Scenes and images into sounds: a taxonomy of image sonification methods for mobility applications

PABLO REVUELTA SANZ [1], **BELÉN RUIZ MEZCUA** [1],
(prevuelt@ing.uc3m.es) (bruiz@inf.uc3m.es)

JOSÉ M. SÁNCHEZ PENA [1], **AND BRUCE N. WALKER** [2]
(jmpena@ing.uc3m.es) (bruce.walker@psych.gatech.edu)

[1] *Carlos III University of Madrid, Leganés, Spain*
[2] *GeorgiaTech, Atlanta, U.S.A.*

This paper presents a survey of existing sonification systems used to represent visual scenes, analyzes their characteristics and proposes a taxonomy of this set of algorithms and devices. This classification is non-exclusive, since many parameters of any sonification procedure work as independent variables and can be recombined in any other set. Although many of these algorithms have been proposed in the field of assistive technology, and most of the examples come from that field, we will only focus on auditory aspects, avoiding an analysis in terms of mobility, rehabilitation or subjective perception. We propose two main categories to classify every sonification algorithm, the psychoacoustic and the artificial, and a third one mixing properties of each one of them. We use classic paradigms such as the pitch, piano and point transform, as well as some new subsets. A final summary of 25 different assistive products is given, with their classification following our scheme.

I. INTRODUCTION

Sonification is the systematic representation of data using sounds. Many types of sonification can be found, such as text-to-speech programs (converting text into audible speech), color readers (color into synthetic voice), Geiger counters (radioactivity into clicks), acoustic radars or MIDI synthesizers, etc. [1]. Sonification is, hence, a translation between two essentially different sets of data. It is a very active field, mostly focused on accessibility, but also on aesthetic and artistic goals, psychoacoustic research, improvement in data analysis or hi-tech commercial devices, among others. In this work, we will focus on the sonification of visual scenes, which is usually applied to the field of accessibility for the visually impaired. However, in this paper we will focus on the sonification procedure itself.

Since the information of the visual world presents much more information than the auditory one, there is a loss of information in the sonification process. This means that some parameters of the image will be ignored in the sonification process.

We will use "image sonification procedure" to refer to the set of rules that make an auditory cue correspond to a visual characteristic of the original image. For example, the following rule can be found in [1]: "Loudness is associated with brightness".

The use of sounds to automatically represent the visual world has been proposed since the last years of the XIX century. Specifically, Noiszewski built the Elektroftalm in 1897 [2]. Some years later, in 1912, d'Albe built the Exploring Optophone [1, 3, 4]. Of course, these two approaches were extremely simple: the Elektroftalm used a selenium cell to discriminate whether there was light or not and, then, produced a sound to inform the user. The optophone, in the same way, produced a sound proportional to the light received by the selenium cell. These first two assistive products received criticism from the blind community, because of the uselessness of such simple information. A more complex proposal came, in those years, also from d'Albe, with the Reading optophone [1]. This new system produced different combinations of sounds, representing characters written in a book (see Fig. 1 from [4]). In this figure, the points represent the part of the image being sonified. Thus, the vertical line of the "E" letter produces all the sounds, while the three horizontal lines that comes after that, only produce the first, the third and the fifth sounds (G, D' and G' in the figure). The vertical scanning line advances from left to right.

> This system allowed some very well trained experts to achieve high reading speeds (around 60 words/min), although usually the rate was much lower [1], p. 139.

A final (pre)historic curiosity was the Radioactive guider from the 1940's [5], which used a radioactive element to

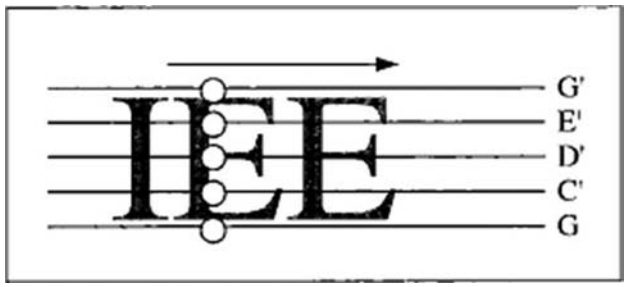

Fig. 1. The d'Albe's Reading optophone [4].

produce a beam that was reflected on the bodies and detected by a Geiger counter, thus, converted into clicks.

