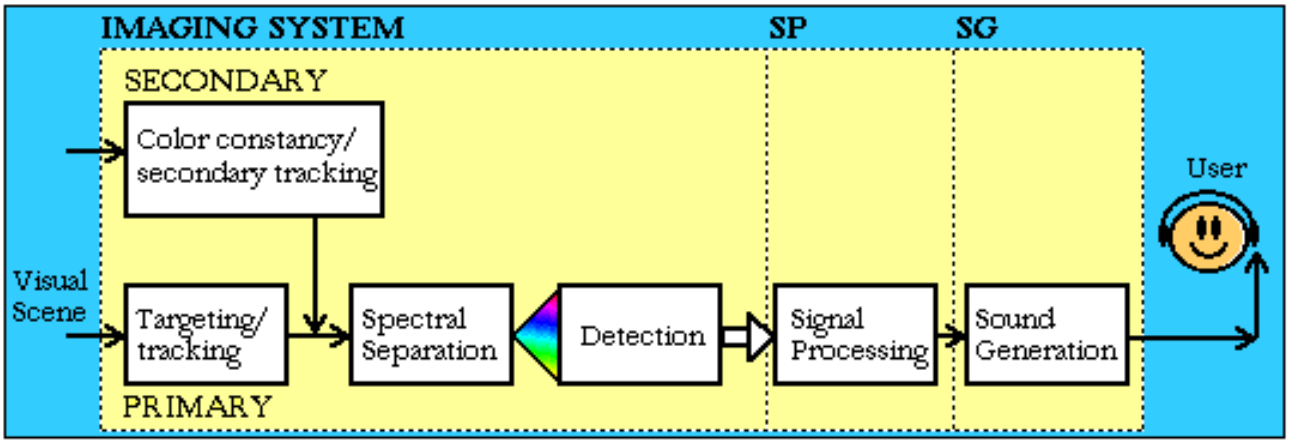
This section presents a general overview of the operation of the Visor, including some of the design decisions made in its optical system to make a wearable system, and discusses many approaches to sonification of the resulting data. For more detail on the actual hardware, consult Appendix A.

### 2.1.1 General overview

The conceptual design of the Visor is straightforward; it's only when it comes to the details that things get tricky. In order to understand why any of the designs presented were examined at all, it may be helpful to understand the rationale in section 1.4 for why this particular mapping of vision to sound was attempted in the first place, which provides some insight into how the current approach was arrived at, and also explains the particular size and placement of the area to be imaged by the Visor.

The basic idea is to take a small, essentially foveal-sized (e.g., about 1 degree of arc) chunk of the user's visual space and map the visual spectrum therein into an audio spectrum. In the best case, this chunk should actually *be* the region currently being imaged by the user's fovea, and should be acquired in an effortless, unobtrusive manner (leading to some interesting implementation tradeoffs; see section 2.2.1).

As shown in the figure below, this means that we must first *image* some part of the scene. We then take the imaged region, break it up into its spectral components, and *detect* what those components are. Finally, we take the detected visual spectra, do some simple signal processing, and *generate* an acoustic signature that corresponds to it. (A small complication, shown as an ancillary piece of the imaging system, includes the necessary extra components to aid in maintaining color constancy, if desired; see section A.4.)




You can find more details about possible designs for each of the subsystems in the following. Note that the most important source of *hardware* design diversity is in the primary imaging system (see section A.1), particularly in its tracking system, although you should read the section 2.2, below, for background first.

## 2.2 Visor optical systems

The Visor optical system is the most difficult part of the hardware design. The bulk of the important design tradeoffs are made here as well. (In contrast, the sonification of the resulting data is the most difficult part of the *software* design.)

### 2.2.1 Implementation tradeoffs

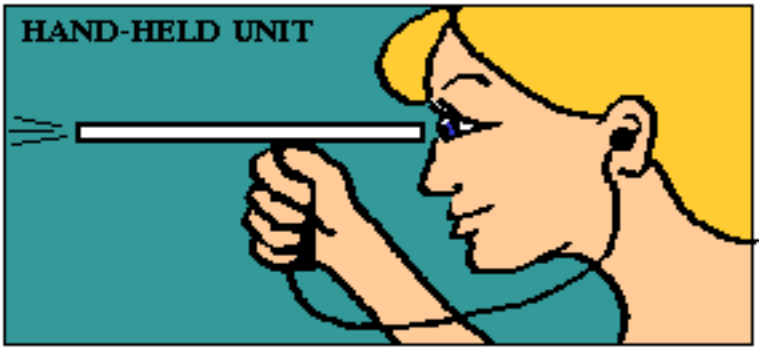
When designing the Visor, two factors tend to pull in opposite directions: *ease of use* and *ease of design*. It seems reasonable that the Visor will be most useful if it is worn most of the time—often enough and long enough that the wearer learns what the visual world sounds like, hence noticing auditory discrepancies (indicating unexpected optical spectra) that would not be otherwise noticed visually. This tends to push for a device is which comfortable and requires little to no effort to operate. Unfortunately, this is exactly the hardest one to design.

This sort of design tradeoff is hardly novel, of course; many devices are like this. In the case of the Visor, it means that most of the hardware design effort goes into the optical system, and has direct implications on wearability and usability. The sound generator (see section 2.3), being exclusively electronic and not electro-optic, is easier to change—but the question of *how* to sonify the data from the sensors is, of course, a big question and a serious design problem in its own right.

### 2.2.2 General approaches

The ease-of-use/ease-of-design tradeoff has several interesting points; we shall examine three major points along that continuum here. All of them concern the fundamental problem of *knowing what the user is looking at*.

*2.2.2.1 Completely manual tracking.* This type of system is the easiest to build. In it, the user looks through a straw or some sort of sighting device at whatever object is supposed to have its spectrum examined. For example, we might have the handheld unit shown below.

