AN AUDIOTACTILE VISION-SUBSTITUTION SYSTEM

David Dewhurst

HiFiVE
www.hfve.org
daviddewhurst@hfve.org

ABSTRACT

This paper describes "work-in-progress" on “HiFiVE” (Heard & Felt Visual Effects), an experimental vision-substitution system that uses verbally-orientated audiotactile methods to present certain features of visual images to blind people.

1.INTRODUCTION

Devices have previously been invented that substitute sight with another sense, particularly hearing and touch. Often such devices convey particular pieces information, such as the presence of obstacles [1]. Relief images such as tactile maps can display unchanging two-dimensional images, and the Optacon [2] used a vibrating matrix to display letters and other printed material. Tone-sound scanning methods have been devised for presenting text [3 & 4], and for general images [5].

Hearing and touch cannot fully replace the vast amount of information provided by sight, and it is difficult to devise audio and tactile equivalents for all visual effects. The HiFiVE system highlights features of visual images that are normally perceived categorically, and substitutes with coded sound effects and their tactile equivalents. It simulates the instant recognition of properties and objects that occurs in visual perception, by using the near-instantaneous recognition of phoneme sounds that occurs when people hear speech.

By smoothly changing the pitch and binaural positioning of the sounds, they can be made to appear to "move", whether following a systematic path or describing a specific shape. Such moving effects are referred to as “tracers”, and can be "area-tracers", which systematically present the properties of the corresponding parts of an image; or "shape-tracers", whose paths convey the shapes of particular items in an image. In the tactile modality, tracer location and movement are presented via a force-feedback device such as a joystick. Moving effects are generally easier to mentally position than stationary ones.

As the system outputs both audio and tactile effects, users can choose which modality to focus on; or both modalities can be used simultaneously, allowing more information to be presented during a certain period of time. Having a degree of “redundancy” of information may result in less tiring usage [6].

This paper focuses on the audio aspects of the system. Note that the example mappings are likely to change.

1.1.Coded phonetics

People can easily recognise speech-like sounds, and rapidly assign meaning to them. Speech is a natural and efficient method of conveying information, and the information content is not greatly effected by distortion. Using real-language words to describe features in images has been investigated before [7], but the HiFiVE system uses new speech-like sounds, consisting of specific “coded phonetics”, that can be rapidly interpreted in a categorical and linguistic way. These sounds convey the categorical properties of an image e.g. the colours, the distribution of those colours, recognised objects etc. Most people are able to retain several such spoken “non-language” words in their short-term memory [8], and the effort needed to learn the coded phonetics is low. (Speech-based coding has previously been used in several mnemonic systems that assign speech sounds to numerals in order to form real words [9].)

The visual effects conveyed to a blind user via heard and felt effects may not be perceived by them in the same way as they are by a sighted person who is directly viewing the same effects. However the visual information may be useful in itself.

Visual properties are normally presented to the user via groups of CV (Consonant-Vowel) syllables.

Figure 1 illustrates the approach. An image (A) is reduced to 8 by 8 pixels (B). The pixels in each square of 4 by 4 pixels (known as a “panel”) are each set to one of the two shades that the system calculates best represent the panel (C). Then the image is presented via audio (D) and tactile (E) methods. For each panel, one CV syllable conveys the two selected shades; and two CV syllables convey the arrangement of those two shades, to the level of detail shown in the pixelated image (C). (The arrangement of the two shades is known as a “layout”).

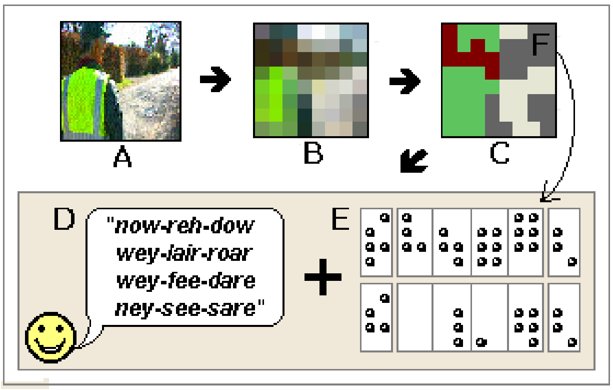


Figure 1. *Diagram illustrating conversion of an image into coded phonetics and braille.*

So for the top right “panel” (F) in the pixelated image (C), the CV syllable “wey” conveys the two colour shades “white and grey”, and the two CV syllables “lair-roar” present the “layout” of the two colour shades as 4 by 4 pixels. The whole image is conveyed by the four spoken “words” shown (D), and by the corresponding 12 braille cells (E), both of which fully describe the 8 by 8 pixels shown in (C). The user can control whether the system outputs audio and / or tactile effects to convey the colour shade pairs and / or the pixel “layouts”.