Since the late 1970's and early 1980's, technology has allowed for the development of more complex, faster and smaller devices.

II. MATERIALS AND METHODS

We carried out a survey of the scientific and commercial literature related to our topic. This retrieval was performed between September 2010 and June 2011 with the following keywords: "ETA", "Electronic Travel Aid", "Assistive product" and "Assistive technology" together with "blind" and "Mobility". Also, a new search with the keywords "Sonification" and "auditory display" was also implemented.

We have used peer and non peer-reviewed English literature, from the "Web of Science" meta-database and from academic and commercial web pages, technical reports, Ph.D. thesis and catalogues.

From these searches, papers that were relevant were studied for the use of sounds to represent visual scenes, especially for blind wayfinding applications.

III. RESULTS

The sonification proposals found in the literature can be gathered in terms of the parameters used to acoustically describe an image, as well as the number of channels and the dimensional complexity of the sonification produced.

Following Milios' study [6], two modes of sonification functions can be identified:

- Proportional mode, where the audio feature is a function of the instantaneous value of part of the image or a data range. More precisely, the nonlinear function is a decaying exponential mapping. In this case, the function follows the so called weberian law [7], and will be used widely in some paradigms presented herein.
- Derivative mode, where the audio feature is a function of the temporal derivative of range. More precisely, changes in consecutive range measurements (an approximation to mapping the temporal variation of a visual cue) are mapped onto the audio domain. This option, although proposed, is not often used. One of the reasons is that our hearing system (and, with some exceptions, every sensory system) works mostly following the first option, i.e. the weberian law. Another reason is that the global accuracy in the first option is higher than that in the second one, as shown in [6]. However, Milios proposed its laser based assistive product able to work in a derivative mode [6].

We propose, in this work, a classification in terms of the number of channels and of how the sonification correspondence is designed.

The sonification electronic travel aids (ETAs) can be grouped by the number of channels they use to transmit the information:

- Monaural: One single channel is used to transmit the information. This information is usually codified in a continuous dimension, such as frequency, amplitude, etc. Thus, most of them are unidimensional. Some examples of this group of ETAs are the Nottingham OD, Ultrasonic Torch, Mims, FOA Laser cane (all these devices are shown in [3]) or the Sidewalk detector [8].
- Binaural: Binaurality permits the transmission of much richer information to the user. Thus, some more complex proposals can be found in this family. Some of the most important ones are the Navbelt [9], Multi and crossmodal ETA [10, 11], Echolocation [12], EAV [13], 3-D Support [14–16], CASBliP [17, 18], vOICe [19], NAVI [20], Sonic Pathfinder [21–25], SonicGuide [26, 27], etc.

Binaurality allows a more complex and richer understanding of the environment, thus, it has been much more commonly implemented for mobility assistive products.

In both mono and binaural solutions, some choices must be made to determine how the ambient information is encoded. Any decision belongs to one of the two possible sonification paradigms that we will call psychoacoustic and artificial paradigms:

- Psychoacoustic: This paradigm exploits the natural discrimination of the source spatial parameters (distance, azimuth and elevation, for instance). It uses functions and curves to simulate, by means of convolution processes, the virtual position of the source, using the natural localization of the source.
- Artificial: When a sonification process does not use the natural sound location skills of the user, a new match between graphical and auditory cues must be made in an artificial way. This happens for some visual characteristics which have no correlation with the auditory world (such as color or textures, for example). Likewise, in some other projects, the natural location accuracy of sounds was not perceived as good enough and researchers have proposed artificial matching of visual and auditory cues which increase the accuracy, as it will be shown in sub-section B.

In [28], Morris proposes a scheme of transition between the abstraction of the sign and the represented object, as shown in Fig. 2.

But there is yet another possibility to link the sign or symbol and the signified object, namely the so-called

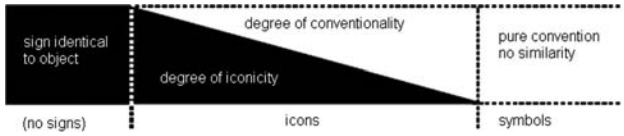

Fig. 2. Degree of conventionality-iconicity of a sign, from [28].

"indexing", which means the representation of something because of a causal link (the typical example is the smoke sign/symbol to represent tabacco). All these terms will be helpful to understand the way that different researchers have represented the world into sounds.

