An Experimental System for Auditory Image Representations

Peter B. L. Meijer

Abstract—**This paper presents an experimental system for the conversion of images into sound patterns. The system was designed to provide auditory image representations within some of the known limitations of the human hearing system, possibly as a step towards the development of a vision substitution device for the blind. The application of an invertible (1-to-1) image-to-sound mapping ensures the preservation of visual information.**

The system implementation involves a pipelined special purpose computer connected to a standard television camera. The time-multiplexed sound representations, resulting from a real-time image-to-sound conversion, represent images up to a resolution of 64 × 64 pixels with 16 gray-tones per pixel. A novel design and the use of standard components have made for a low-cost portable prototype conversion system having a power dissipation suitable for battery operation.

Computerized sampling of the system output and subsequent calculation of the approximate inverse (sound-to-image) mapping provided the first convincing experimental evidence for the preservation of visual information in the sound representations of complicated images. However, the actual resolution obtainable with human perception of these sound representations remains to be evaluated.

I. Introduction

THE arrival of fast and cheap digital electronics and sensory devices opens new pathways to the development of sophisticated equipment to overcome limitations of the human senses. This paper addresses the technical feasibility of replacing human vision by human hearing through equipment that translates images into sounds, which could one day become relevant for the visually impaired. It should be emphasized here that this paper only proves the feasibility of a low-cost system that preserves much visual information in the generated sound patterns. Conclusive results about the actual capability of humans to successfully interpret these sound patterns, and the advantages and disadvantages encountered in practical daily use, are beyond the scope of this paper, and still require much investigation. Even though these points are not evaluated in this paper, anticipations about several of their aspects need to be considered during the early design phases of any prototype system, in order to increase the probability of approaching a practical solution. In the following sections, some of these considerations are discussed, partly based on several overviews and discussions of previous work that can be found in the references [4], [9], [10], [14], [17].

A successful vision substitution system would have to meet many rather stringent requirements, which, taken together, may seem insurmountable. The system must be cheap, portable, and low-power, while taking into account the limited capabilities of a human user and the typical characteristics of his dynamic environment. If the system interferes with other senses, it should be well worth it, by offering an easy-to-learn high resolution mapping of the visual environment. The system must also operate in real-time, giving an up-to-date representation of a changing environment, which is essential under mobility conditions. Still other requirements would become relevant at later stages of system development. Additional ergonomic aspects, including aesthetics, must be dealt with in order to obtain final acceptance by the visually impaired user. Finally, it should be noticed that visual impairment can arise due to many different causes, including many kinds of brain damage. The nature of malfunctioning can have great consequences for the options for a vision substitution system. For simplicity and clarity of discussion, it is assumed here that we need only consider the malfunctioning of the eyes, or their direct neural connections to the brain, as the single cause of impairment. In that case, the problem of finding alternatives to normal vision reduces to the problem of finding other ways of accessing the healthy brain and providing it with useful visual information.

II. Image Mapping Options

In principle, direct cortical stimulation [7], [20] would appear to be the best approach, as it need not interfere with other senses. However, the huge interfacing problems and invasive nature of the approach, together with the low resolution results obtained up to now, probably make it a long-term research item.

As an immediate second choice, when considering interference, a vibrotactile or electrocutaneous skin-stimulating system seems the most adequate because much of the skin normally plays only a subordinate role as an objective communication channel; e.g., under mobility conditions [1], [12], [16], [19], [21], [22]. That very same aspect might indirectly also be its major disadvantage because there is no strong indication yet, nor an evolution-

Manuscript received August 27, 1990; revised April 30, 1991.

The author is with Philips Research Laboratories, 5600 JA Eindhoven, The Netherlands.

IEEE Log Number 9104907.

ary basis, to expect the existence of all the biological "hardware" for a high-bandwidth communication channel, not just at the skin interface, but all the way up to the cognitive brain centres. In order to save on neural wiring, nature could have provided us with coding tricks that lead to a reasonable resolution when discerning a few skin positions, but that lead to severe information loss when stimulating a large matrix of skin positions. Extrapolation of local resolution experiments to global resolution estimates is risky. Pending decisive experimental results, one should not hold these considerations, which are debatable themselves, as an argument against the interesting work on tactile representations, but as an argument to concurrently investigate the merits of other sensory representations.

Therefore, knowing the importance of bandwidth in communicating data, in our case detailed environmental information, an obvious alternative is to exploit the capabilities of the human hearing system, which is the choice made in this paper. Although we cannot claim that this is the best possible choice, it *is* known that the human hearing system is quite capable of learning to process and interpret extremely complicated and rapidly changing (sound) patterns, such as speech or music in a noisy environment. The available effective bandwidth, on the order of 10 kHz, could correspond to a channel capacity of many thousands of bits per second. Yet again there may be restrictions, imposed by nonlinearities in the mechanics of the cochlea and the details of information encoding in the neural architecture [8]. In spite of these uncertainties, the known capabilities of the human hearing system in learning and understanding complicated acoustical patterns provided the basic motivation for developing an image-to-sound mapping system. For experimental work under laboratory conditions, the interference with normal hearing need not bother us yet. Moreover, it is not necessarily an insurmountable problem, provided the system does not block the hearing system to such an extent, that it becomes impossible to use it for speech recognition or to perceive warning signals. If the system only blocks subtle clues about the environment, normally perceived and used only by blind persons, while at the same time replacing those by much more accurate and reliable "messages," it will be worth it. As an example of a subtle clue, one may think of the changes in the perceived sound of one's footsteps, when walking past a doorway.

