G. McMorrow, X. Wang and P.F. Whelan (1997), "A colour-to-speech sensory substitution device for the visually impaired", **Proceedings of the SPIE - The International Society for Optical Engineering**, Vol. 3205 - Machine Vision Applications, Architectures, and Systems Integration VI, Pittsburgh (USA), pp 272-281.

A Color-to-Speech Sensory Substitution Device for the Visually Impaired.

Gabriel McMorrow [a], Xiaojun Wang [b], Paul F. Whelan [a]

[a] Vision Systems Laboratory, Dublin City University, Ireland
[b] Microelectronics and Materials Group, Dublin City University, Ireland

Abstract

A hardware device is presented that converts color to speech for use by the blind and visually impaired. The use of audio tones for transferring knowledge of colors identified to individuals was investigated but was discarded in favor of the use of direct speech. A unique color-clustering algorithm was implemented using a hardware description language (VHDL), which in-turn was used to program an Altera Corporation's programmable logic device (PLD). The PLD maps all possible incoming colors into one of 24 color names, and outputs an address to a speech device, which in-turn plays back one of 24 voice recorded color names.

To the author's knowledge, there are only two such color to speech systems available on the market. However, both are designed to operate at a distance of less than an inch from the surface whose color is to be checked. The device presented here uses original front-end optics to increase the range of operation from less than an inch to sixteen feet and greater. Because of the increased range of operation, the device can not only be used for color identification, but also as a navigation aid.

Keywords: color, sensory substitution, clustering, VHDL, speech

1. Using Audio Tones for Color Recognition.

This project originally intended to use audio tones as the method to convey color to a blind or visually impaired person. This original choice was partly due to the fact that the device was intended to be a small hand-held piece of hardware, and speech synthesis might complicate the design drastically. The first method attempted was the writing of VHDL code to recognize one of 16 colors, by outputting a square wave whose frequency identified each color, i.e. a color-controlled oscillator. These square wave outputs then need to be converted to sine waves with enough current to drive a small speaker. To make these audio tones more recognizable to the user, one might map these 16 tones to two octaves of the diatonic music scale. This could mean that a blind person could be trained to associate musical notes with colors by learning the musical scale. However, the use of two octaves restricts us to the use of only 16 colors. One possible method of increasing this range is to use the intensity of each tone. The intensity could be broken out into four different levels, i.e. dark red, red, bright red, pale red.

But how can we transfer this knowledge of intensity to the user without using more than 16 tones? One method would be to play these 16 tones, or notes, using four different musical instruments, i.e. middle 'C' on the cello might indicate "dark red", whilst middle 'C' on the "harmonica" might indicate "bright red", etc. However, we now need to electronically synthesize four waveforms which would approximate four musical instruments, (note: they would only need to be discernable from each other as a 'cello' is from a 'harmonica'). *What makes middle 'C' on the 'cello' sound different from middle 'C' on the harmonica?* The answer is the number of harmonics of the fundamental frequency of middle 'C' generated by each instrument, coupled with the relative amplitudes of these harmonics. [11]

Rather than electronically synthesize the sound of each instrument, the speech chip used in this project, see Section 5, could be used to record different instruments. We accept that there may be applications where the use of audio tones may be preferable to the spoken color name, but we contend that in this particular application, the use of direct speech is by far the better choice.

2. Color Clustering

What is color clustering and is it necessary? The answer to these questions will become intuitively clear to the reader, by examining the results of trying to identify colors without using a clustering technique. Before proceeding, however, let us examine some fundamental properties of color. Every color we see, is in fact a combination of the three primary colors, red, green and blue. The "color cube" illustrated in Figure 1, is a graphical representation of this fact.

As a first attempt at identifying colors, the following technique was examined. The first step in color identification had to be splitting the incoming light into it's red, green, and blue primary component intensities. In order to use digital circuitry to decide what the incoming color is, these primary intensities need to be converted to analog voltages and then digitized.

This naturally leads to the following question: *What should the resolution of the digitization process be?*, i.e. how many bits should the analog RGB components be digitized into. Suppose that 2-bit digitization were used. This would provide the ability to discern 4 different shades of red, green, and blue. This in-turn implies we have the ability to identify 64 colors. If such a scheme were used, it would be necessary to give names to these colors. To aid in naming these colors, they first needed to be displayed.