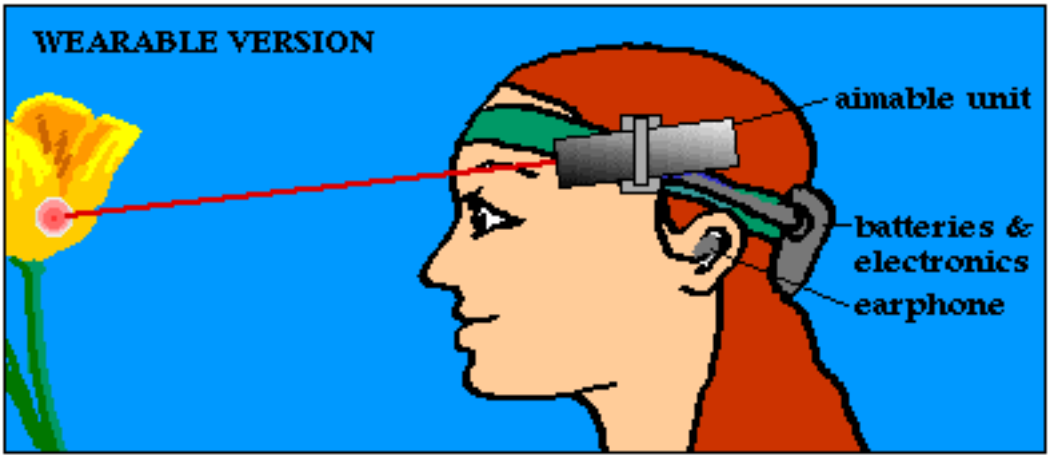



The imager (see details in the description of the primary imaging system in section A.1) can be simply positioned in a fixed relationship inside the collimating tube shown, making the design easy. Unfortunately, this requires one-eyed use (most people close one eye when sighting through something), and also ties up a hand. This is the least desirable configuration from the viewpoint of ease-of-

use.

*2.2.2.2 Pointer-based tracking* In this configuration, the unit is attached to the user's head, and points approximately along the line of sight. An immediate problem with this approach concerns the difference between the orientation of the head and the orientation of the eyes. Given a fixed head position, the user might still be looking anywhere in a very wide field (almost 180 degrees of arc horizontally and at least 100 degrees of arc vertically) simply by moving his eyes; since the imager is only examining a region about a degree wide, we need some way of calibrating the user to this point.

To accomplish this, a laser pointer is built in, in fixed, rigid optical alignment with the imager. The point being imaged[1] is exactly where the laser spot lands. A cartoon representation of the resulting system is shown below.




This approach is more convenient than a completely manual tracking system, since it frees both hands and enables two-eyed viewing. It still requires the user to carefully position a laser spot on the target, hence essentially fixing the orientation of the eyes relative to the head when the user wishes to scan something. While somewhat inconvenient, this is still much easier than a manual tracker, and still relatively easy to build.

The actual optical assembly is attached via a stiff ball-and-socket joint to an adjustable velcro headband. The ball-and-socket joint allows the user to aim the optical system into approximately the center of his or her visual field, without forcing the headband to be adjusted to match (it is unlikely that every user would find that a fixed positioning of the optical system to the headband would be both comfortable, secure, and correct). The joint stiffness is specified such that the unit won't flop or creep, yet can be manually adjusted without excessive force.

1. E.g., The *hearspot*.

The detailed description of this system (Appendix A) describes how to build such a device without having the laser spot itself alter the spectrum being analyzed; there are several easy ways to do this (see section A.1.1 in particular). That description also addresses calibration of the system's optical alignment (see section A.1.3).

**This device is the one explored in the most detail, in the medium-level description of this system in Appendix A.** The actual unit as constructed consists of a small DSP board (pictured in section 2.3.1), mounted on one side of the user's head, and a similarly-small optical system, mounted on the other side, for balance. Batteries can go behind the head, or elsewhere on the body, depending on runtime requirements and hence the size of the power source.

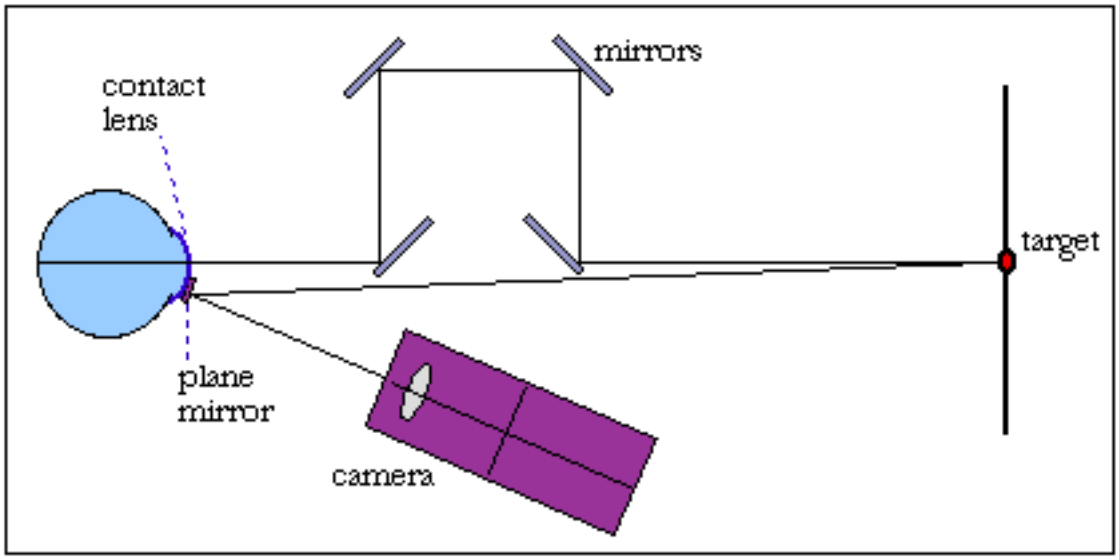
A more sophisticated design would employ flexible PC boards, embedded in the headband, to make the unit less bulky, and would use optical fibers to route the incoming and outgoing optical paths appropriately. An example is shown below.


The use of optical fiber would tend to automatically collimate the incoming light, though it is also important to use a suitable lens to gather enough light, and to match it to the narrow diameter (typically 100 microns or less) of the fiber. Similarly, a fiber optimized for visible-light operation is necessary; most are optimized for the near-infrared.

One interesting advantage of a fiber-optic system concerns the spectral separation step. There exist totally solid-state materials which, when a voltage is placed across them, act as tunable optical bandpass filters. Varying the voltage varies the filter's center point. Such materials are infeasible for optical systems larger than fibers, but for fiber-optic systems such a system saves both the prism and the CCD in one fell swoop—instead, we simply rapidly chirp the filter and observe the output at a single phototransistor. This is essentially a fiber-sized (instead of macroscopic-sized) chirped monochromator, as described in section 1.4.1.2.

