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 an implemented, prototype 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 and results of user testing are 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 sonification—turning light into sound—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 and results of user testing are presented, and many possible extensions to both the sonification and the sensor package are discussed. A prototype of the system has already been implemented, and it 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.

1.1 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. An illustration of this concept is shown in Figure 1.

For a thorough discussion of how this can be experimentally verified using color-matching experiments, see [2]. This ambiguity in visual perception has important practical applications: for example, it makes realistic color rendering possible using CRT's, photographs, or printed images which employ a very limited number of ink or phosphor colors. However, this makes the unaided color visual system unsuitable for tasks in which accurate discrimination of the shape of the spectrum is important, such as determining whether a particular image is a printed or displayed representation of some natural object (employing 3 or 4 colors of phosphor or ink), or is the object itself (with a much more complex spectra due to larger numbers of variously-colored molecular species).

The auditory system operates according to different principles and has different sets of ambiguities [6]. The most important difference for the current discussion concerns the effects of the shape of the presented audio spectrum. Though the audio system can suffer from *masking* [19], these effects tend to be strongest when nearby frequencies are involved (e.g., 2900-3300Hz) or when onsets are correlated in time (e.g., within 200ms) (see [6], p. 70, for example). This means that we can use the differences between the auditory system and the visual system to allow a mapping from visual spectra into audio spectra to reveal phenomena that the visual system alone could not, in particular, to allow the audio system to detect spectra which are shaped very differently but correspond to visual spectra that are indistinguishable.

The device described in this paper—the *Visor*—accomplishes this task, 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 4), one might also be able to say, "Oh yeah—that car *looks* like metal, but it *sounds* like painted plastic."

The Visor's application areas are varied. One thing that should be kept in mind is that it is *not* intended as a replacement for high-resolution visible or infrared spectroscopy. In other words, applications in which it is essential that a high-resolution, well-calibrated readout of an object's visual reflectance spectrum be obtained are inappropriate for this technology. Instead, it is intended primarily to alert its user to unexpected deviations in visual characteristics of a material from what might have been anticipated, e.g., to make it possible for *serendipitous discovery* of unusual properties, by virtue of the extended wear possible with this device and the relative unobtrusiveness of the data presented; such uses could have potential as an entertainment medium.

Other, more utilitarian applications for the Visor do exist, however. For example:

- *Seeing through camouflage.* Because an object which is painted to match a jungle setting is not painted with the jungle itself, but is instead painted with pigments with different spectra, there can be audible differences in the resulting audio-mapped spectrum. See section 3.1 for an example.
- *Melanoma screening.* Certain skin tumors change their spectrum in ways that are hard for the unaided eye to detect as they become cancerous. Nonetheless, the spectral changes can be detected with spectrometers and can lead to simplified diagnosis of suspected lesions [8][13]. A system such as the Visor could make routine screening a regular part of a diagnostician's first glance at the patient.

The examples above are provided to give a flavor of the sorts of applications that this device may have in future work. The research presented here uses the Visor strictly in its serendipitous and entertainment-related applications, although work is currently ongoing [16] in building a simple spectroscopic system for routine use by cancer clinicians.

While the Visor aims to encourage a synesthetic experience, this is not the same as natural synesthesia. Actual synesthetes have an apparently neurologic coupling between various sensory modalities, which causes, for example, a particular color to have an associated sound or texture [3]. Such mappings can be very stable over time, such that, in a given individual, a particular color remains associated with a particular sound for many years [7]. The sort of synesthesia afforded by the Visor is of a different sort, however, because the actual *sensory input* of a synesthete remains that of a non-synesthete. In other words, even a synesthete will see matching colors of greens between a photograph of a leaf and the actual leaf, though the pigments used for the photograph have very different spectra from those in the leaf itself. A user of the Visor, however, can hear the difference, because the actual pigments have different spectra.

1.2 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*. The approach taken here is to take a foveal-sized region (about 1 degree of arc) of visual scene and sonify it.