Once displayed, it became obvious that it was futile to try to give names to these colors without using some kind of quantitative method. For this purpose an RGB Color Table[3] was used. The disadvantage of using this method would be the use of color names not used in daily conversation, names such as "Cinnabar Green", or "Medium Aqua Marine". These are names that most sited people might have difficulty visualizing in their minds, and are not suitable for describing colors to people blind since birth. To a certain extent this is a debatable point, however the next disadvantage is particularly critical.

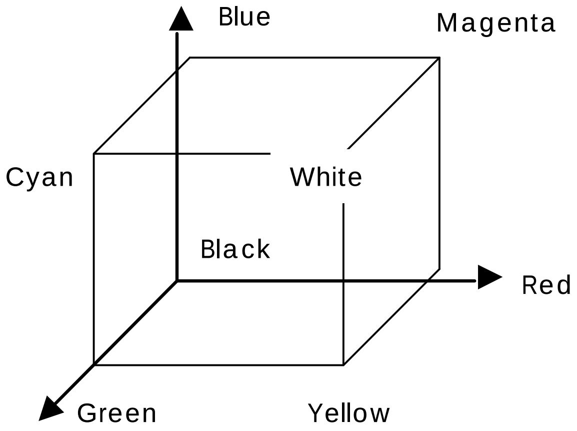

Figure 1. Color cube.

The diagonal of the color-cube shown in Figure 1, is a line in color-space where the red, green, and blue components are equal to each other. This in fact is the grayscale, which extends from black, through dark gray, and finally to white. This diagonal line of grayscale is extremely sensitive, such that points slightly off the diagonal contain color, not grayscale. When other colors other than grayscale are explored the error does not appear as serious, as it is difficult to tell, for example, close shades of green apart, however, for grayscale, small tints of color are noticeable. The only way to reduce this quantization error is by increasing the color resolution, i.e. increasing the number of shades of the primary colors, red, green, and blue. It was found that 4-bit digitization gave a visually acceptable color quantization error for grayscale color bins.

However, using 4-bit digitization provides the ability to discern 2^4 x 2^4 x 2^4 = 4096 color bins. *Does this imply we need to derive 4096 different color names?* The answer is no, because there are large regions or clusters of the same colors. For example, there are large regions of yellow, orange, green etc. This was observed when using MATLAB's Image Processing Toolbox to display colors found at the centroid of 4096 color bins. Having the colors displayed enables the viewer to visually segregate or cluster regions of red, orange, yellow, etc. Figure 2 illustrates 3 of 16 layers of the color cube digitized to 4-bit Red, Green, and Blue.

3. Design of Cluster Logic Using VHDL

As logic designs get larger, gate-level descriptions become unmanageable, making it necessary to describe designs in more abstract ways, and mandatory to adopt a higher-level, top-down design approach.[1]. Logic diagrams and Boolean equations have been used as methods for hardware description, but the complexity of these methods increases rapidly as the system complexity increases. Hardware Description Languages (HDLs) evolved as a solution. An HDL is similar to a high-level software programming language and provides a means of:

1). Precise yet concise description of a design at desired level of detail;
2). Convenient documentation to generate users manuals, service manuals, etc.;
3). Design capture for automatic design systems, e.g. high-level synthesis tools, and silicon compilers;
4). Incorporation of design changes and corresponding changes in documentation efficiently;
5). Design/user communication interface at the desired level of abstraction.

In 1987 the IEEE Std-1076 VHDL (Very High Speed Integrated Circuit Hardware Description Language), compiler became an industry standard with full support of the United States Department of Defense (US DOD). As of 30 September 1988, VHDL is mandated by the US DOD for documentation for all application specific integrated circuit (ASIC) designs developed in military contracts.

In VHDL, a hardware component is represented by the design entity, which in turn is a composite of an entity declaration and an architecture body. The entity declaration provides the "external" view of the component's ports, which are the points through which data flows into the device from the outside world. This includes the component's ports, which are the points through which data flows into and out of the component. [2] . The entity declaration for the clustering device is as follows:

```
ENTITY  cluster IS
    PORT( red: IN bit_vector(3 downto 0);
          green: IN bit_vector(3 downto 0);
            blue: IN bit_vector(3 downto 0);
           color: OUT bit_vector(9 downto 0)
     );
END cluster;
```

This VHDL code defines three input ports to the device, four bits red, green, and blue, and defines the output to be a 10-bit address to be sent to the speech chip, see Section 5.