Concerning the input from the environment, true visual (camera) input has some major advantages over the use of sonar beam reflection patterns. The possibilities of sonar have been investigated rather extensively in the past [3], [5], [15], [23]. The short range of sonar makes it impossible to perceive far-away objects; e.g., buildings, which is essential for mobility guidance. Furthermore, visual input matches some of the most important information providers in our surroundings, such as road signs and markers. Contrary to sonar, visual input could also provide access to other important information sources like magazines and television. Technically, it is difficult to obtain unambiguous high resolution input using a scanning sonar, while any commercially available low-cost camera will do. It should be added, however, that this statement only holds if the camera input is supplemented by depth-clues through changes in perspective, to resolve ambiguities in distance. (Relative) depth information can be derived from evolving relative positions of the viewer and his environment, combined with knowledge about the typical real size of recognized objects, as one readily verifies by looking with one eye covered. Therefore, an equivalent of binocular vision is no strict prerequisite, and doubling of the more expensive hardware parts (such as the camera) can be avoided. Nevertheless, it should be noted that the approach described in this paper, presently involving one camera, lends itself quite well to a future binocular extension, by mapping left-eye images to sound patterns for the left ear and right-eye images to sound patterns for the right ear.

In order to increase the image resolution obtainable via an auditory representation, a time-multiplexed mapping is performed to distribute an image in time. A somewhat related method has been applied by Kaczmarek *et al.* [16] in the design of tactile displays. In this paper, however, the two-dimensional spatial brightness map of a visual image is 1-to-1 scanned and transformed into a two-dimensional map of oscillation amplitude as a function of frequency and time. The same basic functionality was considered by Dallas and Erickson [6]. It is likely that a subdivision of only one of the two spatial dimensions of the visual image into individual scan lines will be more amenable to a later mental integration or synthesis than a subdivision in both dimensions into rectangular patches. Only a one-dimensional mental synthesis will then be required for image reconstruction, in contrast to the two-dimensional scanning scheme proposed by Fish [11], and the sequencing scheme used in [16].

The human brain is far superior to most if not all existing computer systems in rapidly extracting relevant information from blurred and noisy, redundant images. Therefore, no attempt was made to deliberately filter the redundancy out of visual images that are being transformed into sound patterns. From a theoretical viewpoint, this means that the available bandwidth is not exploited in an optimal way. However, by keeping the mapping as direct and simple as possible, we reduce the risk of accidentally filtering out important clues. After all, especially a perfect nonredundant sound representation is prone to loss of relevant information in the nonperfect human hearing system. Also, a complicated nonredundant image-to-sound mapping may well be far more difficult to learn and comprehend than a straightforward mapping, while the mapping system would increase in complexity and cost.

III. Image-to-Sound Mapping

In order to achieve high resolution, an image is transformed into a time-multiplexed auditory representation. Whenever the previous, $(k - 1)$th, image-to-sound con-

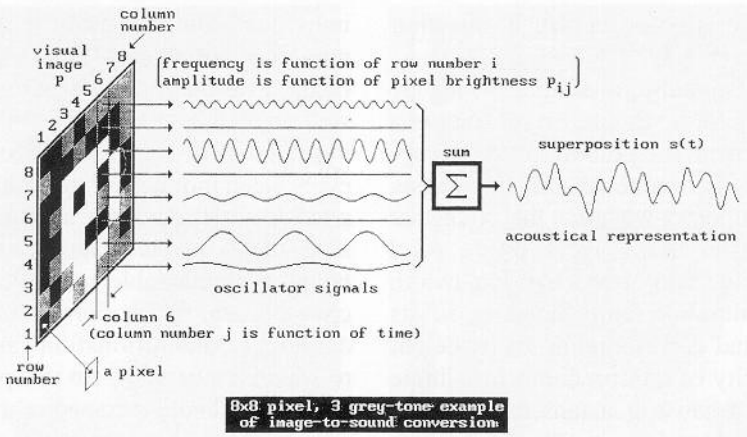


Fig. 1. Principles of the image-to-sound mapping.

version has finished, a new, kth, image is sampled, digitized and stored as an M (height, rows) $\times$ N (width, columns) pixel matrix $\boldsymbol{P}^{(k)}$. This takes τ seconds. During this period, a recognizable synchronization click is generated to mark the beginning of the new image, or, equivalently, the ending of the previous image. The value of any pixel matrix element $p_{ij}^{(k)}$ is one out of G gray-tones, i.e.

$$\boldsymbol{P}^{(k)} = (p_{ij}^{(k)}), \quad p_{ij}^{(k)} \in \{g_1, \cdots, g_G\}$$
$$i = 1, \cdots, M$$
$$j = 1, \cdots, N. \tag{1}$$

Subsequently, the conversion into sound (re)starts, taking one of the N columns at a time, and starting with the leftmost one at $j = 1$. Fig. 1 illustrates the principles of the conversion procedure for the simple example of an 8×8, 3 gray-tone image ($M = N = 8$, $G = 3$). The mapping translates, for each pixel, vertical position into frequency, horizontal position into time-after-click, and brightness into oscillation amplitude. It takes T seconds to convert the whole, N-column, pixel matrix into sound. For a given column j, every pixel in this column is used to excite an associated sinusoidal oscillator in the audible frequency range. Different sinusoidal oscillators form an orthogonal basis, assuming they have frequencies that are integer multiples of some basic frequency. This ensures that information is preserved when going from a geometrical space to a Hilbert space of sinusoidal oscillators. A pixel at a more elevated position i corresponds to an oscillator of higher frequency f_i. The larger the brightness of a pixel, represented by gray-tone $p_{ij}^{(k)}$, the larger the amplitude (''loudness'') of its associated oscillator. The M oscillator signals for the single column are superimposed, and the corresponding sound pattern $s(t)$ is presented to the ear during T/N seconds. Then the next, $(j + 1)$th, column, is converted into sound: this procedure continues until the rightmost, Nth, column has been converted into sound, T seconds after the start of the conversion. Subsequently, a new pixel matrix $\boldsymbol{P}^{(k+1)}$ is obtained, which again takes τ seconds. Meanwhile, the image separating synchronization click is generated, which is essential for a proper lateral orientation. We should have $\tau \ll T$ to continue as soon as possible with the conversion into sound of a new image. Once a new pixel matrix is stored, the image-to-sound conversion starts all over at the leftmost column. The conversion therefore returns to any particular column every $\tau + T$ seconds.

It must be noted that sinusoidal oscillation of finite duration—here at least T/N seconds—is not sinusoidal oscillation in the strict sense. This has significant consequences for obtainable resolution, as will be discussed in a later section.