We will now discuss each paradigm in depth. However, there are almost no purely "psychoacoustic" or "artificial" sonification proposals and most of them present a mix between these two paradigms. Just for presentation purposes, we will present the clearest examples of each case, focusing on the most relevant features regarding this taxonomy.

Before this analysis, one could ask what should the brightness in the original image represent, and here, as before, different options can be found:

- Light intensity: This is the first implementation provided by researchers. This is also the case of the first electronic assistive products developed, as it was seen with the optophone or the elektroftalm. In the literature, this is called a direct mapping [29]. This option has been widely used in modern ETAs as well. This is the case of the vOICe [19], the proportional sonification proposed by Milios et al. [6]. Here problems already present in the reading optophone can be found. If the image is mostly white or bright, the sound pattern is noisy and can hardly be understood. Moreover, the light intensity is not related at all with the significance of the information. For example, the sky is bright, but completely irrelevant for mobility aspects. Another problem is that it is impossible to estimate the distance of an object, so it hardly serves for mobility purposes.
- Depth: An alternative for directly representing the image captured by cameras into sounds, is to represent the depth. In this case, the closer the objects are, the higher the sound is. This is widely implemented, since the importance of this property for the user is obvious. The SVETA [29], for example, implements this approach and, specifically, the musical octave sonification of 2.5D images. The so-called 2.5D images are planar images (2D matrix) in which the intensity of the pixels represents depth information. The most often used ones, due to ease of objective comparisons, are gray scale images in which the brighter the pixel, the closer the point. An important problem of this approach is the loss of planar information, such as text written on a flat surface.
- Edges: To make the system able to recognize information or patterns over flat surfaces, an edge representation has also been proposed. The advantages over the depth map, but also its limitations, can be perceived in Fig. 3.

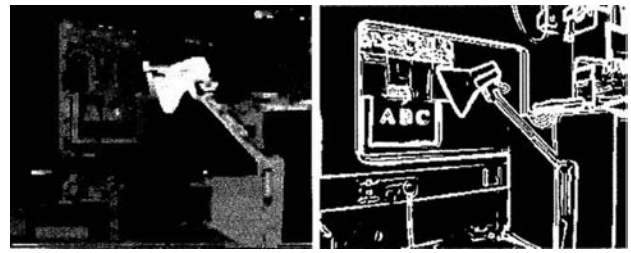

Fig. 3. (Left) Depth map representation (with edge pre-filtering) and (right) edge representation [1].

A. Psychoacoustic sonification

The psychoacoustic paradigm uses, as already mentioned, the natural effects in the hearing system to help locate the source positions.

Rossi *et al.* [30] propose a classification of psychoacoustic implementations into 6 levels based on their sound immersion level scale, presented in Table 1.

Most modern ETAs implement acoustical environments at least at level 2 (with the exception of the laser or ultrasound pointing devices). The most advanced ones do so even at levels 4 or 5. Such is the case of Bujacz's proposed ETA [31].

In these cases, the user perceives different sounds as if they come from different spatial situations. Thus, these systems implement several convolutions for each sound source to virtually place it in the correct solid angle and distance.

The main advantage of this approach is, obviously, the natural localization of the sound sources. Although the authors state that no training is needed at all, and that the system can be used by untrained users from the beginning, this statement can be debated, given that the use of sounds to perceive spatial structures needs to produce new neural networks by means of training (see, for example, [32]).

Table 1. Rossi's acoustical immersion levels [30].

Level	Techniques/Methods	Perceptions (results)
0	Monoaural "dry" signal	No immersion
1	Reverberating, echoes	Spaciousness ambience
2	Panning (between speakers), stereo, 5.1	Direction movement
3	Amplitude panning, VBAP	Correct positioning in limited regions
4	HRTF (Head related Transfer Function), periphony (ambisonics, WFS. . .)	Stable 2D sound fields
5	HRTF, periphony (ambisonics, WFS. . .)	Stable 3D sound fields, accurate distance and localization