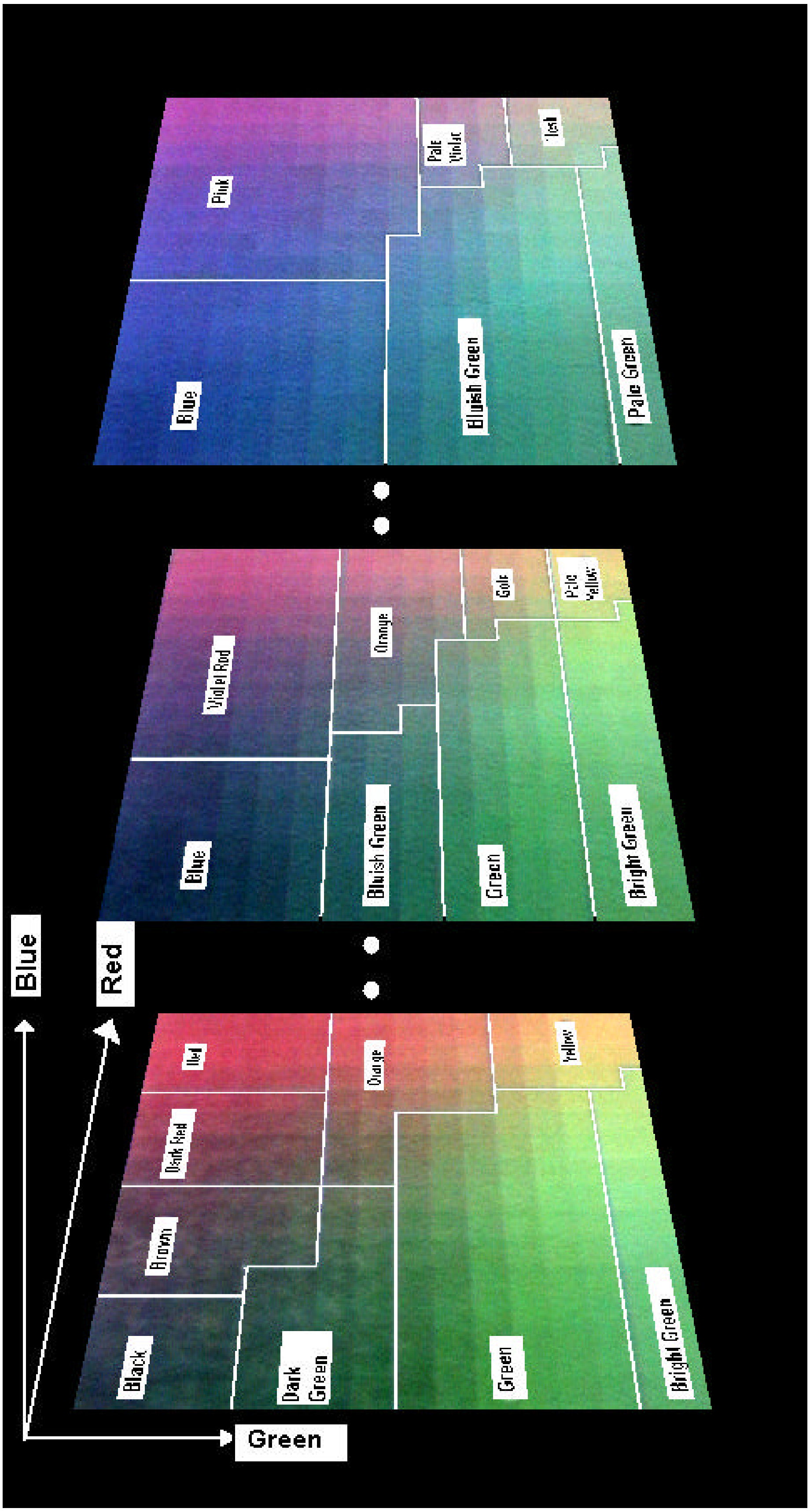

Figure 2. 3 of 16 layers of the color cube digitized to 4-bit Red, Green, and Blue.