Using $\omega_i = 2\pi f_i$, the general transformation can be concisely written as

$$s(t) = \sum_{i=1}^{M} p_{ij}^{(k)} \sin(\omega_i t + \phi_i^{(k)}) \tag{2}$$

during times t that satisfy the condition that some column j of a pixel matrix $\boldsymbol{P}^{(k)}$ is being transformed; i.e., with t, j and k related by

$$t_k + (j-1) \cdot \frac{T}{N} \le t < t_k + j \cdot \frac{T}{N}$$
$$j = 1, \cdots, N; \; k = 1, 2, \cdots \tag{3}$$

where t_k is the starting moment for the transformation of the first column $j = 1$ of $\boldsymbol{P}^{(k)}$. Therefore, if the first image ($k = 1$) was grabbed starting at $t = 0$, we have

$$t_k = (k-1)(\tau + T) + \tau. \tag{4}$$

Additional requirements must be fulfilled for monotonicity (to avoid row permutation) and separability

$$\omega_i > \omega_{i-1}, \qquad i = 2, \cdots, M. \tag{5}$$

According to (2)–(4), $s(t)$ is not defined during the image renewal periods of duration τ, i.e., for t in $t_k - \tau \le t < t_k$. As indicated before, a recognizable synchronization click is generated within each of these periods. The pixel brightnesses are in (2) directly used as amplitude values, which is always possible by using an appropriate scale for measuring brightness. The phases $\phi_i^{(k)}$ are just arbitrary constants during the image-to-sound conversion, but they

may change during the generation of the synchronization clicks.

The sound patterns corresponding to simple shapes are easily imagined. For example, a straight bright line on a dark background, running from the bottom left to the upper right, will sound as a single tone steadily increasing in pitch, until the pitch jumps down after the click that separates subsequent images. Similarly, a bright solid rectangle standing on its side will sound like bandwidth limited noise, having a duration corresponding to its width, and a frequency band corresponding to its height and elevation. The simplicity of interpretation for simple shapes is very important, because it means that there is no major motivation diminishing learning barrier that has to be overcome before one begins to understand the new image representation. More realistic images yield more complicated sound patterns, but learning more complicated patterns can be gradual and, because of the simplicity of the *mapping*, accessible to conscious analysis. Of course, in the end the interpretation should become "natural" and "automatic" (i.e., largely subconscious) for reasons of speed.

It is worth noticing that the above image-to-sound mapping does not exploit the ability of the human hearing system to detect time delays (causing phase shifts) between sounds arriving at both ears. Because this is the natural way of detecting sound source directions, the perception of the jth column could be further supported by providing a time delay of the sound output. Using a monotonically increasing delay function $d(\cdot)$ of column position, with $d(0) = 0$, one can take $d(j-1)$ for the left ear and $d(N-j)$ for the right ear. Although CCD's allow for a straightforward implementation, a delay function has not been implemented, because the time-after-click method seemed sufficient and could be realized without additional hardware cost.

In the following section, the values for M, N, G, and T will be discussed. In order to prevent (significant) loss of information, a (virtually) bijective—i.e., invertible—mapping must be ensured through proper values for these parameters.

IV. Basic Design Considerations

Communicating $M \times N$ brightness values, each one taken from a set of G possible values, within T seconds through a communication channel, amounts to sending data at a rate of I bits per second, where I is given by

$$I = \frac{MN \cdot {}^2\log(G)}{T}. \tag{6}$$

Several restrictions to the values of M, N, G, and T must be considered to prevent an unacceptable loss of information within the communication channel, which in this case includes the human hearing system that has to deal with the output of the image-to-sound mapping system.

In the transformation described by (1)–(5), the G allowed brightness values correspond to the amplitudes of individual components of a superposition of—in principle—periodic signals. The amplitude values convey the image information. The periodic signals themselves, as well as their superposition, may have a large, even infinite, number of signal levels for a given amplitude. However, when monitoring the amplitudes of the Fourier components, as the human hearing system does to some extent, the G levels reappear. An estimate for the upper bound for obtainable resolution can be derived from the crosstalk among Fourier components, caused by the finite duration of the information carrying signals in the image-to-sound conversion. (In this paper, the term "crosstalk" is used to denote the spectral contribution, or leakage, to a particular frequency that is caused by limiting the duration of a sinusoidal oscillation of another frequency.) To facilitate the analysis, a very simple image sequence will be considered: all images are dark, except for one, the k'th image, in which only a single pixel (i', j') is bright; i.e. $p_{ij}^{(k)} = 0$, $\forall i, j, k \backslash \{i = i', j = j', k = k'\}$. By dropping the accents, (2) now becomes equivalent to

$$s(t) = \begin{cases} p_{ij}^{(k)} \sin(\omega_i t), & t^{(1)} \le t < t^{(2)} \\ 0, & \text{otherwise} \end{cases}$$
$$\text{with } t^{(2)} - t^{(1)} = \frac{T}{N}. \tag{7}$$

When neglecting the contributions of the horizontal synchronization clicks, one can derive, e.g., via Fourier transformation, that to avoid significant crosstalk with sinusoidal signals of duration T/N s, the frequency separation Δf should be on the order of $2/(T/N)$ Hz. In that case, it can be shown that the oscillator frequency associated with a vertically neighboring pixel on row $i \pm 1$ already lies beyond the maximum of the first spectral side lobe corresponding to the pixel on row i. For an equidistant frequency distribution we would have $\Delta f = B/(M-1)$, using a bandwidth of B Hz, such that the crosstalk restricting criterion becomes

$$(M-1)N \le \frac{B \cdot T}{2}. \tag{8}$$

In our application, using sequenced "frozen" images, the allowed conversion time T is restricted by the fact that the information content of an image soon becomes obsolete within a changing environment. Furthermore, the capacity of the human brain to assimilate and interpret information distributed in time is also limited. One may expect that biological systems are well-adapted to the processing and anticipation of major environmental changes to an extent that corresponds to the typical time constants of these changes and the time constants of physical response. For humans, although having a minimum response time of a few tenths of a second, most events, significant motorial responses, etc., take place on the time scale of a few seconds. One may think of events like a door being opened, or a cup of coffee being picked up. For longer time scales, humans have the option to think

and contemplate an evolving situation, but the brain still seems highly dedicated and optimized towards the time scale of seconds, as when dealing with, for example, speech or mobility. Therefore, we should have a conversion time T on the order of one second or less. Similarly, we should have $\tau << 1$ s, to avoid blurred images, as well as to have most of the time available for sending image information through the auditory channel. Furthermore, the total available bandwidth of the human hearing system amounts to some 20 kHz, but the useful bandwidth B of the human hearing system is not much more than 5 or 6 kHz. Above these frequencies, the sensitivity drops rapidly, especially for elderly people. Therefore, taking $B = 5$ kHz and $T = 1$ s, we find from (8) for a square matrix $M = N$ that $M \leq 50$.