The example shown in Figure 1 shows one way of systematically presenting a section of an image. Many similar configurations can be devised.

1.2.Tactile effects and user interaction

The HiFiVE system's audio effects have tactile equivalents, which can be presented by using standard low-cost force-feedback devices to convey location and shape; and braille or other touch-based methods to convey the categorical properties.

If 16 consonants and 16 vowel sounds are used, 256 (i.e. 16 x 16) combinations of CV syllables are available. This is the number of different dot-patterns that can be displayed on an 8-dot braille cell (refreshable / programmable 8-dot braille cells are available commercially [10]). Figure 1 (E) shows one way in which the information conveyed by the spoken sounds could also be displayed on 12 braille cells.

A force-feedback joystick makes an effective pointing device with which the user can indicate areas of the image, as it can also be programmed to tend to position itself to one of a number of set positions, so that a "notchy" effect is felt as the joystick is moved, giving a tactile indication of location.

Sections of an image can be selected by the user via the pointer / joystick, so that only those parts are presented by the audiotactile effects, but at a higher resolution. The user can instruct the system to "zoom in" to present a smaller area, but in more detail, as well as to "zoom out" to present a low-resolution representation of the whole image.

A force-feedback joystick can also be moved by the system, pushing and pulling the user's hand and arm, both to convey any shapes that are to be presented (by tracing them out), and to indicate the area within an image that is currently being described via the audiotactile effects. (The user can override the joystick forces at any time, for example if they wish to change the section of the image that is being presented.)

Two force-feedback devices can be used : the main joystick can be used as a pointer by the user; and by the system to indicate the location and size of the area being presented. The other device, for example a force-feedback mouse, can be used by the system to present any shapes, the “tracer” being expanded in size to better convey the details of such shapes.

2.MAPPINGS BETWEEN IMAGES AND SPEECH

.

2.1.Selection of colour shades

The HiFiVE system generally allows a maximum of two colour shades (e.g. “blue and yellow”) to be used when presenting any particular “panel”. This approach allows the arrangement of those two colour shades to be effectively displayed via braille, and via a modest number of “layout” codes / mappings.

The two-colour-shade approach is analogous to painting a picture with one colour on a differently-coloured background.

Figure 2 shows an example of an image being reduced to two colour shades in each image quarter / “panel”.

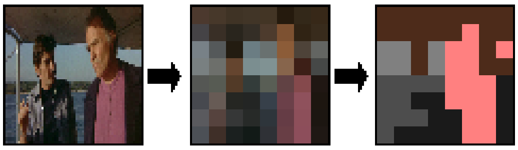


Figure 2. *Example of an image being reduced to pairs of colour shades within each “panel”.*

When one colour predominates (known as a “monochrome” panel) or more than two colours predominate (a “polychrome” panel), the system can convey this information, then present the “layout” of lighter and darker pixels within the panel.

2.2.Colour shade palettes

The HiFiVE system can convey monochrome versions of images, when a palette of five brightness levels is used. However colour gives users additional information about an image, and so it is useful to be able to include it.

Most cultures tend towards classifying any colour shade as one of eleven “Basic Colour Categories”, namely red, orange, yellow, green, blue, purple, pink, brown, black, white and grey [11]. This set of colours is available as a palette option.

However it is found that having more colour shades available in a palette can result in the clearer presentation of images. Currently the HiFiVE system also offers 15-colour, 22-colour and 24-colour palettes (the exact choice of colours shades can be varied according to the preferences of the user).

The 15-colour-palette format adds to the 11 “Basic Colour Categories” and allows two colours to be easily presented via a single “CV” syllable or via a single 8-dot braille cell.

The 22-colour-palette format can also be presented via a single CV syllable or braille cell, but only if more complex mappings are used : when two shades are conveyed, they can be presented in either order. The number of combinations of two of 22 colour shades (presented in either order) is 231 (22 times 21 divided by 2). The CVs for these mappings are allocated in a fairly arbitrary manner, though certain patterns can be devised. However as in this case no attempt is being made to match the mappings to particular English words, clearly-distinguishable CV pairs can be chosen.