Sonification was chosen primarily for the following reasons:

- *To avoid cluttering the user's visual field.* If the device required a visual readout, it would also require some sort of heads-up display if it was intended to be used in a hands-free manner, as would be appropriate for a device intended to be used many hours a day. However, this display would be both bulky

and distracting—many users do not wish to sacrifice a portion of the field of view in their normal activities.

- To allow ignoring the system's output when it is unwanted. Phenomena such as the *cocktail party effect* ([6], p. 189) make it clear that people can attend to one or another audio source with very little effort; the unattended source for the most part vanishes from perception. By using audio, at low amplitude, through an earphone which does not also block hearing in that ear, the Visor's output can be made unobtrusive to a user who is not actively attending to it.

1.3 Why a foveal-sized imaged region?

The choice of a foveal-sized region is in some sense arbitrary. One could argue that the eye naturally perceives most (but not all) of the colors in any image in this region [2], and hence that this is somehow a natural choice, but in fact the nature of the visual system is such that people are never aware, in normal circumstances, that their color vision is so severely restricted. Instead, the very small angle imaged by this system is intended to make the system less overwhelming and more comprehensible to its user, by restricting all the possible parts of the scene that could be imaged, and hence mapped into sound, to a small patch. This allows the user essentially to aim at a single, small region, and hear a sound associated with that region, without being distracted by all the other colors that appear elsewhere in his or her visual field.

The dimensions of the area imaged are important; we want an area that is likely to contain only one object or color in it. 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) presents problems because there is not enough light coming back from such a small region to be imaged well with convenient sensors (including the associated optics required to gather sufficient light). 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 spot of a degree or two has been experimentally determined to be acceptable in averaging out such effects.

2 Design and implementation of the system

2.1 General overview

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.

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. To do this, we first image some part of the scene. We then take the imaged region, break it up into its spectral components, and detect the components. Finally, we take the detected visual spectra, do some simple signal processing, and generate an acoustic signature that corresponds to it. The details of this process are described below.

2.2 Visor hardware implementation

The Visor is designed to be used continuously. Three possibilities for its physical construction are:

- A handheld unit; the user sights through it or aims it manually at targets.
- A headmounted unit that looks where the user's head is aimed.
- A headmounted unit that looks where the user's eyes are aimed.

The first option is simple to build, but requires continuous attention from the user, who must also sacrifice the use of a hand at all times. The third option is quite complex and also quite intrustive—eye tracking systems tend to be bulky, expensive, and some of the highest-accuracy devices even require the use of, e.g., contact lenses with plane mirrors embedded in them, which are very uncomfortable [20].

For these reasons, the second option—a headmounted unit that images a spot related to where the user's head is aimed—was chosen. A sketch of the general idea appears in Figure 2, and photographs of the unit in operation appear in Figure 3. The major hardware components, described in more detail below, are the spectrometer, the DSP, and the power source.

The current system uses a commercially-available, very compact fiber spectrometer, namely the Zeiss MMS-1, as its input device. This spectrometer measures only 7 by 6 by 4 cm and masses under 200g. It has a 24 cm fiber-optic bundle as its input; this bundle is then matched to an appropriate lens, to provide light-gathering power and restrict the light cone admitted by the system to the desired width of 1-2 degree. The resulting spectrometer and lens assembly is sensitive from 1150nm, in the near-infrared, up to 305nm, beyond the limit of human blue vision (about 400nm).

The more slowly the spectrometer is sampled, the longer each pixel has to integrate incoming photons, and the greater its resulting sensitivity. The Visor's DSP examines each complete set of samples (one 256-pixel scan line) and determines whether the spectrometer appears under- or overexposed; it then either increases or decreases, respectively, the sampling period for the next scan line. This means that the sampling rate (total scanlines per second) can vary between around 5 Hz in a dark environment to nearly 100 Hz in

bright daylight. The sampling rate is never allowed to drop below 5 Hz; if it was allowed to get arbitrarily slow in low-light conditions, response of the unit would appear to lag because updates could happen quite slowly compared to the speed at which the user's head might scan across its environment.