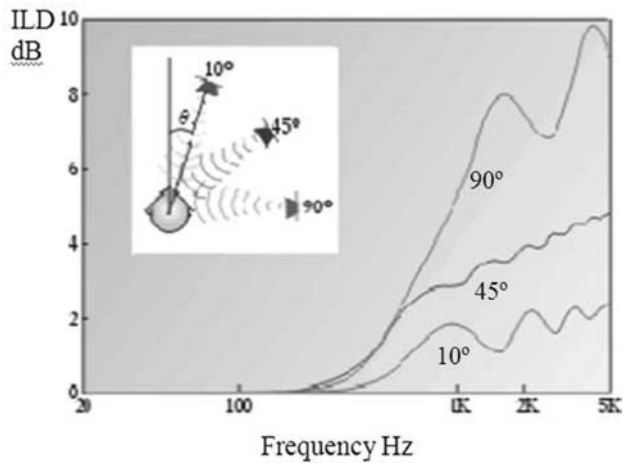

Fig. 4. Loudness-angle relation [33]. Vertical axis represents the IDL (Interaural Level Difference) in dB, i.e. the difference of loudness received in both ears. Although it is dependent of the frequency, the mappings exploiting this will generate a simpler curve to implement the panning (such as the MIDI protocol [34], for example.).

Anyway, the training needed seems to be shorter than that needed for other paradigms we will present.

For example, regarding the azimuth localization, Fig. 4 shows the intensity mapped in each ear versus the virtual angle of the source.

However, some smaller problems remain unresolved, such as "to determine if the most naturally perceived direction information is obtained when (a) referenced to the person's head orientation, or (b) referenced to the person's body orientation." [35], p. 195.

More problems can be found after analyzing existing psychoacoustic proposals: in vertical localization, HRTF proposals present high errors, as shown in [36], who found errors between 9° and 15° in vertical localization. This is a quite low error, and some researchers recommend not implementing vertical information at all [37]. Some projects follow this recommendation, like the CASBliP [38]. In this case, the user must move his head, since the processing is done in the horizontal plane defined by the eyes.

Likewise, the computation load is high, because of the two complex convolutions needed for each sound source. As Castro Toledo stated in 2006, "In order to avoid the poor perception of the elevation, the use of individualized HRTFs functions is recommended. An individualized HRTF function is computationally-complex and cannot be used for real-time spatial rendering of multiple moving sources." [39]. This statement makes sense if the HRTF complexity is different due to individual characteristics. If only weighing changes are done over a generic HRTF, the complexity remains the same. Anyhow, different sources increment the computational complexity of the virtualized sound, since their localization also depends on the frequency among other parameters [40, 41]. The effect of several sound sources (as it may happen in a mobility task) also make this paradigm hard to be understood and managed by the users, as shown in [42] or even decreases the performance [43, 44].

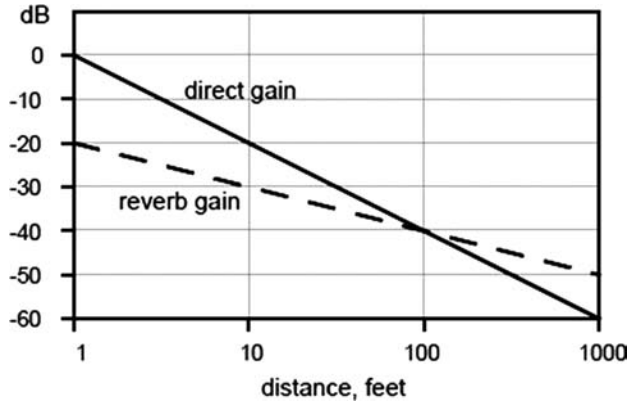

Fig. 5 Loudness-distance perception curve, from [46]. Direct sound source and reverb does not behave in the same way.

In the case of distance sonification, the psychoacoustic paradigm is followed in every project found; however, some issues can be pointed out:

- Some of them link, at the same time, the frequency to the distance [9,31]. These systems should then be in the "mixed" category.
- Others propose an inverse coding of the distance, following the psychoacoustic curve of distance perception, as done in [45].

These data must be carefully read, since distance detection (as it happens with vertical discrimination) is strongly related to other aspects of the sound source, as explained in [41].

Finally, examples of the ETAs mainly implementing this paradigm are, among others, the Multi and cross-modal ETA [10, 11], the Echolocation [12], the EAV [13], the 3-D Support [14–16], the FIU Project [47], the SWAN [48] or the CASBliP [17, 18].