The architecture body provides the "internal view"; it describes the behavior or the structure of the component. The following VHDL is a shortened version of the architecture body used to define the behavior of the cluster device, and is almost a direct transcription of the visual clustering shown in Figure 2.

```
ARCHITECTURE arch1 OF cluster IS
BEGIN combin1: PROCESS(red, green, blue)
                CASE  blue IS   -- cluster the first 6 layers of the RGB color cube
WHEN "0000" | "0001" | "0010" | "0011" | "0100" | "0101" =>
IF green <= "0011" THEN IF red <= "0011" THEN color <= "0100110111";
ELSIF ( ( red = blue) and (red = green) )  THEN color <= "0101000010";
ELSE color <= "0100101010";
END IF;    etc. etc .
END PROCESS; END arch1;
```

The architecture body has a process sensitivity list. This means the device is inactive, and only becomes active when it senses a change of the input signals red, green, and blue. The entire architecture essentially makes decisions of the form: if there is this amount of blue, then if there is this amount of green, and finally, if there is this amount of red, then the color is pink, say. Therefore, output the appropriate 10-bit address to enable the speech chip to play back the voice recorded color name, "pink".

The VHDL code was compiled and exhaustively simulated by stimulating with all 4096 possible combinations of red, green, and blue.[8] The code was debugged until all simulation outputs were as expected.

4. Front End Optics and Sensors

Of the two color-to-speech conversion systems available on the market at the time of writing, both need to be operated within an inch of the surface whose color is to be checked. A main goal of this research was to design a device, which would operate over much larger ranges. The use of a color camera to convert the incoming light to current was discarded in favor of a trio of photodiodes. This was done since another goal of this research was to keep the cost of the device to a possible end-user as low as possible to enable it to be purchased by Health Authorities.

In order for the device to operate at greater distance ranges, it is necessary to use a lens to gather the incoming light, and project it onto the light sensitive surface of the photodiodes. In addition, the lens is required to have a narrow field of view (FOV). The narrow field of view is important so that the color of an object's surface is tested at a minimal surface area. This is to avoid the identified color being an averaged color, i.e. the result of checking a large FOV encompassing multiple surfaces of different colors.

It would be preferable to have the minimal surface area being checked to remain constant over a large range of operating distances. This would be possible only through the use of an auto-zoom lens. Consideration of the increase in overall cost of the device with an auto-zoom lens resulted in the use of a fixed single lens. The disadvantage of using a fixed single lens is that the minimal area being checked grows in proportion to the distance the surface being checked is from the lens.

Therefore, using a single fixed lens implies that the continuity of color of the surface in question determines the range the device must be operated at to get a correct color reading. For example, pointing the device at a cloudless sky should return blue because although the minimal surface area being checked is large, the sky should be a constant blue over all the area being checked. Similarly, the green grass in a field should return green when the device operates at large ranges. However, for surfaces whose color changes over areas of, say, an inch diameter, the device should be within 8 feet of the surface, to get correct color readings, etc.

As already indicated, it is necessary that the light gathering lens of the device have a narrow FOV. Figure 3 illustrates how a plano-convex lens can be used to limit a detector's FOV.[5] The angle β subtends half the field of view of the lens, and is given by:

$$b = \frac{D2}{2 * f} \textit{ radians}$$

where D2 is the diameter of the photodiode used to convert the light to current, and f is the focal length of the lens. Placing the detector one focal length behind the lens ensures focus of objects one focal length in front of the lens. However, a focussed image is not important in this application. It was found that by placing his eye one focal length behind the lens of choice, using plastic tubing between eye and lens to exclude all light but that which comes through the lens, that objects one focal length in front of the lens were sharply focussed. Surfaces at distances greater than one focal length in front of the lens become increasingly blurred. However, this did not appear to affect the color of the surface. It was also seen that the reduction in FOV resulted in a single incoming color at a time, scanning over an average sized room.

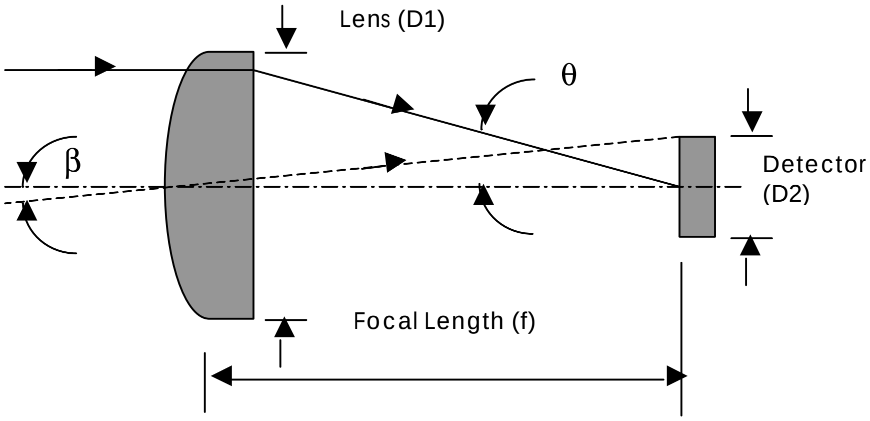

Figure 3. Using a plano-convex lens to limit a detector's FOV.