Since this system looks where the user's head is pointed, and not where the user's eyes are pointed, some feedback is necessary to show the user what region the system is imaging; this is most useful for new users. To accomplish this, three 5mW, 670nm laser diode assemblies surround the lens, in fixed, rigid optical alignment with it. Their aim is calibrated to surround the imaged region with three laser spots; and their intensities vary with the current exposure setting (integration time) for the spectrometer—hence, in a dim room, the spots also will be dim. Note that this strategy simplifies the optics considerably compared to a single laser spot that is instead centered in the imaged region.[1] A representation of the resulting system is shown in Figure 2, with the three beams collapsed into one to simplify the drawing.

Processing of the spectrometer's output is performed by a simple DSP chip, currently a TI 320C50, 28 MIPS, fixed-point unit, and is integrated with the A/D and D/A converters and some memory on a 6 by 10 cm board. (The DSP's software is described below, in section 2.3.) Power for the entire system is provided by a single Sony NP-F530 Li-ion half-height camcorder battery, which provides 1350mAh at 7.2V and weighs a bit less than 100g, and is run through a pair of Maxim voltage converter IC's to generate the required positive and negative regulated supply voltages. A fully-charged battery will run the unit for several hours depending on the volume of its headphones and the intensity selected for the laser diodes.

The spectrometer itself is mounted on the back of the user's head, with its sensor assembly (lens and laser diodes) on one side of the head and connected by the fiber optic cable. The DSP mounts on the other side of the user's head, and, currently, the battery sits on the top of the head. This approach distributes the hardware's mass relatively evenly around the head so that no net torque is exerted on the user's neck; this makes the unit comfortable enough to wear for extended periods.

2.3 Sonification

The spectrometer provides the DSP with a byte stream, consisting of 256 8-bit samples of the incoming spectrum (one detector scan line) at 5-100 Hz depending on exposure. Each scan line from the detector

1. If the laser spot was actually in the imaged region, its energy would have to be subtracted from the resulting spectrum, either by ignoring that particular pixel's value, by time-slicing the beam so that the detector was never sensitive when the beam was on, by oppositely polarizing the beam relative to the detector, or via a narrowband notch filter, such as a holographic optical element. All of these techniques have various disadvantages, and also show the user only the *center* of the imaged region, and not its full extent.

must be converted into a chord. In principle, 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.[1]

In detail, there are several additional steps which raise the quality of the output. First, we need to re-scale 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 detector at each wavelength
- Log-scaling of overall brightness and normalization of total brightness to some standard
- 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 spectrometer bandwidth.

Several of these steps are combinable into a smaller number of lookup tables which have been precomputed when the DSP is programmed. In practice, this has not yet been necessary.

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 [6], but is easiest to program; see section 4 for some alternatives.

3 A sample application

The Visor was designed enable easy, serendipitous discovery of material properties, by allowing the user to learn a mapping from visual spectrographic data into sound. A typical example is presented briefly below, followed by a very brief summary of other experiments.

---

1. The actual system running on the 320C50 is capable of producing 32 frequencies simultaneously without saturating the processor, so the current prototype maps 8 adjacent wavelength buckets (each about 3nm wide) into one generated tone. Any of a number of faster DSP's would have no trouble mapping all 256 input amplitudes directly.

3.1 Differentiation of materials

A good color photocopier can produce output that is virtually indistinguishable from its input—but only to the typical human color vision for which the machine was designed. A single green leaf was photocopied on a Canon 700L color copier and the leaf was then mounted on the output sheet of paper next to the copy. Informal polling showed that other observers could distinguish the two images at no better than chance if carried out immediately after photocopying, while the leaf was still fresh.

However, when the two images were compared with the Visor, they did not sound similar: the color copy had distinct low overtones which mapped to the infrared, and not all of the higher tones sounded quite the same, though it was difficult to describe how verbally. This study was conducted single-blind, e.g., an assistant randomly imaged either the leaf or the copy, without telling the subject (the author) which was being imaged and, it is hoped, without other cues. Under these conditions, the subject guessed with excellent accuracy which was which.[1] To investigate this, spectroscopic data from the leaf and its copy were captured from the spectrometer used in the Visor[2] and then plotted; this data appears in Figure 4.