There is an interesting distinction defined by Walker and Kramer [49] which separates *intentional* sounds as those "purposely engineered to perform as an information display", from *incidental* sounds "which are non-engineered sounds that occur as a consequence of the normal operation of a system". In many psychoacoustic based sonifications, it can be seen how clicks and other *incidental* sounds are used to transmit the information (i.e. the information does not travel in the click itself but in the spatialized transform of this click). It can be found, then, another branch which tries to transmit information through the sound itself. This is what we call *artificial* sonification.

B. Artificial sonification

In general terms, artificial sonification uses some other characteristics of the sound, such as frequency, brightness or timbre, formants, saturation, time intervals, etc. which, not being related to physical characteristics or parameters of objects or of surroundings, can be linked artificially to them in order to transmit more information. This linking is, somehow, artificial, i.e., there is no physical/mathematical

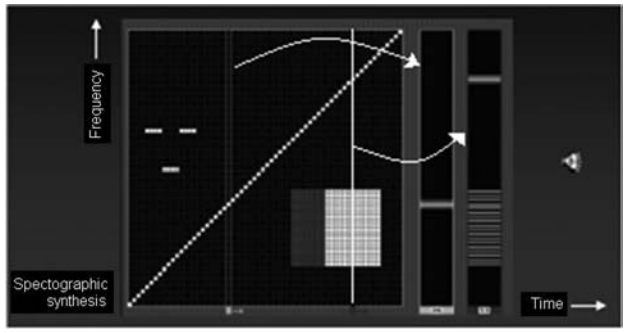

Fig. 6. Scene in two consecutive moments (the first scanning bar in gray, the second one in white). The two columns at the right represent the notes excited by the two scanning bars. The higher is the pixel detected, the higher is the pitch. Extracted and modified from [19].

reason why one option and not another one should be followed. When using the artificial sonification, users must learn how real world objects and the sounds that represent them are related, and how the mapping was designed. We were talking about indexes (causality), icons (similarity) and symbols (conventionalism), to represent objects and, now, we can apply these concepts to this paradigm. Oswald [50] states that the learning effort increases from indexing (the easiest representation) to the symbolic representation (the hardest one). There can be "intuitive" links between image cues and sounds. Indexing can be used, for example, whenever larger objects producing larger (or louder) sounds or some relations about the frequency and the distance could be easily understandable. Iconicity was used by so-called *auditory icons* [51]. Some other proposals implement purely conventional representations (symbols), which were defined as *earcons* [52]. In fact, systems can be found employing mappings simply because the system designer(s) felt that the mapping would be easier to understand. Some of them will be presented in the following paragraphs.

One of the deepest studies dedicated to this field was done by O'Hea during his Ph.D. research [53], after which three paradigms in artificial sonification can be defined:

- Dallas [54] mapped vertical position to frequency, horizontal position to time, and brightness to loudness. This mapping is an example of the *piano transform* (shown in Fig. 1). In this mapping, the system needs some time to scan the image from left to right and transform it to sequential sounds. Because of that, this mapping does not allow real-time processing in any case. A modern example of this transform can be found in the vOICe project [19]. In this case, a click represents the beginning of the scene description, as shown in Fig. 6.
- There is a version of the vOICe which maps left sounds (i.e., the first ones that are reproduced) to the left ear, and the right ones to the right one, thus, helping the understanding of the scene by means of psychoacoustic transformations. The main advantage of this option is the temporal structure of the scene information, sequentially perceiving the information of the scene from the left to the right. However, we have to point out that this mapping does not allow real-time implementations, since a representation of the scene requires a gap of time. In the case of the vOICe, "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" [19], p. 115. The scan time T is set between one and two seconds for 64×64 images.
- Another option can be added to this mapping which maps distance to frequency, proposed by Milios: "to be mapped to higher frequencies thereby stressing their importance of objects nearby. The maximum frequency in this mapping (4200 Hz) corresponds to a range measurement of 0.30 m whereas the minimum frequency (106.46 Hz) corresponds to a range measurement of 15 m" [6] (p. 417). We will call this option "pitch transform". The AudioGuider [45], the Sonic Pathfinder [25] or the K-sonar [55] could be also classified as pitch transforms. The main problem of this approach is the learning process for a somewhat intuitive dimension such as the distance given and advantage in the case of highly noisy environments, where differences of loudness might not be perceived and differences of pitch, at constant volume, work much better. As a modification of this paradigm, implemented in the case of car parking systems, decreasing time interval between impulses also provide accurate perception of distance.