*2.2.2.3 Eye-tracking.* The most sophisticated optical tracking system would employ eye-tracking to directly ascertain the point being observed by the user. No laser pointer system would be required.

Unfortunately, eye-tracking with any accuracy is expensive, bulky, difficult, and often fairly intrustive (see [11] for details). One approach, approximately in the middle of the range in terms of accuracy, bulk, and intrusiveness, involves the typical image-stabilization apparatus shown below.




The original purpose of such an apparatus is to stabilize an image on the retina no matter how the eye moves; such a stabilized image almost immediately fades out. By turning the former projector into the imaging system instead, we can always see what the user is seeing, no matter how his eye moves—with some important provisos.

First, this system depends on a fixed range to the target. Those four mirrors at the top of the diagram are designed to exactly double the optical path compared to the path on the right; this compensates for the way mirrors work and hence compensates for eye motion. (For more details, see [2].) While we could attempt the use of clever holographic diffraction gratings to change the relationship between the angle of incidence and the angle of reflection, such gratings would be wavelength-dependent.

Second, this system requires that the user wear a contact lens with a *plane mirror* in it. Plane mirrors on spherical surfaces have corners. Such lenses are hideously uncomfortable.[1]

While there are other potential ways to do eye-tracking, those which are less intrusive in this way

1. ...though not as uncomfortable as *some* lenses worn in different eye-tracking setups, which can have posts projecting from them which interfere with closing the eyelid!

tend to require a large amount of optical and computational power simply to figure out where the user is looking. Such a system would have to be wedded to a second system which sighted back along that vector to actually image the target. The result is virtually guaranteed to be expensive, heavy, bulky, and fragile.

For these reasons, until eye-tracking systems improve substantially, a Visor system employing eye-tracking was dismissed as a viable option.

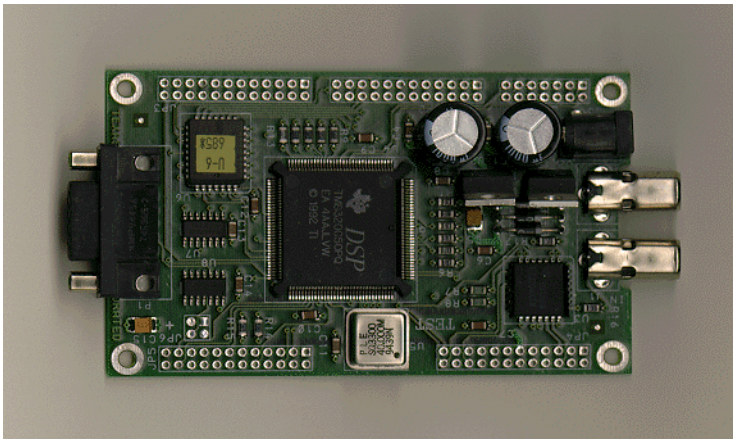
## 2.3 Sonification

This is where things all come together. This section briefly describes how the sensor data gets to the DSP, which is the heart of the sonification process. It then discusses the sonification of the resulting data.

### *2.3.1 Getting the sensory data to the DSP*

The chromatic detection system (section A.2) furnishes us with a byte-stream, consisting of 256 8-bit samples of the incoming spectrum (one CCD scan line) at regular intervals (on the order of at least 5-10 Hz to reduce system latency to changes in targets).

This byte stream is processed by a simple DSP chip, currently a TI 320C50, 28 MIPS, fixed-point unit, and is shown below integrated with the A/D and D/A converters, some memory, etc. The picture is about




75% of actual size; the actual 6 by 10 cm unit is one-half of what gets head-mounted (the other half, of course, being the optical system, on the other side of the head for balance).

The required computations are almost all table-lookups, multiplies, and adds, and there are not very many of them per second (a somewhat more detailed analysis is below), putting this well within the range of medium-end DSP's. The sinewave signals generated can be produced via a lookup table, and 16-bit nu-

meric precision appears to be adequate dynamic range, making floating-point unnecessary. (Note that more complicated sonification could well drive up the computation requirements of the DSP.)

### 2.3.2 Sonification of the spectral data

Each scan line from the CCD must be converted into a chord. In outline, we divide up the audio spectrum into 256 buckets, adjust the amplitude of each audio bucket to correspond to the appropriate visual bucket's current amplitude, and sum the resulting 256 different generated frequencies. (Other approaches are discussed below.)

In detail, there are several additional steps which raise the quality of the output. First, we need to rescale various quantities to match the detector used to human sensory characteristics. These are most conveniently done by running the samples through a number of lookup tables:

- Optical response of the eye at each wavelength
- Optical response of the CCD at each wavelength
- Log-scaling of overall brightness and normalization of total brightness to some standard (requires a discrete sum across all buckets)
- Audio response of the ear at each wavelength
- Audio response of the headphones at each wavelength (though this can be flat enough to ignore for reasonable headphones)
- Log-scaling of audio output
- Overall scaling of visible spectrum limits to audio spectrum limits. In other words, deciding how wide an input visible spectrum should be mapped to how wide an output audio spectrum—this determines the width of the *wavelength buckets* and is also influenced by the design of the chromatic detection system; see section A.2.

Several of these steps are combinable into a smaller number of precomputed lookup tables, if necessary, to achieve the desired DSP performance; in practice, this has not yet been necessary. (Presumably, they would be generated from a higher-level description during DSP programming, and hence could be arbitrarily complicated.)

In sonifying the sensed spectrum, the current system uses pure sinewaves, with simultaneous onset, whose amplitude is a direct map (scaled as above) to the amplitudes of visual wavelengths. This has a number of known disadvantages [4], but is trivial to program. Obvious further extensions include:

- *Timbre control*. Establishing the pitch of a pure sine wave is usually more difficult for people to do than for sounds with other overtones [4]; furthermore, those other overtones could pack a lot more information into the resulting audio output than a simple sum of sines. (For example, we could take the top few peaks in the spectrum and use those to control the timber in some clever way—although we must be careful to take *more than three* such peaks or we have laboriously recreated normal trichromatic vision!) Simply replacing sinewaves with, e.g., a waveshape like that of a reed instrument has relatively modest computational requirements—if minus any transients due to attack/decay of a real instrument, and minus any characteristic small amplitude variations, etc [4]—since the actual waveshape can be precomputed and stored as a table (this is currently how the sinewave signal is generated quickly with a fixed-point DSP). On the other hand, dynamic modification to the timbre based on the data is significantly more challenging, though with correspondingly large payoff in utility. One might also be able to employ speech-like sounds, such as vowel formants.