It should be noted that obtainable resolution also depends on the particular image being transformed. The criterion (8) for obtainable resolution holds for situations with a few bright pixels per column on a dark background. The strongest crosstalk occurs to the nearest neighbors. However, many bright pixels positioned at larger distances could together also contribute significantly. Therefore, the obtainable resolution for detecting a small bright object against a dark background will be somewhat higher than for a small dark object against a bright background.

Criterion (8) is the consequence of an uncertainty relation between frequency and time, see also [18]. Yet even for signals of long duration, the ability of humans to distinguish different sound spectra is known to be limited. It must be verified that the separability of some 50 frequency components, estimated from uncertainty principles, does not (significantly) exceed other known limitations of human perception. The limits of frequency discrimination in human auditory perception are highly dependent on the particular sound patterns involved in the experiments. From the width of the critical bands of the human hearing system one would predict the separability of at most several tens (20 to 30) of components [13], [24]. On the other hand, the difference limen or just noticeable difference (jnd), determined for a single sinusoidal tone with slowly varying frequency, would suggest the separability of hundreds of components [24]. The image-to-sound mapping can, in principle, and dependent on the particular artificial image being transformed, yield almost any sound pattern used in the literature for frequency discrimination experiments (as long as phase relations are neglected). Moreover, it is virtually impossible to characterize what constitutes a "normal" image, and thereby characterize the corresponding sound patterns. Therefore, it seems reasonable that our present resolution estimate with 50 frequencies is not much more optimistic than expected from measurements of critical bands. Lowering the number of frequencies to the number suggested by critical bands would have disadvantages. It would enhance the undesirable perception of discontinuities in frequencies obtained with images containing, for example, widely separated continuous slanting bright lines on a dark background. In such cases the difference limen becomes the relevant limit for perception.

V. Design and Implementation

A prototype system has been designed and built for $M = N = 64$ and $G = 16$. The use of 16 gray tones ensures sufficient information preservation in shading, while causing smooth changes in brightness to be perceived as smooth changes in loudness. The latter is important to avoid hearing distracting artificial discretization boundaries within, for example, near-uniform image backgrounds. The system design involved a 64×64 pixel matrix instead of a 50×50 matrix, because powers of 2 are more convenient for a digital implementation. In order to be able to improve the quality of the sound representations in less time-critical situations (e.g., static images), the design also included a user-switchable 1.05 s or 2.10 s conversion time T. According to (6), we find a system throughput of up to 16 kb/s for $T = 1$ s, and 8 kb/s for $T = 2$ s.

The $M = 64$ sinusoidal oscillators have not been implemented as physically distinguishable analog components, as was the basis of [6]. Instead, the behavior of the oscillators is emulated via fast digital computations. In this way, the implementation can be made much more compact (portable), more robust (against detuning) and more flexible (programmable), but also cheaper and lower in power dissipation (for battery operation), than a large set of independent and precisely-tuned analog oscillators could. In [6], the tuning problem was treated by using digital frequency dividers, but the other aspects remained unsolved due to the 64-fold hardware involved in signal integration, modulation, and summation. Fig. 2 depicts schematically how a sound sample is calculated for M oscillators, by updating the phase ϕ_i of oscillator i with a phase step $\Delta\phi_i$, calculating the sine of the resulting phase, scaled by the brightness $p_{ij}^{(k)}$ of the pixel at vertical position (row) i, and adding the result to a superposition accumulator. The phases, phase steps, scaled sine values, and pixel brightnesses all reside in memory modules. Thereby, expensive hardware multipliers or sine evaluators could be avoided, while increasing the flexibility for the implementation, via programming, of alternative mappings. The frequencies of all 64 individual oscillators, determining the applied bandwidth, are programmable. This provides sufficient flexibility for further optimization of the system. It may also be used to take care of individual differences—and malfunctions—in hearing among users, by using a personalized frequency distribution. In the present prototype, a 16K-word $\Delta\phi_i$ memory module allows for 256 different and independent frequency distributions for the 64 oscillators, without reprogramming.

To reduce system cost further, only commercially available components of standard speed were used. Apart from CMOS memory components, most of the system logic was implemented using standard LS TTL technol-

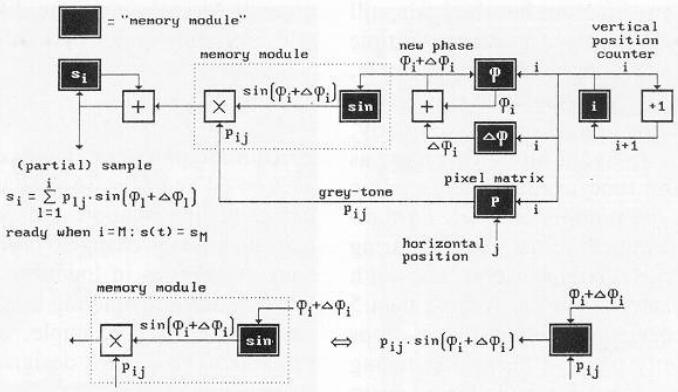


Fig. 2. Principles of the sound sample synthesis.