The plotted data shows each spectrometer pixel in the range 305nm to 1000nm; each pixel is 3.125nm wide. Absolute amplitudes between plots are not significant. A plot of the leaf and its copy are shown, along with the difference between these spectra, and a single 670nm laser as a calibration point.[3]

The majority of the difference heard between these samples corresponds to a strong IR emission peak in the copied image of the leaf. White paper elsewhere on the same sheet also exhibited this IR reflectance; therefore, the conclusion is that the dyes used in the copying process fail to attenuate IR. Since the copier is designed to have its output viewed only in the visible spectrum, this characteristic of the dyes goes unnoticed by ordinary users. The rest of the auditory difference is attributable to the slight differences in the shape of the visible-light peak in the green; the leaf has a much sharper cutoff at the red end, for example.

---

1. When presented with both spectra in alternation, separated by under ten seconds between them, the subject could always differentiate the two and could specify which was which. When presented in isolation (two minutes between trials, several hours after the alternation test), accuracy for the subject tested was approximately 70% over 7 trials, with improving accuracy as the test continued, apparently due to memory effects.
2. The spectra captured in this test were obtained by illuminating both targets with a white xenon lamp with output spanning approximately 400-900nm.
3. The laser's output is not perfectly one pixel wide all the way down to the noise floor, due to bleed effects between detector pixels for such a bright source. Careful attenuation could eliminate this effect.

Similarly, anecdotal results outdoors during Fall in New England indicate that not all plants which turn similar-looking colors sound the same, even when identically illuminated.[1] This is unsurprising, since not all plants employ exactly the same chemistries or they would all look identically colored.

3.2 Other experiments

The Visor was presented to approximately 30 users informally in an office environment during a conference poster session. Lighting consisted of a mix of fluorescent and daylight through ordinary window glass (which is a strong IR attenuator [1]). This demonstration used a speaker system to enable everyone nearby to hear the resultant audio, and not just the wearer of the device. Some users actually wore the device; others observed where it was aimed and listened to the resulting audio, but did not actually don it.

Targets employed consisted of a variety of materials, including a brightly-colored poster (3 by 4 feet) printed on an HP 650C DesignJet inkjet printer, a red nylon jacket, blank pieces of white paper, interior carpeting, and other common objects. The poster had large red, green, blue, and yellow areas printed on it at high saturation; no listeners had difficulty identifying differences in the resulting sonification of these areas. This result held true even for those listeners who did not look at where the user was aiming the device, e.g., in a single-blind trial. Approximately half of all listeners were able to differentiate between the red of the poster and the red of the nylon jacket after several (one to five) alternations between the two (the jacket was not as reflective in the IR as the poster, among other differences). The neutral gray carpet and the white paper were substantially more challenging, with results approximating chance, since both of them tended to have fairly flat emission spectra (recall that two spectra with approximately the same shape but different absolute amplitude will sound very similar due to the gain-normalization of the device).

In short, we demonstrated that objects with substantially different spectra can be recognized as such even by users with training of a few minutes or less. These objects were easily differentiable by normal human color vision, so, in this experiment, the Visor did not enable users to perceive anything they could not already do—though several users said that they found the experience entertaining.

4 Future Work

The Visor is a research prototype, constructed to evaluate whether sonification of spectral data was a worthwhile application in the first place. To make it more useful, there are several ways that one could proceed. One of the most useful additions would be the addition of *color constancy* [2] to the device, such that

1. Since the Visor does not yet have color constancy (see section 4), objects could only be safely compared which were in the same scene at the same time of day, lest changing daylight colors throw off the calibration

its output would be related to the ratio of illumination colors versus object colors, in the same way that human color vision uses a *center/surround* strategy to also compensate for varying illumination spectra—hence making the same object, whether lit by incandescent or fluorescent light, sound the same because it looks the same. Such an addition requires a relatively small hardware change, and small amounts of software, but space does not permit a detailed design description here.