- *Staggered attack*. People often find it much easier to identify chords if not every note in the chord starts at the same time, e.g., if the chord is arpeggiated. Thus, it may be advantageous to have a frequency-dependent delay in the onset of each frequency before summing; this can be accomplished by simulating a simple delay line in software, with a different delay at each frequency. Such computation is fairly simple even for a low-end DSP, since it primarily affects a table offset in the sinewave generation computation.

- *Spatialization*. This extension requires a much more capable DSP, and also requires (of course) a two-channel audio system. This has implications beyond increasing the dimensionality of the data presented [9]. For instance, many people, especially those without perfect pitch, find it difficult to accurately sense (much less accurately recall) absolute pitches, yet the position of peaks in the visual (hence sonic) spectrum are excellent clues to identifying materials. On the other hand, people often remember *where in space* a sound originated. Thus, spatializing the sensor data may allow the user to more accurately recall any given spectrum based on its apparently physical map, rather than the tones used to represent it.

- *Melody*. As in the discussion of spatialization above, tunes are often much easier to recall than particular pitch spectra. Indeed, melody was used in a laboratory measurement system (see the discussion at [6], p.46, which referred to [7]). Of course, it should be observed that, while melodies can be easy to remember, they take *time*. This leads to one of the same problems demonstrated by the chirped-monochromator solution in section 1.4.1.2, namely that one must fixate on the object of interest long enough

to hear the entire melody.

- *Plucked notes*. Unless the acoustic signature for most surfaces is both pleasing and unobtrusive, it is going to be irritating to wear it continuously. Imagine the case, for example, of someone who spends a while reading a book or looking at a text editor. Since the surface being observed isn't changing much, the sound shouldn't either, and one would imagine that the user would stop attending to it. One might aid in the user's backgrounding of the sound by actively decreasing its amplitude in such a case; the limiting case of this is to represent the interface as one of plucked notes instead of continuous tones: any change in optical spectrum larger than a threshold triggers a pluck. Hence, whereas scanning across a rainbow in the original system leads to a continuous scale, scanning a rainbow in this system might lead to a finger-walk on a piano keyboard.

# 3 Related work

The Visor stands at the intersection of several usually disparate fields: laboratory instrumentation, sensory augmentation, ubiquitous computing, and sonification. It seems, in general, that pairs of these points are usually addressed. For example, [7] hints at, and [6] discusses in more detail, a system for augmenting the (missing) visual sense in a blind subject so he could use laboratory instrumentation. Indeed, [6] contains an impressive list of sonification projects, which I have no wish to reproduce here. In addition, [5] as a whole is the obvious jumping-off point for discussions of sonification, both in the abstract and in the particular projects presented. And even though the Visor is not a warning device per se, the results of [8] could be quite useful in making it less objectionable and increasing the likelihood that its output is understood.

To my knowledge, however, no single device has simultaneously attempted to provide both ubiquitous (hence wearable, implantable, or otherwise unobtrusive) sensory augmentation (whether to compensate for damaged "normal" senses or to extend existing ones) via sonification, with a few interesting semi-exceptions:

- *Hearing aids*. Are these sonification? Not really. They do not really *transform* the data so much as amplify it, with possible frequency equalization adjustments.

- *Seeing-eye dogs*. If the dog barks to tell its owner something he or she cannot see, this can be considered sonification after a fashion. However, unless one counts the dog's training, this is hardly a *designed* system.

It is interesting that both of the systems above are aimed at users who are disabled (relative to the normal human population) in some way, rather than at extending the sensory range of nondisabled users. Systems aimed at nondisabled users are generally either visual (e.g., binoculars, equipment readouts), essentially simple amplifiers (e.g., stethoscopes), or transducers (e.g., Geiger counters). Those systems which attempt to actually transform data into sound via computation (e.g., [7] or [3]) are not generally designed with portable, ubiquitous use in mind. And most practicers in the field of wearable computing have generally dismissed the audio channel as fit for only voice communication, or simple alerting (e.g., beeps).

# 4 Conclusions

A system has been presented which is designed to be constantly and ubiquitously worn by its user, and can extend the sensory range of even normally sighted individuals by sonifying the complete visual spectra of objects in the visual field. Careful attention to system design yields a device small enough to be convenient, and hence used most of the time. This presents a rich set of sensory data, making creative sonification a challenging and rewarding task.

Many potential avenues for improvement remain to be explored, primary in two areas:

- Extending the sensory capabilities of the instrument, either into nonvisual electromagnetic bandwidths, or by adding different kinds of sensors entirely.
- Improving the sonification of the resulting sensor data.

Work is currently continuing on both these fronts. It is my hope that this could spawn a number of efforts for the routine sensory augmentation of ordinary humans with extraordinary senses, and that it also spurs further research into sonification and auditory display in general.

# A Appendix A: Visor sensory system details

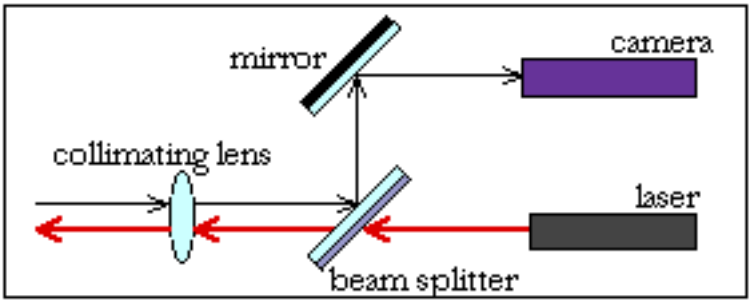
This appendix details how the Visor collects its information from the outside world, and some other details of its construction. This material may be interesting in its own right to the curious reader; it also shows some of the tradeoffs necessary on the way to sonification of the data, and suggests other area where sonification may pay off.