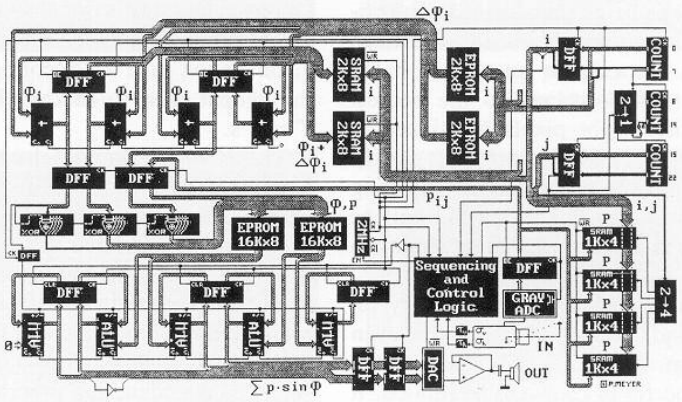


Fig. 3. The pipelined system architecture.

ogy. Using a system clock of $F = 2$ MHz, a serial computation of samples of the superposition of oscillator signals can take place at a frequency of $F/M = 31.25$ kHz, which is sufficiently high for very good audio quality (cf. CD players using 44.1 kHz). The superposition samples are represented by 16-b values, again to achieve high audio quality. The proper processing of two million pixel oscillator samples per second at a 2 MHz system clock requires significant parallelism. Because the transformation itself is fixed, a pipelined design was made to provide the required data flow. Several oscillator contributions are being processed simultaneously, but in physically different stages of the necessary set of operations. For example, the phase step value of oscillator $i + 1$ is read from memory at the same time as the scaled sine value belonging to oscillator i (of which the phase step value has been read 500 ns earlier). The resulting special purpose computer is optimized towards the image-to-sound conversion. The whole conversion system, including 20 ms frame grabbing and 16 gray-tone digitization hardware for input, and the analog output stages for the headphones, has been implemented on a single 236 × 160 mm circuit board, dissipating a measured 4.4 W. In addition, a commercial 2.7 W Philips 12TX3412 vidicon observation camera is presently used for input. This standard 625-line PAL camera delivers 312- and 313-line interlaced images every 20 ms, of which only 64 lines are used for later conversion, by neglecting three out of every four lines in a centered subset of 256 lines. The digitization hardware applies Gray code encoding in a 4-b flash AD-converter, and includes a sample-and-hold circuit for extremely rapid video signal transitions. The sampling, digitization and storage of a 64 × 64 pixel image takes place within a single 20 ms frame time, thereby avoiding blurred images. Because the camera frame generation runs independently, vertical and horizontal (delayed) synchronization signals are used to synchronize the conversion system with the camera. Due to the required synchronization, the conversion system may first have to wait for a period lasting less than 20 ms, before the 20 ms frame grabbing of a new camera image can actually start. Still, 20 ms $\leq \tau <$ 40 ms easily satisfies $\tau << T$. The user only notices a characteristic synchronization click, as needed to delimit subsequent sound patterns. Fig. 3 illustrates the system architecture with a more detailed overview of the system design, close to the component level. The analog stages for input (camera IN) and output (headphones OUT) are only indicated symbolically. The drawn width of data and address buses is proportional to the number of bits. The two oversized 2K × 8 SRAM's for the 64 16-b ϕ_i values were used only

because discrete 64 × 16 b memories are non-standard, and consequently far more expensive than these larger memory chips. Using divisions of the system clock by 2^{21} or 2^{22}, T can at any moment be switched by the user between a 1.05 or a 2.10 s image-to-sound conversion time.

The proper width of the data and address buses has to be selected very carefully. For example, truncation of digital phase values to a smaller number of bits leads to a kind of frequency noise. The bandwidth of this noise should be much less than the frequency differences between neighbouring oscillator frequencies. In the present design this condition is met by a frequency noise less than $F/2^{18} \approx 7.6$ Hz, due to a division of F by $M = 64 = 2^6$ to arrive at the sound sample frequency, and an additional division by 2^{12} corresponding to the truncated 12-b phase values used for calculating sine values within one whole sine period. The truncation to 12 b can lead to at most a $1/2^{12}$ period phase step error when going from one sound sample to the next. The different oscillator frequencies are freely programmable integer multiples of $F/2^{22} \approx 0.5$ Hz, because the phase values are updated and stored as 16-b values before their truncation to 12 b. This ensures the assumed orthogonality of the oscillators, and provides a very fine scale for selecting and programming a particular set of frequency distributions.

For the purpose of experimentation, an observation monitor is presently an integral part of the experimental setup, to provide visual feedback to the sighted experimenter or teacher on the field of view and on the quality of the camera image, concerning contrast and definition.

VI. Results

Before experimenting with human perception, it is worthwhile to objectively evaluate the performance of the image-to-sound conversion system itself. In order to verify the correct functionality of the system, and to prove that the sound patterns still contain the expected image resolution, an inverse mapping was applied. The filtered analog signal, arising from the 31.25 kHz 16-b DA-conversion, was itself sampled, digitized and stored for off-line computer processing. The sound sampling was done with 8 b resolution at a 20 kHz sampling rate. Subsequently, Fourier transforms were applied using moving (overlapping) rectangular time windows of width $T/64$ s. No image enhancement techniques were applied to improve the perceived quality of the spectrograms. The spectrograms were directly obtained from the foregoing calculations by representing time by horizontal position, frequency (of the Fourier component) by vertical position, and amplitude (for the Fourier component) by gray tone. Figs. 4–11 show CRT screen photographs using a 192 × 192 pixel matrix (i.e., 192 Fourier components, 192 time window positions). The positions of frequencies corresponding to the $\Delta\phi_i$ set, as programmed into the EPROM's, are indicated by small dots along the left sides of images. The image separating click is visible as a somewhat blurry vertical line on the (lower) left side of each image, within the artificial bounding frame. In spite of the lower-quality resampling, the photographs confirm the basic expectations. The restoration of images proves that their content was indeed preserved in the image-to-sound conversion.