Obviously it is important to center the photosensor on the axial line of the lens. Knowing where the light will go is, however, only the first step in the design of a light projection system; it is just as important to know how much light is transmitted. The light gathering power of a lens is given by:

$$F\_number = \frac{f\ (focal\ \ length\ \ of\ \ lens)}{D\ (diameter\ \ of\ \ lens)}$$

$$NA\ (numerical\ \ aperture) = \frac{0.5}{F\_number}$$

Figure 3 shows how the numerical aperture controls the angle of light accepted by the lens. As the F_number decreases, θ (the acceptance angle) increases and the lens becomes capable of gathering more light. In the design of any lens system it is important to consider the term throughput (TP), a quantitative measure of transmitted light energy. Because the photodiode is an area and not a point, lens diameter affects throughput even when the F_number remains constant. The following equations were used to calculate the throughput of the lens.

$$X = \frac{D2}{2f}\ ;\ Y = \frac{2f}{D1}\ ;\ Z = 1 + (1 + X^2) * Y^2$$

$$G = 0.5 * (Z - \sqrt{Z^2 - 4X^2Y^2})$$

$$Throughput\ \ (TP) = G * p^2 * (\frac{D1}{2})^2$$

Because the device is to be physically small, a small lens diameter is required whilst at the same time maintaining a large throughput. For example for a particular lens system the TP's for each of three lens choices are: TP (6mm lens) = 16.77, TP(12mm lens) = 19.04, and TP(25mm lens) = 19.69. In this example the 12mm diameter lens is the best compromise between diameter and throughput. The following table is a subset of calculations that resulted in a final choice of the plano-convex lens used in this application:

Focal Length F	Lens Outer Diam (mm) D1	½ FOV degrees β^o	TP	Object Area Diam Inches at 1 ft.	Object Area Diam Inches at 2 ft.	Object Area Diam Inches at 3 ft.	Object Area Diam Inches at 4 ft.	Object Area Diam Inches at 8 ft.	Object Area Diam Inches at 16 ft.
75	18	1.03	0.255	0.216	0.432	0.647	0.863	1.72	3.45
100	19.8	0.77	0.175	0.161	0.323	0.484	0.645	1.29	2.58
100	25	0.77	0.277	0.161	0.323	0.484	0.645	1.29	2.58

Using the table as a guideline, and given a restricted choice of C-Mount lens holders led to the choice of the 100mm focal length lens with diameter 19.8 mm. The relatively small value of TP = 0.175 imply that the photodiodes will output smaller currents, but this may be compensated for by electrically amplifying the photodiode currents.

At this stage in the design, it becomes necessary to choose a particular photodiode from a vast array commercially available devices. All photodiode specification sheets contain a graph showing the relative spectral response characteristics of the sensor. This plots the relative sensitivity of the sensor versus the wavelength of incoming light in nanometers (nm). The visible spectrum of light ranges from 400nm to about 700nm. The color bands of the visible spectrum span the following wavelengths: [11]

From 400-450nm	Violet
From 450-500nm	Blue
From 500-550nm	Green
From 550-600nm	Yellow
From 600-650nm	Orange
From 650-700nm	Red

It is known that the human eye's response to light is most sensitive or peaks in the Green, or 550-600nm wavelengths. Searching through the Toshiba Photo Sensor Catalog, in particular, it appeared that most devices peak in the near if not infrared. However, a photo-diode whose response matches closely the human eye was found. The device is the BPW 21 photodiode, whose main application is exposure and color measurement. It comes in a hermetically sealed case, has a radiant sensitive area of 7.5 mm^2, and it's typical sensitivity is 7 nA/lux. Typical room light is about 700 lux.