## A.1 Visor primary imaging system

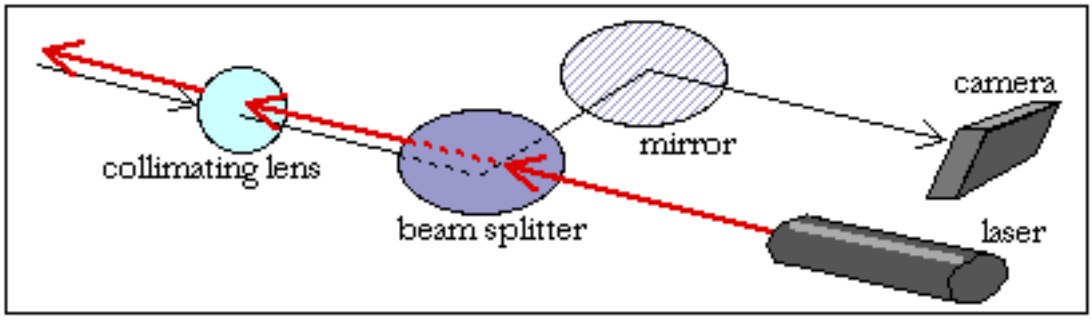
The primary imaging system is designed for the head-mounted, pointer-based tracking system, as described in section 2.2.2.2. This is a compromise position between ease of use and ease of design; see section 2.2.2 for details.

The simplest design would place a camera (with appropriate collimating optics, so that the camera images only a 1-degree patch) side-by-side with a laser. However, since the resultant optical axes would be offset from each other, this would lead to an aiming error at short ranges. (This is exactly the same effect that makes simple Instamatics hard to use at short range; the viewfinder does not look out through the same lens that the film sees.)

Such an aiming error would be fairly objectionable at short ranges; for this reason, the actual optical system design works somewhat like a single-lens-reflex [SLR] camera, in that it aligns both the laser and the camera along a single optical axis. Such a colinear arrangement reduces the aiming error to whatever mechanical misalignments remain when the unit is assembled, and it is straightforward to mount the two components such that this misalignment is substantially less than a degree (and pot them both in optical epoxy to maintain this). An SLR uses a moving prism to time-multiplex the optical path between the viewfinder and the film plane; the design here uses a beamsplitter to multiplex the two paths, sacrificing a small amount of beam power to do it. This is illustrated (in the plane and in perspective) by the diagrams below.




**Note well** that, in the diagram above and in most of the diagrams here, the box labelled *camera* is actually a combination of a prism or grating of some sort and a camera.




### *A.1.1 Keeping the laser from interfering with the spectrum measurement*

Since we are shining a laser exactly at the point being imaged, a natural question to ask is, "How do we keep from being affected by the laser?" There are a multitude of simple remedies.

A.1.1.1 *Duty-cycle control.* The most straightforward approach is to modulate the laser such that it is out of phase with respect to the imager. Thus, whenever the laser is on, the imager is disabled, and whenever the laser is off, the imager is enabled. One easy way to do this is to throw away every other line coming in from the imager (equivalent to throwing away every other frame if this was a 2D CCD instead of a 1D CCD), and fire the laser only when we are throwing away CCD lines. (We must throw away lines coming in from the CCD to bleed charge out of the individual pixels; simply stopping the sampling clock would allow charge from incoming photons to build up, which would not work very well.)

Note that the constraints of the imager and that of the laser are working for us in this case. If the target is very bright (e.g., the user is outside on a sunny day), the imager needs very little time to be properly exposed. On the other hand, the laser needs to be on for as long as possible, so that it can be seen at all. (We assume here that the laser, when on, is on at full power—this makes the modulation circuit trivial.) In such a situation, we may want to skip three or four lines of input from the imager, firing the laser all this time, such that the laser's duty cycle approaches 80% or more. On the other hand, in low-light situations, we may want to only throw away every fourth or fifth CCD scan line, and run the laser at 20% or lower duty cycle. (The primary constraint on the number of scan lines discarded is the latency of the rest of the system, since, if we are throwing away many scan lines, it could be many hundreds of milliseconds in between each sampled line, leading to possibly-perceptable lags. Similarly, if the laser duty cycle is too low, it will appear to flicker.)

This scenario also makes the laser less obnoxious when in low light: indeed, the user can be given a knob that controls the ratio of the beam's brightness to the ambient light, and can set this relative

brightness down far enough that the spot is just visible enough to see, and no brighter. (This has two ancillary benefits as well: the effects of accidentally nailing someone else in the eye are reduced [although a typical 5mW Class III-A laser is not particularly dangerous anyway, simply annoying], and the power required to run the laser is reduced [although this is not a major drain: a 5mW 670nm diode laser typically consumes only around 15-20mW in continuous duty].

A.1.1.2 *Polarization control.* Another way to handle the influence of the laser sight is to use polarization. If we were to put a horizontal polarizer at the laser emitter, and a vertical polarizer at the camera aperture (actually, just in front of the prism), then the camera cannot see the laser unless the beam's polarization is rotated by the target. (A polarizing beamsplitter is another potential solution.)