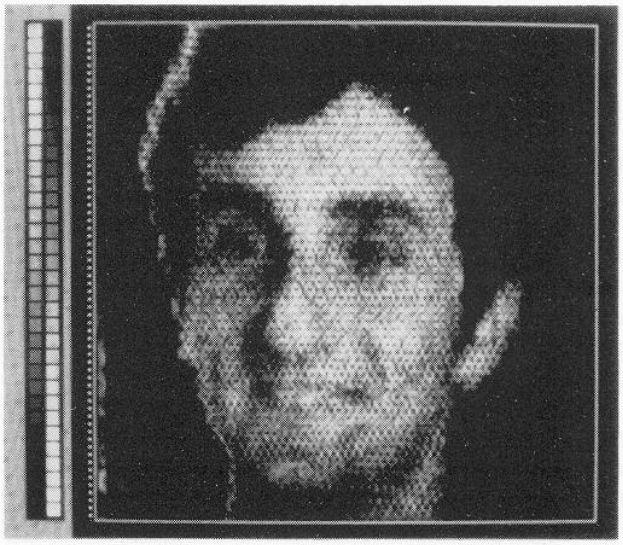


Fig. 4. Spectrogram of a sampled sound pattern for a human face, using a linear frequency distribution and a 1.05 s image-to-sound conversion time.

The resolution estimates discussed before were based on the assumption of an equidistant frequency distribution, as given by

$$f_i = f_l + \frac{i-1}{M-1} \cdot (f_h - f_l) \qquad i = 1, \cdots, M \tag{9}$$

where f_l and f_h are the lowest and highest frequency, respectively. In the experiments described here, the parameters are $f_l = 500$ Hz, $f_h = 5$ kHz, and $M = N = 64$. Figs. 4 and 5 show that with the 64 × 64 matrix, neighboring frequencies are barely separable for a $T = 1.05$ second image-to-sound conversion time, which is as intended and expected, because the design was meant to push towards the theoretical limits. Using a $T = 2.10$ second conversion time, subsequent frequencies are clearly separable, as shown by Figs. 6 and 7.

Although the resolution estimates were partly based on the assumption of an equidistant frequency distribution, experiments were also done using an exponential distribution, given by

$$f_i = \left(\frac{f_h}{f_l}\right)^{(i-1)/(M-1)} \cdot f_l \qquad i = 1, \cdots, M. \tag{10}$$

With the exponential distribution, subsequent frequencies differ by a constant factor instead of a constant term. An equidistant distribution does not match the properties of the human hearing system very well, because the ear is more sensitive to changes in the frequency of a single sin-

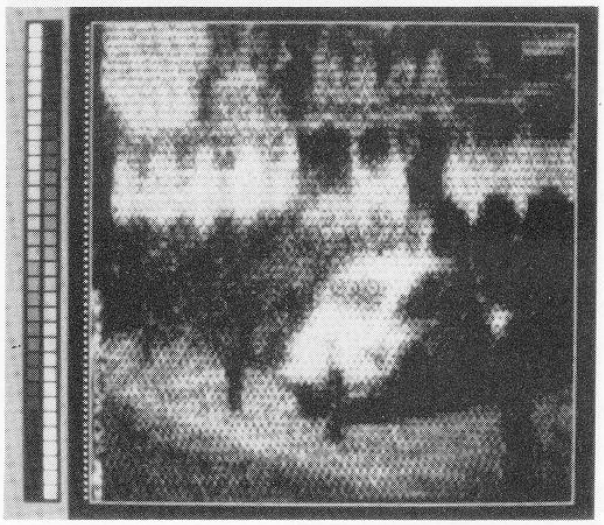


Fig. 5. Spectrogram of a sampled sound pattern for a parked car, using a linear frequency distribution and a 1.05 s image-to-sound conversion time.

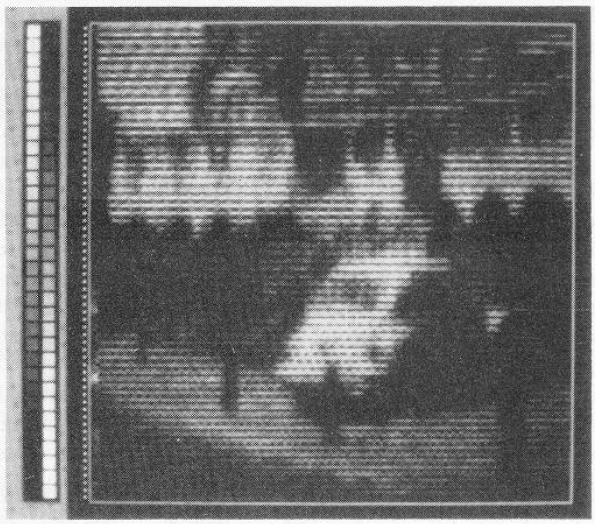


Fig. 7. Scene of Fig. 5, but using a doubled, 2.10 s, conversion time with the linear frequency distribution.

Fig. 6. Scene of Fig. 4, but using a doubled, 2.10 s, conversion time with the linear frequency distribution.

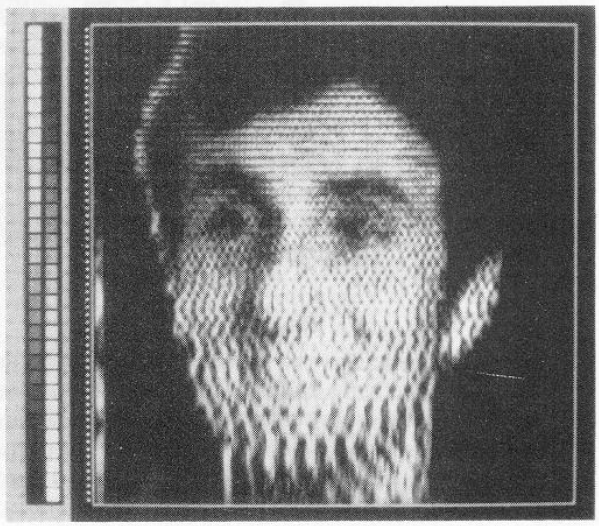


Fig. 8. Scene of Fig. 4, but using an exponential frequency distribution with the 1.05 s conversion time. The vertical axis is logarithmically compressed.