The 24-colour palette format uses CV combinations chosen from five consonant and five vowel sounds for each shade, and can be used with any language's phonemes. However it needs two CV syllables or braille cells to convey a colour shade pair.

For all of these palettes, certain CV syllable combinations are reserved for signifying when a single colour shade predominates (“monochrome”), and for signifying when more than two dominant colour shades are present (“polychrome”).

2.3.Mapping between colour shades and speech sounds

One approach to mapping colour to speech is to base the speech on English colour names, as shown (for the first fifteen shades of the colour palettes) in the table in Figure 3.

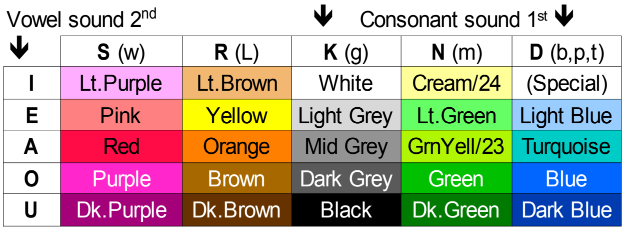
Colour Name	Short Name	“Split” Name	Vowel Sound	Phonemes
1 Red	Reh	S-eh	bed	S-EH
2 Orange	Joh	J-oh	bob	JH-AA
3 Yellow	Yow	Y-ow	boat	Y-OW
4 Green	Gee	N-ee	bee	N-IY
5 Blue	Boo	B-oo	boot	B-UW
6 Purple	Puh	Z-uh	bull	Z-UH
7 Pink	Pih	P-ih	bid	P-IH
8 Brown	Bow	R-ow	bout	R-AW
9 Black	Bah	K-ah	bad	K-AE
10 White	Wuy	W-uy	buy	W-AY
11 Mid Grey	Mey	M-ey	bay	M-EY
12 Dark Grey	Dey	D-ore	bore	D-OR
13 Light Grey	Ley	L-air	bear	L-XR
14 Turquoise	Toy	T-oy	boy	T-OY
15 Dk.Brown	Dow	V-ar	bar	V-AR

Figure 3. *Table showing colour-to-speech mapping for the first 15 colour shades, using “English” coding.*

This “English” coding makes use of the relatively large number of vowel sounds available in English; and the surprisingly varied set of vowel (V) and consonant (C) sounds used in the English names for basic colours. The “Short Name” column shows single syllable “names”, of C-V format, that are similar to the English colour names. Two syllables can be used to present colour pairs, for example “boo-yow” for “blue and yellow”. These sounds can be further contracted : notice how the vowels for the short-format names of the 11 “Basic Colour Categories” are all different. By changing the C or V of certain colours so that every C and V is different (see “Split Name” column), single-syllable colour pairs can be produced by taking the consonant of the first colour and adding the vowel of the second colour, for example “bow” for “blue and yellow”.

This mappings shown in Figure 3 are based on the phonemes found in spoken English, which could make comprehension difficult for people whose mother-tongue is not English, especially as they will not be able to use the context to help with understanding, as would occur with normal speech. Many languages use only five different vowel phonemes, centred around the vowels A, E, I, O and U.

The “International” format shown in the table in Figure 4 can use just five consonant and five vowel sounds (alternatives shown in brackets). For example the colour pair “blue and yellow” would be presented as “doh-reh”. A disadvantage of using this double-syllable format is that the sounds take longer to speak, though the individual syllables can be spoken faster than for the “English” format, as the consonant and vowel sounds are slightly shorter on average, and more distinctive.



Vowel sound 2nd ↓ ↓ Consonant sound 1st ↓

	S (w)	**R** (L)	**K** (g)	**N** (m)	**D** (b,p,t)
I	Lt.Purple	Lt.Brown	White	Cream/24	(Special)
E	Pink	Yellow	Light Grey	Lt.Green	Light Blue
A	Red	Orange	Mid Grey	GrnYell/23	Turquoise
O	Purple	Brown	Dark Grey	Green	Blue
U	Dk.Purple	Dk.Brown	Black	Dk.Green	Dark Blue

Figure 4. *Table showing colour-to-speech mapping using “International” coding.*

As the phonemes do not attempt to match the English colour names, other factors can be used when designing the mappings. There is a slight “synaesthetic” effect between colour shades and speech : if the vowel sounds are arranged in the order I-E-A-O-U, most people tend to find the “I” sound “lightest”, then the following vowels “darkening”, through to “U”, which gives the “darkest” impression. This effect is used when allocating vowel sounds to shades, as shown in Figure 4. There seems to be a similar, but milder, effect with consonants, with, for example, “R” better matching “warmer” colours, and “N” matching “cooler” colours. Apart from this factor, consonants have been chosen which are found in most languages, and which are clearly distinguishable when spoken.