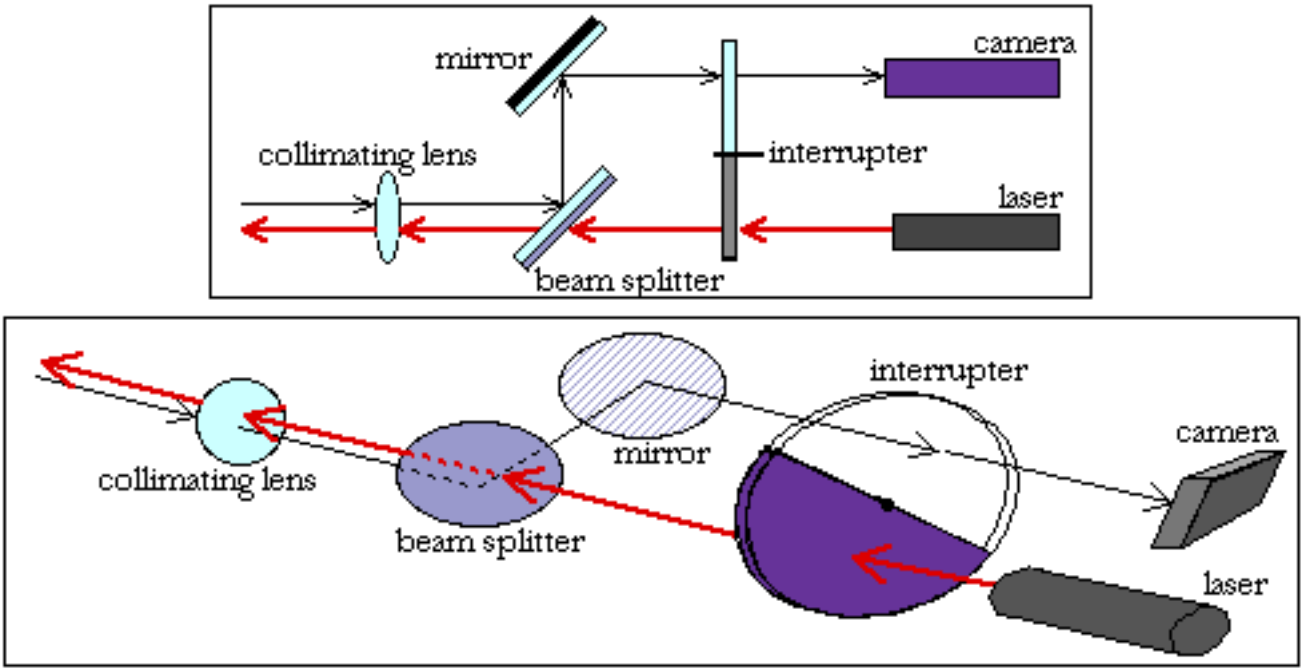
This approach is simple, requiring no modulation control, but depends on the target to be well-behaved with respect to polarization rotation. Further, it also depends on the target having the same emitted spectrum regardless of its polarization; this assumption may or may not be correct in the face of, e.g., roads and bodies of water. Also, it wastes half of the laser power (since diode lasers do not normally produce polarized light) and half of the target's light (this is in addition to the approximately 50% wasted by a non-holographic beamsplitter for both paths—a total 75% reduction).

Using polarization does have the great advantage that it is totally passive and trivial to implement. If the laser is nonetheless too bright and obnoxious in low light, we can still implement modulation control (though in this case, there would be no reason not to chop the beam in the tens of kilohertz, meaning that it would not appear to flicker no matter how low it was turned down—this might be an important advantage).

A.1.1.3 *Camera bloom.* If we use duty-cycle control instead of polarization to avoid seeing the laser sight with the camera, there is an unfortunate effect that may make life difficult. This is *camera bloom*, in which certain pixels of the camera are so overexposed that all the charge cannot be transferred off in one cycle. While generally a much more serious problem with, e.g., vidicon tubes than CCD's, this may still be an issue.

If this is a serious problem, we can employ a *mechanical interrupter* which simply prevents the imager from ever seeing the laser's spot in the first place, as shown in the illustrations below. Here, we simply spin a disk in the beam paths. Part of the disk is transparent, and part is opaque:

Obviously, this approach constrains the degree to which modulation control can be used, since we can never go beyond 50% ON for either the laser or the camera. (If the transparent/opaque ratio is not




50%, then this argument changes in the obvious way.) We would also have to phase-lock the interrupter to the imager; this might require a phase-locked-loop and a position encoder (trivial to do, given this spinning light-and-dark disk), or possibly some sort of brushless DC motor driver.

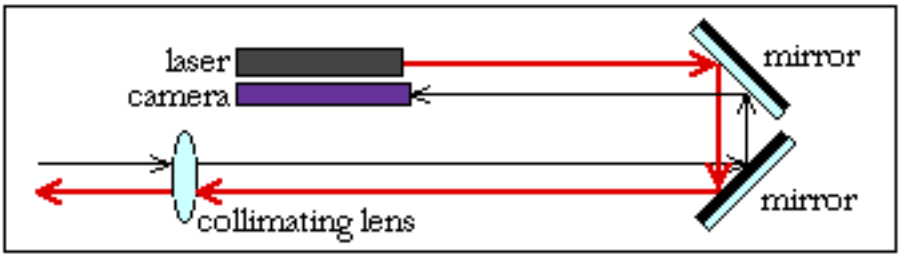
Another approach, which might actually be simpler considering the necessity of phase-locking the above system and its duty-cycle inflexibility, would be to replace the mechanical interrupter above with an LCD shutter (as used in, e.g., some 3D glasses) in front of the camera only (since the laser is easy to modulate electronically), or to use some sort of resonating mechanical mirror or acousto-optic modulator.

A.1.1.4 *Simple subtraction.* Another approach is simply to calibrate the unit to subtract a certain amount of light, in the signal processing stage, at the wavelength bucket which contains the laser light. This amount of light is easy to predict based on the current laser output energy. Unfortunately, it may be the case that in many regimes, that particular pixel is completely saturated by the laser, such that subtracting the laser's contribution leaves us with zero even though the target had energy at that wavelength. Active control over beam output energy might help this (since darker scenes need little laser power), but a bright outdoor scene which happens to be low in reds is still a problem.

### *A.1.2 Making the system more compact*

The beam path of the pointer-based tracking system described in section 2.2.2.2 has a lot of components in it: a collimating lens, some distance for the collimating to happen, a beamsplitter, perhaps an interrupter, and either a prism and a CCD imager, or a laser diode with its own collimation lens.

The resulting optical path length can make the system as a whole somewhat unwieldy and uncomfortable to wear; it would be no fun to be wearing something long enough to whack on random obstacles. At the very least (and without going to the high-tech version of the pointer-based system described in section 2.2.2.2), we can halve the beam path by some careful component rearrangement. For example, consider the setup below:




This diagram is simplified; for example, it doesn't include the beamsplitter, associated mirror, and possible interrupter (all of which would be in the upper optical path), and the pair of mirrors at the right would more likely be a single right-angle prism. This setup could put the bulk of the components parallel to the collimation path, hence shortening the whole thing.