usoidal tone at lower frequencies. The same applies to the discrimination of multiple simultaneously sounding sinusoidal tones. In fact, an exponential frequency distribution maps to an approximately linear spatial distribution of excitation along the basilar membrane in the human cochlea [2]. Subjectively, an exponential distribution is perceived as approximately equidistant, and it is the basis of the tempered musical scale, which has become a de facto standard scale in music. Another reason for abandoning the equidistant frequency distribution was the observation of a rather dominating and hence annoying $(f_h - f_l)/(M - 1)$ Hz tone, as might be expected from the cross-talk components of the image-to-sound conversion and the nonlinearities of the human hearing system (known to give rise to combination tones). Experiments with linearly increasing frequency differences gave little improvement in this respect. However, no annoying spurious sounds were heard using the exponential frequency mapping. Because the frequency differences at the lower end of the spectrum are now smaller than with the equidistant scheme, a resolution less than required for preserving a 64×64 pixel image with $T = 1.05$ s should be expected at the bottom part of reconstructed images. This is indeed observed from the Moiré-like distortion patterns in the bottom parts of Figs. 8 and 9. These patterns are due to the cross-talk among neighboring frequencies. The vertical (frequency) axis was logarithmically compressed in these spectrograms to compensate for the exponential frequency distribution. As expected, the loss of resolution at the lower end of the spectrum, corresponding to the

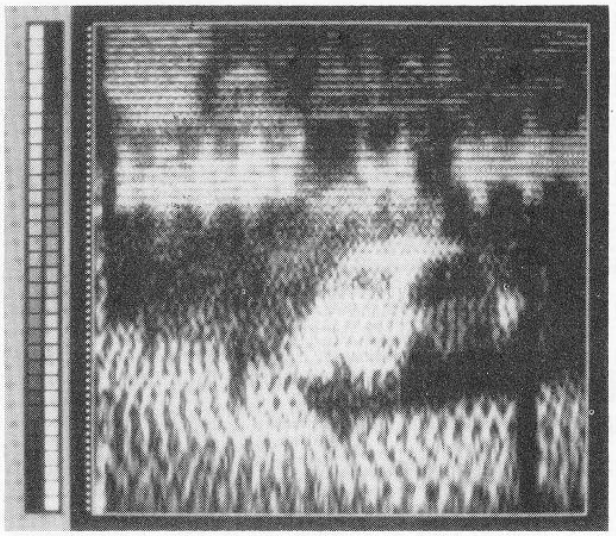


Fig. 9. Scene of Fig. 5, but using an exponential frequency distribution with the 1.05 s conversion time. The vertical axis is logarithmically compressed.

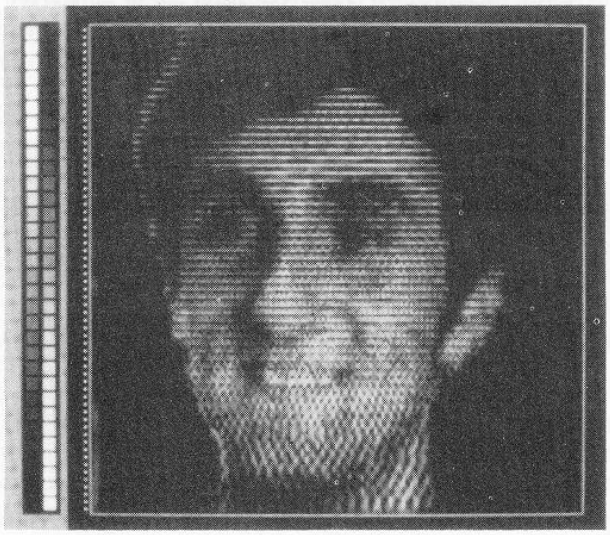


Fig. 10. Scene of Fig. 4, but using an exponential frequency distribution and a doubled, 2.10 s, conversion time. The vertical axis is logarithmically compressed.

lower part of the spectrograms, is most noticeable for $T = 1.05$ s, and much less for $T = 2.10$ s, as is observed by comparing Figs. 8 and 9 with Figs. 10 and 11, respectively. In spite of the partial loss of resolution with the exponential distribution, the gain in perceived equidistance and loss of spurious mixing sounds may appear to be more important in practice. Moreover, the situation is in this respect somewhat comparable to normal vision, in which the human retina has a region of high resolution (the central fovea), while other regions provide much less resolution. As with the eye, one can adapt the camera orientation for a higher resolution at items of interest. In all cases shown, the image-to-sound conversion is obviously

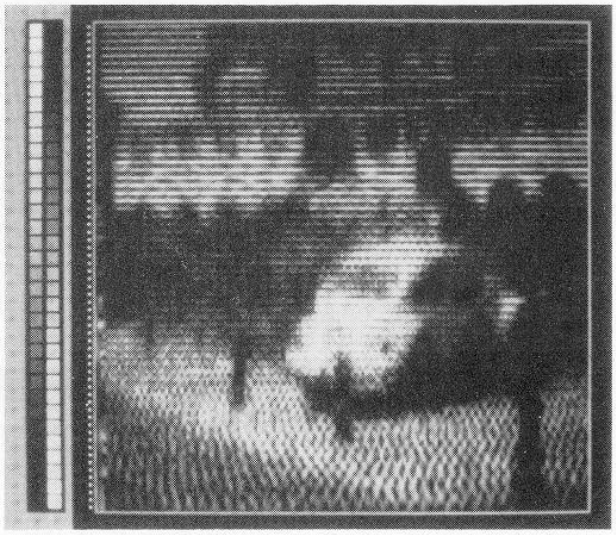


Fig. 11. Scene of Fig. 5, but using an exponential frequency distribution and a doubled, 2.10 s, conversion time. The vertical axis is logarithmically compressed.

of sufficient quality to give a useful account of the visual environment.

VII. Conclusions

A new system has been presented for the conversion of images into sound patterns. The hardware implementation employs a special pipelined architecture for the real-time digital calculation of sound samples. The design aimed at obtaining a portable, low-power, and low-cost system. To this purpose, the prototype system was constructed entirely out of standard, commercially available components. The technical feasibility of the approach has been proven by the construction of a fully functional system, which, by means of inverse mappings, was shown to preserve image information up to a resolution corresponding to estimates for achievable resolution. Some practical parameter restrictions were derived from known limitations of the human hearing system. The resolution provided by the system itself appears sufficiently high to deal with many practical situations normally requiring human vision, as the system was designed towards possible later use by the visually impaired.