Another related proposal is that based on *spearcons* [56], but although classifiable, we will not treat them as an image or scene sonification since it is mainly used in auditory menus. The main advantage of spearcons is that no training is needed. The main disadvantage is the loss of a huge amount of information, since this kind of transform is limited to recognize simple forms or structures which are linked to specific verbal messages. A comparison in the performances of both auditory icons and earcons can be found in [57]. Another study demonstrated the superiority of spearcons against auditory icons and earcons [58]. We also found in recent researches other proposals, such as the *morphocons*, which "allow the construction of a hierarchical sound grammar based on temporal variation of several acoustical parameters" [59]. The NAVIG project [60] is based on this proposal. Finally, and forcing the mapping concept, we will propose the verbal transform, i.e., translating the scene into words (synthetic or recorded) which allows the user to form a mental representation of the surroundings. For example, the Mini-radar [61] produces messages like "Stop" or "Free way" as soon as an obstacle of middle height (1.2 m) is in front of the system, or it is not. However, Loomis *et al.* [62] demonstrated that the verbal instructions had lower performance (in terms of orientation accuracy and cognitive load) than the HRTF. Likewise, complex scenes are hardly convertible automatically to speech. Moreover, it has been found [63] that some external references are needed to perform a "straight walking", and "free way" messages may not be enough.

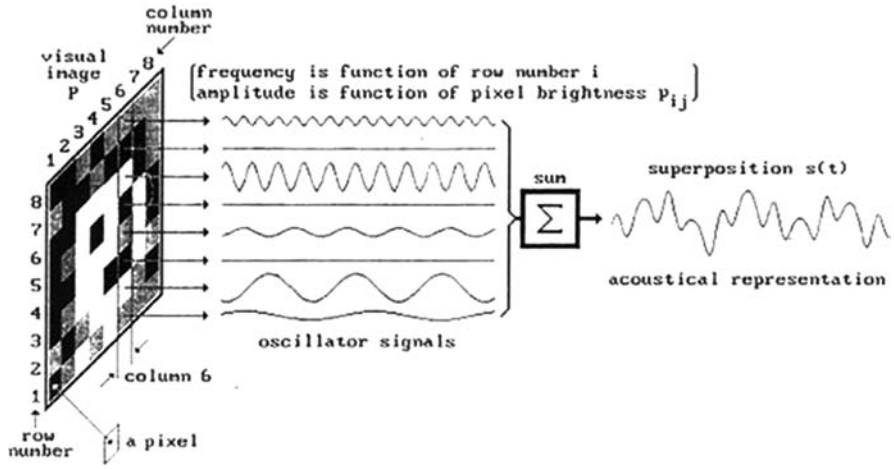

Fig. 7. Scheme of the point mapping.

Table 2. Accuracy determining spatial parameters by means of two sonification methods: direct and musical octave [66].

Object Characteristics	Direct Mapping Method	Musical Octave Method
Position (Top/Bottom)	66%	78%
Position (Left/Right)	100%	100%
Shape	72%	88%
Size	77%	91%
Distance	32%	98%
Pleasantness	40%	91%
Distance	32%	98%

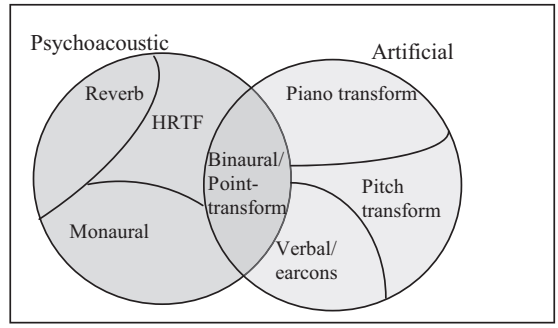

Fig. 8. Taxonomy of image sonification procedures.