### *A.1.3 Calibration procedures*

It is useful, when assembling the unit and perhaps during adjustment later, to be able to be assured that the imager is indeed optically aligned with the laser, e.g., that they are both targeting the same point.

One way to do this would be to create a test target, consisting of a spot of some color surrounded by a field of some other color. Land the laser on the spot, and ensure that perturbing the laser from the spot leads to a change in sound. This might even be usable in the field, since it may not be too difficult to find an appropriately-sized spot, or a set of edges, that can serve the purpose.

Another way to accomplish this would be to remodulate the laser emitter; this would only be usable if we are using duty-cycle control (section A.1.1.1) without either polarization control (section A.1.1.2) or an interrupter (section A.1.1.3). In this mode, we would modulate the laser slowly (e.g., 1-5 Hz) while ob-

serving *every* scan of the CCD. If the CCD detects synchronous, in-phase changes in total brightness, it is looking at the laser.

Further, if the correct wavelength bucket is being modulated, then the prism is correctly aligned with the CCD. This is information that is not available in the static, test-target approach; to do this well with that approach, we would require a target with known spectral peak(s) upon which to calibrate. Such a target would be unlikely to be made of ink (since ink tends to have broad spectra and also fades over time); instead, it would likely have to be made of a gas-discharge lamp viewed through a small hole, or be a laser itself.

### *A.1.4 Optical efficiency*

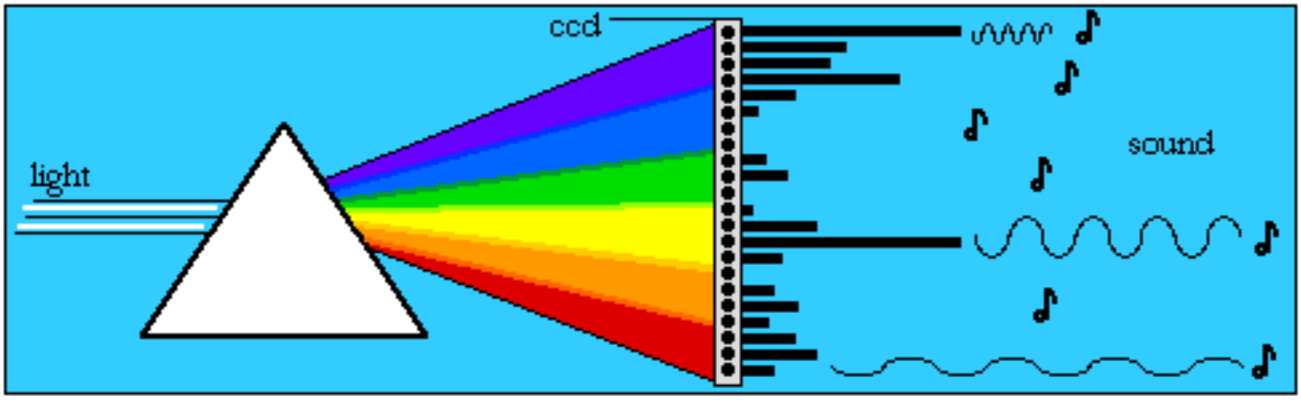
The designs illustrated above lose some light due to a couple of small effects. While a reasonable choice of lens and CCD can compensate for a great deal of light loss, if we were truly concerned about it, we could either:

- Use a *holographic optical element* for the beamsplitter, which could be manufactured to transmit (or reflect) only light of the wavelength in use by the laser, and reflect (or transmit) all others. This would sacrifice less beam power, while taking only a *very* thin slice out of the spectrum being imaged: essentially, we could double the optical efficiency of both the laser and the imager over the nonpolarized case, and quadruple it in the polarized case (half of the loss occurs at the non-HOE beamsplitter, and each resulting path also has a 50% loss due to inserted polarizers).
- Rearrange the beam path so the camera sights only through the beamsplitter, while it is the laser that must also reflect off the pure mirror. (A decent front-surface mirror should have a reflectance of at least 99%, however, so the extra reflection hardly hurts us; on the other hand, using a dielectric mirror calibrated for the laser's wavelength can give us truly excellent reflection if it is the laser that hits the mirror instead of the light from the target.) Note again that even the naive design only loses on the order of 50% to 75% of the light; for reasonable lens and CCD choices, this may not be a problem. Note also the discussion of sensitivity (section A.2.2) in the discussion of the chromatic detection system (section A.2).

## A.2 Chromatic detection

The chromatic detection part of the Visor is quite straightforward. The general idea, as shown by the icon at the top of the first page, and in slightly more detail below, is to take a beam of light from the target,

run it through a prism to spread its spectrum, and then run the resulting fan of light into a one-dimensional CCD imager.




### *A.2.1 The prism*

It is the diagram above, with the prism and the camera, which is shown as just the *camera* in most other diagrams, such as the ones describing the primary imaging system.

In order to get a wide dispersal of light (the CCD we are using has an active region 3mm long, and other CCD's can be longer) in the shortest possible distance (since that distance makes the total path length longer and therefore increases the size of the overall unit), a diffraction grating is a better choice than a simple glass prism. There are a wide variety of such gratings available; selecting the appropriate grating for the geometry of the whole system is easy.

### *A.2.2 Sensitivity*

Since the actual light from common lenses will be a round beam of some nonzero cross section, we need to squeeze down the beam cross-section before running it through the prism or the spectrum will not be as sharp as it could be. One way to do this would be to use a simple slit in front of the prism; however, this approach sacrifices whatever light spills over the slit. A better approach is to use a cylindrical lens instead, which makes use of all of the available light.

The actual imager is typically a 256-element linear CCD, with 8 bits per pixel. Light in a particular wavelength bucket will *always* on the same pixel of the CCD, assuming that the geometry of the system is sufficiently rigid. (Some metal and/or optical epoxy can go a long way here.) Since the light's spectrum is being subdivided 256 times, the brightness of any one pixel (assuming white light input) is 1/256 as bright as the target, *not* including losses from the rest of the primary imaging system (which can be as

much as 75% in some designs; see section A.1.1.2). This means that we may take a hit in brightness of up to three orders of magnitude between the target and any particular pixel.