One advantage of the “International” format is that users can get an impression of the lightness or darkness of an area from the vowel sounds alone.

For people who can easily distinguish English phonemes out of context, the “International” format vowels can be “coloured”, wherein the standard five vowel phonemes are replaced by similar “R”- or “Y”-sounding vowels when “warm” or “cool” colours respectively are being presented. For example “blue and yellow” could alternatively be presented as “doy-rare” (instead of “doh-reh”), allowing a user to know that the first colour (blue) is “cool” and the second colour (yellow) is “warm”, without fully interpreting the coded sounds.

For 5-level monochrome shades, the two levels can be conveyed via a single syllable even in “International” format, with the consonant sound conveying the first level and the vowel sound conveying the second level.

2.4.Mapping between “layouts” and speech sounds

As there are no standard English terms for the arrangements of colour-shaded pixels within a panel, the mapping of “layouts” to speech sounds is fairly arbitrary.

However some method can be used : for English speakers, a single-syllable CV format can be used to convey the patterns of 8 dots, with four dots conveyed by the consonant and four dots conveyed by the vowel. The vowel sounds can be chosen so that similar vowel sounds are applied to patterns that have the same number of dots raised (0 to 4), and the consonants chosen so that similar-sounding consonants are used for similar dot patterns. Two CV syllables are needed for a 4 by 4 panel.

For “International” format, six consonant and five vowel sounds are used, and one CV syllable is used to present four dot settings, the vowel sound signifying the number of raised dots (0 to 4). Four CV syllables are needed for a 4 by 4 panel.

So for example if the first four dots were all raised, and the next four dot positions all blank, the speech sounds to convey such an arrangement would be “s-oo” for “English” format and “see-doo” for “International” format.

When presenting the layouts of 4 pixel by 4 pixel panels, the system can output the arrangements in “column-by-column” or “row-by-row” order, or ordered as four squares of 4 pixels. However it is hoped that the users can soon learn to interpret the meaning of the layout sounds directly, without having to calculate the layouts from the individual phonemes.

(No mapping is required for the braille display of layouts, as the braille dot patterns can correspond directly to the colour-shaded pixel arrangements, as illustrated in Figure 1 (E & F).)

2.5.Cultural factors

The given mappings can be changed to allow for cultural issues (for example to avoid phoneme combinations that happen to produce unacceptable words). The speech sounds can also be adjusted to match the phonemes found in the users' languages.

3.OTHER FEATURES

Some of the other features of the HiFiVE system are described below (at the time of writing several of these features have not yet been implemented) :-

3.1.Conveying image “textures”

The textures of an area or entity can be conveyed via small fluctuations in the volume of the speech-sounds. These volume effects combine the effects of changes in brightness, colour etc., to give a single volume-conveyed "texture" effect. This simulates the effect found in vision whereby the overall properties of an area tend to be perceived categorically, while the minor variations across it are perceived as varying textures. The user does not need to follow the precise detail conveyed by the volume effects, but gets a general impression of the textures of an area or entity from the volume fluctuations.

3.2.Audiotactile entities

As well as conveying general visual properties and "layouts", the HiFiVE system can attempt to simulate the way in which features and objects are perceived in vision. Conveying basic properties does not do much to identify "entities", separate "figures" from the background, or assist with the other processes that occur naturally when people see things.

The system can highlight identified entities and features within a scene by exhibiting their size and shape via audiotactile "shape-tracers", at the same time as presenting their categorical properties etc. via coded phonetics and braille.

3.3.Audiotactile objects

"Audiotactile objects" are items in an image that have been identified to the extent that they can be presented as specific entities rather than being described in terms of their properties, shapes and features. They are signified by special CV syllables and braille patterns.

Initially, audiotactile objects will mainly be presented as part of pre-processed images or movies, but in the future automatic recognition of certain objects may be possible.