C. Mixed solutions

Although we dedicate a specific section to the mixed solutions in sonification, we have already seen some of them in the artificial approach. The mixed approach tries to overcome the problems of both the previously presented options:

- On the one hand, these systems take the psychoacoustic approach from the binaural capacity of the hearing system, to apply it to the azimuth localization of the sound sources. Moreover, the distance is usually related to the loudness of the sound, following a weberian law.
- On the other hand, elevation and other non psychoacoustically related characteristics (texture, color, edges, velocity. . .) are mapped into artificial parameters, trying to reduce the error of the localization (in the case of elevation) or to represent extra information, which could be relevant. Moreover, as shown in [64], blind people encounter bigger problems to localize objects in the vertical axis, regarding the sighted ones.

The principal advantage is, mainly, a reduced learning process compared to the strictly artificial approaches. The main drawback is that, however, training is then mandatory because of the artificial implementation of some parameters.

Fish [65] used frequency to map vertical position, and a binaural loudness difference for horizontal positions (here a psychoacoustic implementation for this second dimension can be seen). Brightness is mapped to loudness. All these three parameters vary frame-by-frame and this is usually called point mapping. The basic process of this transform (for an 8×8 image) is shown in Fig. 7, modified from that of [19].

In this figure, each row produces a different pitch (represented as longer or shorter wavelengths before the sum), while the horizontal position is mapped in the stereo space. The amplitude of each produced wave is proportionally related to the brightness, as can be seen in the figure. Two final complex waves (one for each stereo channel) are produced from each frame.

The size of the sonified image (the 64×64 in the case of the vOICe, or the 8×8 of Fig. 7) is not limited by technical issues, but rather by the ability of the user to derive useful information from the sonification.

For example, the SVETA project [29] generates the sound image according to the following expression:

$$S(j) = \sum_{i=1}^{N} I(i, j)M(i, j)$$

Where *S(j)* is the sound pattern produced from column j of the image, j = 1,2, . . . ,16 and j = 32,31, . . .17 for stereo type scanning, *I(i, j)* is the intensity value of $(i, j)^{th}$ element, and *M(i,j)* is the sample of musical tone for $(i, j)^{th}$ pixel. In this case, the brightness is related to the depth (distance) of the pixel. The main advantage of this transform is the instant representation of the scene. For any static image, a stationary sound pattern is created. The problem is that,

Table 3. Summary of ETAs and main characteristics.

Device Name and Reference	Sonification input	Sonification group (Art: artificial, Psy, psychoacoustic, Mix: mix of them)	Notes
Nottingham OD [3]	Ultrasounds	Art: Pitch transform	Musical notes for each distance (thus, discrete measurement). Lower pitch means closer.
Sidewalk detector [8]	2D images	Art: Auditory icons	Edge detection in mono-vision. Alerts when not walking straight.
vOICe [19,70–72]	2D images	Art: Piano transform	Direct light mapping, mono and stereo versions. Another version with 2.5D images and, thus, depth perception).
Sonic Mobility Aid [73]	Ultrasounds	Art: Pitch transform	The further is the object, the higher is the pitch.
K sonar [55]	Ultrasounds	Art: Pitch transform	The closer the object is, the lower the pitch is.
Milios [6]	Laser beam	Art: Pitch-Point transform	Discrete distance sonification. Distance mapped to frequency and loudness. Musical instruments to make sounds softer. Proportional and derivative modes.
Mini-radar [61]	Ultrasounds	Art: Verbal	Two messages: “Stop” and “Free way”. Other options, distance measurement, directional stability, detection of light and 8 sectors compass.
NAVIG project [60]	2.5D images (and other systems)	Mix: Spatial Audio and Morphocons	In V1 only text-to-speech and spatialized sounds. In further versions (in 2012), also morphocons.
SWAN [48]	RF-RFID-Inertial sensors-Infrared	Mix: HRFT-Verbal	Synthetic voice, distance mapped to loudness.
Bujacz [31]	2.5D images	Mix: HRTF	Distance mapped into loudness and frequency.
Sonic Pathfinder [21–25,74]	Ultrasounds	Mix: Pitch transform	5 transducers and 3 directions. Only the nearest object is sonified. Musical scale used to map distance, descending with it.
SVETA [29,66]	2.5D images	Mix: Point transform	Pitch limited to musical scale.
AudioGuider [45]	2D images	Mix: Verbal-Pitch transform	Synthetic voice, monaural sounds for pitch transform.
Navbelt [9]	Ultrasounds	Mix: Pitch-Amplitude	Distance mapped to loudness and pitch. Sounds virtually placed in the horizontal axis.
CASBliP [17, 18,38]	1.5D images- Head positioning sensor	Psy: Binaural	1.5D image: one line with depth information. No vertical sonification, distance mapped into loudness.
NAVI [20]	2D images	Psy: Binaural	Object-background processing. No information about distance is given.
Echolocation [12]	Ultrasounds	Psy: Binaural	Distance mapped to loudness. Ultrasounds received are downconverted and directly presented to the user.
Cross-modal ETA [10, 11]	2.5D images	Psy: HRTF	It also uses reverb to enhance the distance perception.
EAV [13]	2.5D images	Psy: HRTF	Clicks virtually emitted from the real objects directions.
3-D Support [14–16,75]	2.5D images	Psy: HRTF	It uses the same sound of the real object, if the system arrives to recognize it.
FIU Project [47]	Ultrasounds	Psy: HRTF	6 different directions sonified.
Mims [3]	Infrared beam	Psy: Monaural	Direct transmission of noise detected when an object enters in the field of view.
FOA Laser cane [3]	Laser beam	Psy: Monaural	One channel transmission, to make it more usable.
Ultrasonic torch [3]	Ultrasounds	Psy: Monaural	One channel transmission and two ranges: 2.1 m and 6 m.
Sonic Guide (prev. KASPA) [27,76]	Ultrasounds	Psy: Monaural	Ultrasounds received are downconverted or lowpass filtered and directly presented to the user, with the emitting pulse. Problems with multiple objects.
FIU Project [47]	Ultrasounds	Psy: Monaural	Distance mapped to loudness.