On the other hand, unlike a typical 2D camera, the light entering the beam path is integrated over the entire 1-degree target area, instead of coming from smaller, pixel-sized subareas. Also, a 2D CCD would be spreading the light from the image out over, say, 400 x 600 = 240,000 pixels (or even three times that many, for a 3-CCD color camera); here, we can take all of the light, squeeze it along one dimension with a cylindrical lens, and land it on only 256 pixels. Finally, there is no particular need to clock the CCD as rapidly as a normal video camera (e.g., 30 or 60 Hz). Instead, the major limiting factor on CCD clock rate is overall system latency and flicker of the laser sight (if using duty-cycle modulation; see section A.1.1.1).

Thus, the required sensitivity is easily achievable even with low-light 2D CCD imagers, and 1D imagers can do much better due to the larger pixel size available on the die and the small number of illuminated pixels.

### *A.2.3 Calibration*

A gas-discharge lamp with known peaks is the most obvious way to calibrate the system, albeit somewhat inconvenient in the field; such a procedure would also establish linearity (or the lack of it). A single laser can also be used to at least register a point in the spectrum on a known, particular pixel of the CCD—the section of the primary imaging system which discusses calibrating the laser sight to the imager (section A.1.3) mentions a way in which this can be done without any external light sources.

## A.3 Glue logic

There are only two more steps required after the CCD and before acoustic generation:

- Correct interlocking between CCD scans and the laser modulation, if using duty-cycle control (section A.1.1.1), and
- An 8-bit A/D converter. (A very modest one suffices; even at a sampling rate of 30 Hz, we only need to convert 30 x 256 = 7680 pixels/sec, which gives us 130 microseconds/pixel.) In the DSP system currently being used, the A/D and D/A converters are integrated on-board.

## A.4 Color constancy

The Visor's primary imaging system, as specified, will produce different results and hence different sounds, if presented with two identical targets which are lit differently. The human visual system, under most conditions, does not behave this way; it exhibits color constancy [2].

It is not clear whether this behavior of the Visor is a bug or a feature; an argument can be made either way, depending on the application. A reasonable argument might be made, however, that the user should have a choice of operating in "absolute mode" or "color-constancy mode", depending on his goals. The discussion below examines how this could be achieved, and describes the *secondary imaging system*.

The way the human visual system achieves color constancy is by comparing the *ratio* of color intensities from the fovea to those of a surrounding, much larger area (see [2] for a thorough treatment of this effect and its experimental verification). This so-called *center/surround* system can be approximately simulated by the Visor as follows.

Consider putting a spinning *holographic optical element* in the collimating beampath, probably at the very start (e.g., the input to the whole device). This element would look something like the interrupter disk described as a potential solution to the problem of camera bloom (see section A.1.1.3) in the discussion of the primary imaging system. However, instead of one half being transparent and the other opaque, the HOE would be designed such that one half is a lens of the right power to image a 1-degree patch (e.g., the normal collimator), while the other half images a much larger area (e.g., 30 degrees or more).

We phase-lock the imaging system to this spinning disk, so that the DSP knows which scan lines are the "center" and which are the "surround" data. It can then take appropriate ratios to cancel the contribution of lighting spectra to the target spectra.

Note that this disk doesn't have to spin very fast, since lighting changes are usually slow. (And if the system is retargeted, *both* the center and the surround data will change simultaneously, so this is not a problem either.) Also, since illumination data changes slowly, there is no reason to make the HOE 50% center and 50% surround; instead, it makes more sense to make the HOE mostly center and only a thin pie-slice of surround—this could make, say, every tenth scan line a "surround" scan line instead of every other scan line as in the 50% system.

## A.5 The final step in the signal path

The final step in the signal path is the audio amplifier. This is trivial; it consists of an audio power op-amp with a volume control, connected to either an earphone or a set of Walkman headphones.

## A.6 Power supply

The power supply required for the visor is quite modest. The major electronic components include:

- The laser emitter (about 15-20mW for a 5mW output 670nm laser diode, which is as bright as allowed under Class III-A safety regulations).
- The CCD imager (tens of milliwatts).
- The DSP chip (hundreds of milliwatts depending on type; in the tens for some).
- The audio amplifier (at most 200mW, and usually much less, depending mostly on the efficiency of the earphones and the deafness of the user).

If using some sort of interrupter system, such as a spinning disk, an LCD shutter, or some sort of acous-to-optic modulator, the driver for that (tens of milliwatts?). This gives a total power budget of an amp or less at 5V. Rechargeable NiCad or NiMH batteries of reasonable size work fine for several hours at a stretch; also, they can be carried anywhere (e.g., in the pocket, not the head) if total mass is a problem. (The major problem here would be running a cord from a pocket to the head in a comfortable fashion.)

# References

[1] Anderson, Herbert, ed., *A Physicist's Desk Reference*, American Institute of Physics, 1989.

[2] Cornsweet, Tom, *Visual Perception*, Harcourt Brace Jovanovich, 1970.

[3] Fubini, E., De Bono, A., and Ruspa, G., "System for Monitoring and Indicating Acoustically the Operating Conditions of a Motor Vehicle," US Patent #4,785,280, US Patent and Trademark Office.

[4] Handel, Stephen, *Listening*, MIT Press, 1989.

[5] Kramer, Gregory, ed., *Auditory Display: Sonification, Audification, and Auditory Interfaces*, Addison-Wesley, 1994.

[6] Kramer, Gregory, "An Introduction to Auditory Display," *Auditory Display: Sonification, Audification, and Auditory Interfaces*, Gregory Kramer, ed., Addison-Wesley, 1994.

[7] Lunney, David, and Morrison, Robert, "High Technology Laboratory Aids for Visually Handicapped Chemistry Students," *Journal of Chemical Education*, volume 58, 1981, pp. 228-231.

[8] Patterson, R. D, "Guidelines for Auditory Warning Systems on Civil Aircraft," Civil Aviation Authority, London, 1982.

[9] Wenzel, Elizabeth, "Spatial Sounds and Sonification," *Auditory Display: Sonification, Audification, and Auditory Interfaces*, Gregory Kramer, ed., Addison-Wesley, 1994.

[10] Winston, Mark, *The Biology of the Honey Bee*, Harvard University Press, 1987.

[11] Young, Laurence, and Sheena, David, "Survey of Eye Movement Recording Methods," *Behavior Research Methods & Instrumentation 1975*, Volume 7, number 5, pp. 397-429.