A shape-tracer can present the shapes of objects as they are found in an image. However it may be better to convey the distinctive "classic" shapes of objects, allowing the "shape constancy" visual effect [12] to be simulated, instead of the outline that happens to be formed by the object at its current distance and orientation.

3.4.Image "pre-processing"

When completed, the system will be able to convey prepared programmes of material. To create "pre-processed" images and movies, a sighted designer selects key entities within images, then specifies appropriate methods for presenting them.

Images and movie sequences prepared in this way could be transmitted through currently available media, for example via DVDs, the Internet or broadcasts, with the prepared presentation instructions being embedded in, or accompanying, otherwise standard video material.

3.5.Activity-based processing

It is useful to be able to specify and store activity-based filtering parameters that can be applied when users are doing a particular task, for example seeking items of a particular colour. A "selection filter" can decide which (if any) entities should be highlighted to the user. The filter's parameters specify how the system should value each of several visual features, for example colour shade; size; closeness to an ideal shape etc. Each visual feature is assigned an importance, which can include "essential" i.e. those entities whose features do not fall within certain bands of values are excluded.

The system presents as many of the qualifying entities as it can within the time available, according to their score. Normally such entities will replace the colour arrangement information in the areas in which they occur.

3.6.Highlighting vertices

When sighted people see things, vertices (such as the corners of a rectangle) produce a considerable effect in giving the impression of the shape, and it is useful if this visual effect can be reproduced in the audio and tactile modalities.

It is particularly important that vertices are highlighted when they are essential features of an entity but are not sharp angles, as is the case for an octagon, for example.

In both the audio and tactile modalities, to emphasise a vertex the system momentarily stops the movement of the "shape-tracer" (for example by momentarily stopping a moving force-feedback joystick); and in the audio modality the system also alters the volume of the sound.

4.SUMMARY

When fully implemented, it is intended that the HiFiVE system will allow a continuum of visual features, from basic visual properties, to fully-identified objects, to be conveyed to blind and deafblind users. At the time of writing several of the features described above can be demonstrated, as well as "work-in-progress" on some of the others.

5.REFERENCES

[1] L. Kay, "An ultrasonic sensing probe as a mobility aid for the blind". in *Ultrasonics*, 2, 53, 1964.

[2] P. Bach-y-Rita and B Hughes, "Tactile vision substitution: some instrumentation and perceptual considerations." in *Electronic Spatial Sensing for the Blind. (Eds. D.H. Warren, and E.R. Strelow)*, Matinus Nijhoff, 1985.

[3] E. E. Fournier d'Albe, "On a Type-Reading Optophone" in *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, Vol. 90, No. 619 (Jul. 1, 1914), pp. 373-375.

[4] M. Jameson, "The Optophone, or How the Blind May Read Ordinary Print by Ear" in *And There Was Light*, 18. Vol 1, No. 4, 1932.

[5] See website http://www.seeingwithsound.com.

[6] G. Jansson, "Implications of perceptual theory for the development of non-visual travel aids for the visually impaired" in *Electronic Spatial Sensing for the Blind. (Eds. D.H. Warren, and E.R. Strelow),* Matinus Nijhoff, 1985.

[7] C.C. Collins, "On mobility aids for the blind." in *Electronic Spatial Sensing for the Blind. (Eds. D.H. Warren, and E.R. Strelow)*, Matinus Nijhoff, 1985.

[8] G.A. Miller, "The magic number seven, plus or minus two: Some limits on our capacity for processing information" in *Psychological Review*, 63, 1956, pp. 81-93.

[9] The earliest such mnemonic system appears to be by J.J. Winkelmann (using the pseudonym "Stanislaus Mink von Wennsshein"), *Relatio Novissima ex Parnassus de Arte Reminiscentiae*, published in Marburg, Germany, 1648. (See http://diglib.hab.de/drucke/202-74-quod-4/start.htm)

[10] Refreshable/programmable braille displays. See, for example, websites http://www.kgs-america.com and/or http://www.metec-ag.de.

[11] B. Berlin, and P. Kay, *Basic color terms: Their universality and evolution*. University of California Press, Berkeley, CA, USA, 1969.

[12] S. Coren, L.M. Ward, and J.T. Enns, *Sensation and Perception (Fourth Edition)*. Harcourt Brace & Company, 1994, pp. 487-501.