Before considering the amplification used with these photodiodes, the important topic of color filtering needs to be addressed. The incoming light needs to be split up into it's primary color components, (red, green, and blue), prior to being converted to electrical signals. Various color filter options are available. Acrylic filter material is cheap but the filter responses were found to be substandard for this application. Glass color separation filters have excellent filter responses and are used in machine vision applications. Two choices of design for color separation was examined, see Figure 4.

Choice B requires three lenses. These would have to be small in diameter and arranged in a mosaic structure similar to a fly's eyes. Choice A is by far the better choice as only a single lens is used whose diameter may be large. Choice A is made possible by the availability of Color Separating Dichroic Filters. A light filtering system using a 45 degree blue reflector/corrector, 45 degree red reflector/corrector, and green corrector set allow a single source of light to be split into it's component parts. An aluminum structure was designed and manufactured to hold the glass filters at their appropriate angles, and support the three photodiodes.

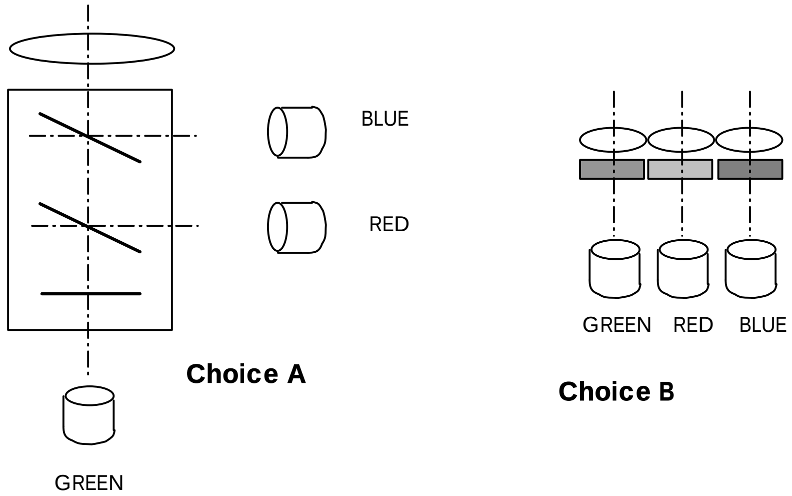

Figure 4. Color separation design choices.

5. Hardware Design

Analog to Digital Conversion

Assuming that the incoming primary component intensities have already been converted to analog voltages, it is then necessary to convert the RGB analog signals to 4-bit binary logic levels. This will form the appropriate input into the cluster device described previously. The A/D device chosen was Analog Devices' AD7575, which uses the successive approximation conversion technique. This provides a necessary high conversion speed of 5 microseconds. A standard 555 timer was used to control the start of conversion and the reading of data from the AD7575.

Light Sensor Amplification

Next in consideration is the amplification of the individual red, green, and blue signals coming from the photodiodes. It is desired that the A/D converters output 4-bit binary patterns spanning the range of 0 to 15. To accomplish this it is necessary for the output signals of the photodiodes to span the analog voltage range 0 to 2Vref, i.e. 0 to 2.4V. Single supply TLC271 operational amplifiers were chosen for amplification using the following configuration: [6]

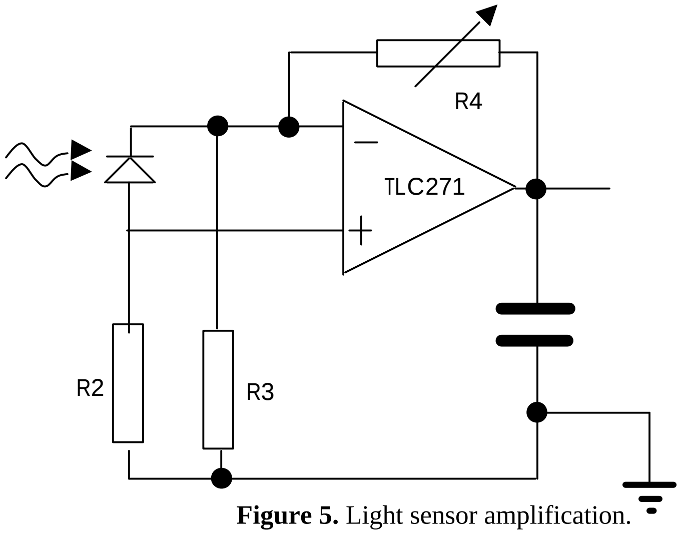

Figure 5. Light sensor amplification.