The sonification employed by the Visor is the simplest possible mechanism that could be devised. It is well-known from experimental data, however (for example, [6], p. 270) that individuals can only identify five to nine sonic stimuli that vary along only a single attribute, such as pitch. Employing multiple attributes simultaneously greatly enhances identification—see, for example, [15], in which eight dimensions were used to enable observers to identify about 125 different sounds. Using varying timbre [6] would help. Spatialization of the audio [17] could also be valuable; when using broadband sources, such as that produced by the Visor's monster chord, subjects can often accurately determine the point of origin of sounds to within 6 degrees in either horizontal or vertical directions ([5], in Middlebrooks, p. 84). Finally, melody (see the discussion of a laboratory measurement system at [10], p.46, which referred to [12]) or plucked notes may also help, though listening to an entire melody takes more time to use.

Finally, the Visor could be extended to other senses, such as the *near ultraviolet* and *near infrared*, where a great wealth of chemical and compositional data is available [1]. A *polarimeter* could allow the user to navigate using polarized light in the daylight sky, like a honeybee [18], or to notice the polarization effects of some materials under mechanical stress [1]. *RF field sensors* or a *magnetometer* would allow the user to perceive electronics and electrical systems, such as powerlines in walls.

5 Conclusions

An implemented 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.

Work is currently proceeding on extending the sensory capabilities of the instrument, improving its sonification of the resulting data stream, and manufacturing several copies to enable larger-scale user testing over much longer time scales.

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] Cytowic, Richard E., *Synesthesia: A Union of the Senses (Springer Series in Neuropsychology)*, Springer Verlag, 1989.

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

[5] Gilkey, Robert H., and Timothy R. Anderson, eds., *Binaural and Spatial Hearing in Real and Virtual Environments*, Lawrence Erlbaum Associates, 1997.

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

[7] Harrison, John E, and Simon Baron-Cohen, eds., *Synaesthesia: Classic and Contemporary Readings*, Blackwell Publishers, 1996.

[8] Hertzman, Clyde, Stephen D. Walter, Lynn From, and Adrienne Alison, "Observer Perception of Skin Color in a Study of Malignant Melanoma," *American Journal of Epidemiology*, Vol. 126, No 5, The Johns Hopkins University School of Hygiene and Public Health, 1987.

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

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

[11] Lide, David R., editor-in-chief, *The CRC Handbook of Chemistry and Physics*, 76th ed, CRC Press, 1996.

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

[13] Marchesini, Renato, Marco Brambilla, Claudio Clemente, Massimo Maniezzo, Adele E. Sichirollo, Alessandro Testori, Daniele R. Venturoli, and Natale Cascinelli, "In Vivo Spectrophotometric Evaluation of Neoplastic and Non-Neoplastic Skin Pigmented Lesions—I. Reflectance Measurements," *Photochemistry and Photobiology*, Vol 53, No 1, pp.77-84, Pergamon Press plc., 1991.

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

[15] Pollack, I., and L. Ficks, "Information of elementary multidimensional displays," *Journal of the Acoustical Society of America*, volume 26 no 2, pp. 155-158, 1954.

[16] Pentland, Sandy, Thad Starner, and Kevin Pipe, personal communication.

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

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

[19] Yost, W. A., and D. W. Nielsen, *Fundamentals of Hearing*, 2nd edition, Holt, Rinehart, and Winston, 1985.

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

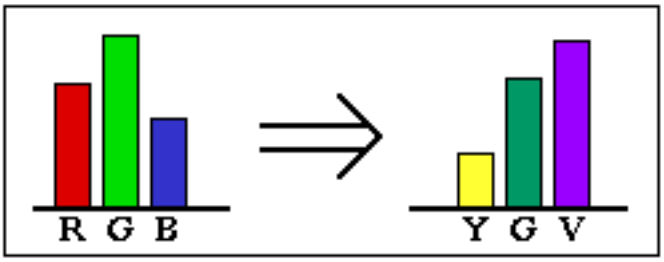

Figure 1. Different sets of wavelengths, with amplitudes properly adjusted, can appear as the same color.

Figure 2. Wearable version with fiber spectrometer.