However, the further development towards a practical application of the system still awaits a thorough evaluation, with blind persons. This is needed to determine and quantify the limits of attainable image perception, before any firm conclusions about the usefulness of the system can be drawn, because only the technical feasibility has now been established. Such an evaluation should involve both (young) children and adults, while distinguishing the congenitally blind from the late-blinded, because neural adaptability may drop rapidly with age, and the neural development may strongly depend on prior visual experiences.

ACKNOWLEDGMENT

The author would like to thank P. A. Kuppen for preparing the screen photographs for this paper, and H. E. M. Mélotte and L. F. Willems from the Eindhoven Institute for Perception Research for providing initial spectrographic evidence for the correct functionality of the image-to-sound conversion system.

REFERENCES

[1] P. Bach-Y-Rita, C. C. Collins, F. A. Saunders, B. White, and L. Scadden, "Vision substitution by tactile image projection," *Nature*, vol. 221 pp. 963–964, Mar. 8, 1969.

[2] G. von Békésy, *Experiments in Hearing*. New York: McGraw-Hill, 1960.

[3] T. Bower, "Blind babies see with their ears," *New Scientist*, pp. 255–257, Feb. 7, 1977.

[4] J. Brabyn, "Developments in electronic aids for the blind and visually impaired," *IEEE Eng. Med. Biol. Mag.*, pp. 33–37, Dec. 1985.

[5] M. Brissaud and G. Grange, "Système ultrasonores d'aide à la localisation pour non-voyants," *Acustica*, vol. 66, pp. 53–55, 1988.

[6] S. A. Dallas and A. J. Erickson, "Sound pattern generator," Patent: Int. Publ. No. WO 82/00395, Int. Appl. No. PCT/US81/00847.

[7] W. H. Dobelle, M. G. Mladejovsky, and J. P. Girvin, "Artificial vision for the blind: Electrical stimulation of visual cortex offers hope for a functional prosthesis," *Science*, vol. 183, pp. 440–444, Feb. 1, 1974.

[8] H. Duifhuis, "Current developments in peripheral auditory frequency analysis," in *Working Models of Human Perception*, B. A. G. Elsendoorn and H. Bouma, Eds. London: Academic 1989, pp. 59–65.

[9] S. Ferguson and S. D. Ferguson, "High resolution vision prosthesis systems: Research after 15 years," *J. Visual Impairment Blindness*, pp. 523–527, Jan. 1986.

[10] R. M. Fish, "Visual substitution systems: Control and information processing considerations," *The New Outlook for the Blind*, vol. 69, pp. 300–304, Sept. 1975.

[11] ——, "An audio display for the blind," *IEEE Trans. Biomed. Eng.*, vol. BME-23, pp. 144–154, Mar. 1976.

[12] S. F. Frisken-Gibson, P. Bach-Y-Rita, W. J. Tompkins, and J. G. Webster, "A 64-solenoid, four-level fingertip search display for the blind," *IEEE Trans. Biomed. Eng.*, vol. BME-34, pp. 963–965, Dec. 1987.

[13] D. D. Greenwood, "Critical bandwidth and the frequency coordinates of the basilar membrane," *J. Acoust. Soc. Amer.* vol. 33, pp. 1344–1356, Oct. 1961.

[14] A. D. Heyes, "Human navigation by sound," *Phys. Technol.*, vol. 14, pp. 68–76, 1983.

[15] T. Heyes, "Sonic pathfinder," *Wireless World*, pp. 26–29, 62, Apr. 1984.

[16] K. Kaczmarek, P. Bach-Y-Rita, W. J. Tompkins, and J. G. Webster, "A tactile vision-substitution system for the blind: computer-controlled partial image sequencing," *IEEE Trans. Biomed. Eng.*, vol. BME-32, pp. 602–608, Aug. 1985.

[17] L. Kay, "Electronic aids for blind persons: an interdisciplinary subject," *IEE Proc.*, vol. 131, pp. 559–576, Sept. 1984.

[18] V. Majerník and J. Kalužný, "On the auditory uncertainty relations," *Acustica*, vol. 43, pp. 132–146, 1979.

[19] F. A. Saunders, "Information transmission across the skin: High-resolution sensory aids for the deaf and the blind," *Int. J. Neurosci.*, vol. 19, pp. 21–28, 1983.

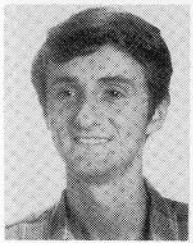
[20] K. Smith, "Seeing with microcircuits," *New Scientist*, pp. 258–260, Feb. 7, 1977.

[21] A. Y. J. Szeto and F. A. Saunders, "Electrocutaneous stimulation for sensory communication in rehabilitation engineering," *IEEE Trans. Biomed. Eng.*, vol. BME-29, pp. 300–308, Apr. 1982.

[22] S. Tachi, K. Tanie, K. Komoriya, and M. Abe, "Electrocutaneous communication in a guide dog robot (MELDOG)," *IEEE Trans. Biomed. Eng.*, vol. BME-32, pp. 461–469, July 1985.

[23] Y. Yonezawa, "Binaural sensitivity to direction cue in an acoustic spatial sensor," *Acustica*, vol. 61, pp. 140–144, 1986.

[24] E. Zwicker, G. Flottorp, and S. S. Stevens, "Critical band width in loudness summation," *J. Acoust. Soc. Amer.*, vol. 29, pp. 548–557, 1957.



Peter B. L. Meijer received the master's degree in physics from the Delft University of Technology, Delft, the Netherlands, in 1985.

Since 1985 he has been with the group CAD for VLSI Circuits at the Philips Research Laboratories in Eindhoven. His current research interests include nonlinear multidimensional modeling, circuit simulation techniques, neural networks, and self-organizing systems.