Through examination it was found that to increase the ambient light rejection, i.e. current flows through the photodiode when all light is absent, that R2 should be small. This implies that the gain is essentially R4 / R3. The following values were used: R2 = 1 kΩ, R4 = 2 MΩ potentiometer, R3 = 1 kΩ. R4 was set with the sensor receiving pure daylight, i.e. light reflected from bright cloud cover to yield 2.4V, the maximum voltage range of the A/D converter. This resulted in all LED's on the outputs of the A/D converter to be lit. By increasing covering the face of the photodiode, the LED's were seen to go through the 4-bit binary pattern from 15 to 0.

The ISD2590 Single-Chip Voice Record/Playback Device

The ISD2590[4], manufactured by Information Storage Devices provides an excellent method for verbal description of identified colors required in this project. It allows cueing of recorded color names, which can then be directly addressed. Therefore, once the incoming red, green, and blue intensities have been digitized, and clustering performed by the Altera Programmable Logic Device, one of 24 colors (addresses) can be input to the speech chip to cause it to "speak" out the name of the color. This EEPROM device allows analog data to be written directly into a single cell without A/D or D/A conversion. This results in increase in density over equivalent digital methods, and non-volatile storage. The ISD2590 allows a total of 90 seconds recording time. It provides excellent voice quality, and the two output pins provide direct speaker drive capability of about 12.5 mW RMS (25 mW peak) into a 16 Ω speaker. This is enough to be clearly heard from the other side of an average room. The external components; a microphone, loudspeaker, switches, a few resistors and capacitors, and a power source, are all that is required to build a complete voice record/playback system. All other functions; preamplifier, filters, AGC, power amplifier, control logic and analog storage, are performed on-chip, see Figure 6.

The ISD2590 provides a total of 90 seconds of recording time, and a total of 600 individual messages. This gives rise to the term message resolution, which is 90/600 = 150 milliseconds. The address bits control the message start pointer (MSP). The value of an address to access a particular message x seconds into the recording is given as follows:

$$\text{Address (in decimal)} = \text{(x seconds)} / \text{(message resolution)}$$

i.e. to access a message 1.5 seconds into the recording use address = 1.5 / .15 = 10. To access a message 3.0 seconds into the recording use address = 3.0/.15 = 20, etc. These addresses, 10, 20,30 etc. require that each message be exactly 1.5 seconds long. However, it is impossible to record each different color name with a microphone, and have each recording exactly 1.5 seconds long. It was for this reason that it was decided to use the Windows Digital Speech Record/Processing package COOL. COOL provides a variety of editing features such as amplification and message length adjustment of the digitally sampled and stored recording. This allowed the calculation of the addresses needed to be presented to the speech chip to access each of 24 different color names, for example:

COLOR NAMES	**Address (binary) into chip**	**Address (Decimal) into chip**	**Time in seconds into the total recording**
DARK RED	"0000000000"	0	0
RED	"0000001110"	14	2.1
VIOLET RED	"0000011100"	28	4.2
PINK	"0000100111"	39	5.85
Etc.	Etc.	Etc.	Etc.

6. Conclusions

At the time of concluding this report we were still awaiting delivery of the red 45-degree reflector color filter. This restricted the full testing of the device. However, by manipulating light falling on the unfiltered red sensor, various intensities of RED could be achieved. When the device was presented with full blue, full red, and no green, pressing a button on the speech chip resulted in it "playing back" the word MAGENTA. Full red only generated the word RED, etc. This demonstrated correct operation of the A/D converters, the cluster logic and the speech chip. By tying the inputs of the A/D converters together, the grayscale was tested using a variable DC level in lieu of the sensors. For 0V

DC the device said BLACK. When the DC level was increased it said DARK GRAY, GRAY, and finally WHITE at DC level 2.4V.

This pseudo testing of the overall device brought to light many areas of improvement. The gain of the sensors is definitely a calibration issue. Lens choice provides a tradeoff between shorter focal length (larger FOV) versus lens diameter to provide larger throughput (TP). These tradeoff and calibration issues require further field testing of the device.