Figure 3. The Visor in actual use.

.

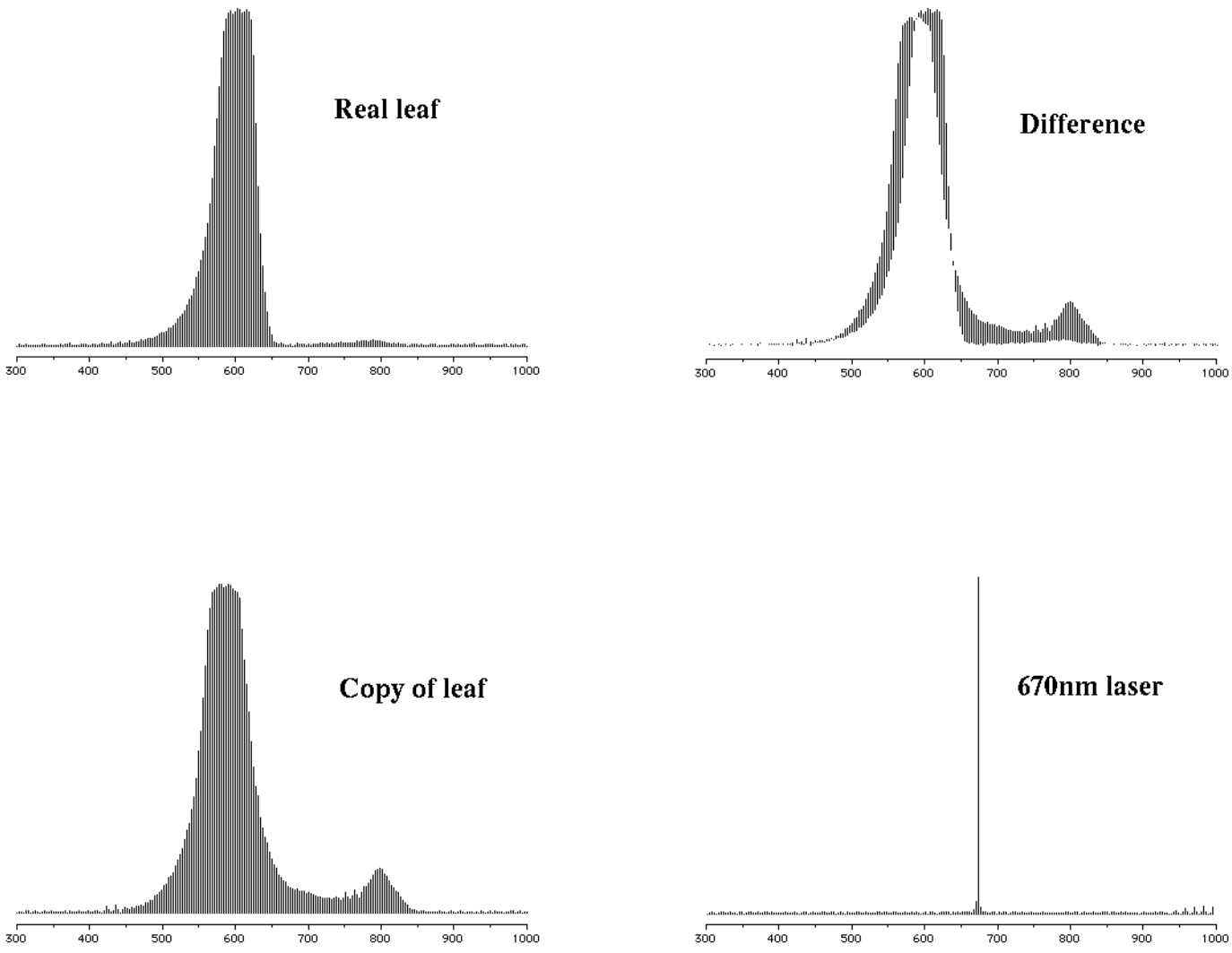

Figure 4. Measured spectra of a leaf. Wavelengths are in nm; amplitudes are arbitrary. See text for details.