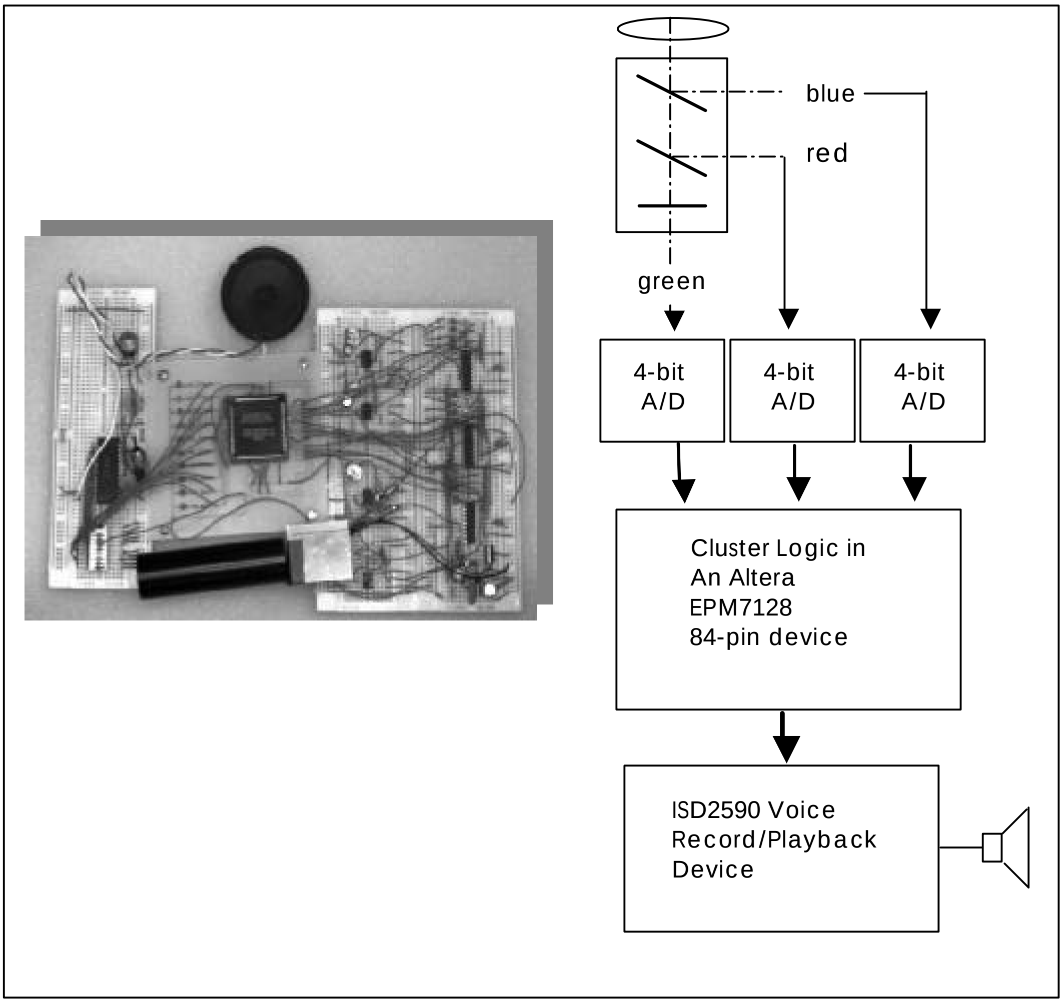
blue
red
green
4-bit A/D
4-bit A/D
4-bit A/D
Cluster Logic in An Altera EPM7128 84-pin device
ISD2590 Voice Record/Playback Device

Figure 6. Block diagram of complete system, and photograph of prototype.

7. References

1. X. Wang, "VHDL and High-Level Logic Synthesis", Course Notes, School of Electronic Engineering, Dublin City University, 1994.
2. R. Lipsett, *VHDL : Hardware Description and Design*, Kluwer Academic Publishers, 1989.
3. C. A. Lindley, *Practical Ray Tracing in C*, Wiley 1992.
4. Information Storage Devices', "Application Notes and Design Manual for ISD's Single-Chip Voice Record/Playback Devices", July 1994.
5. Edmund Scientific, *Annual Reference Catalog for Optics, Science, and Education*, 1994
6. Horowitz and Hill, *The Art of Electronics*, 1992
7. E. Parr, "IC 555 Projects", Bernard Babani 1994
8. Mentor Graphics Corp., "QuickVHDL User and Reference Manual", April 1994.
9. Altera Corp., "MAX+PLUS II Getting Started", 1994
10. Altera Corp., "Altera Data Book", 1993.
11. Sears, Zemansky, and Young, *University Physics*, Addison Wesley, 1976