since time is not a function of the scene, the information is embedded in the same pattern and it might be more difficult to discriminate small details in the perceived pattern. The advantage of using musical notes and not direct frequency representation is shown in Table 2.

There are also proposals around the point transform mixing psychoacoustic and artificial rules even for the same cue, implementing some redundancy to enforce the localization of the perceived objects [67]. There are in this group a lot of assistive products, such as that presented in [20], the Sonic Pathfinder [25], the Sonic Guide [26, 27], the Single Object Sensor [3], the KASPA (a binaural implementation of the K-sonar) [68], the SVETA [29,66], or the AudioGuider [45].

Finally, we can find a mixed solution based on the pitch transform: the Navbelt [9] uses both pitch and amplitude to represent the distance, with higher amplitude and pitch when the object is closer.

Fig. 8 shows the final taxonomy derived from the analysis and discussion presented thus far.

Classifying a sonification procedure is not an automatic process, since the definition of success may vary depending on the application, scenario or even design goals. Likewise, even if some objective characteristics of the sonification procedures, such as accuracy, speed or memory consumption could be used, it is extremely rare to find such information in the scientific or commercial literature. Thus, it is very difficult to compare and classify the different proposed paradigms and algorithms even though they are proposed to solve the same problem, mainly scene sonification.

Some characteristics that should help to classify algorithms are the following:

- Rules used to sonify (and classification of the algorithm in some of the main groups)
- Time to compute the sonification
- Technology used
- Time to represent the scene
- Number of visual cues represented
- Minimum lateral change represented
- Minimum lateral change perceived by users (if not psychoacoustic based sonification)
- Minimum vertical change represented
- Minimum vertical change perceived by users (if not psychoacoustic based sonification)
- Minimum distance change represented
- Minimum distance change perceived by users
- Minimum area represented
- "Empty space" representation (time, frequencies, delays, etc.)
- "Full space" representation (time, frequencies, delays, etc.)

A summary of the retrieved proposals and devices is given in Table 3.

IV. CONCLUSIONS

The sonification process requires some artificial mapping decisions, since almost no automatic or natural code exists to translate images into sounds. However, no standard method has been found.

The sonification methods have been classified according to how the visual information is translated to sounds, along with the decisions taken to perform such transformation.

Each one of them has been analyzed in terms of advantages and disadvantages regarding the ease of use and the precision of the information provided.

This research field is open and should be developed in the coming years to overcome the limitations and constraints found. One important aspect, not discussed because it doesn't belong to the sonification world, but which is very relevant for the users, is the use of headphones or earphones. Most of the ETAs found, implementing HRTF algorithms, use this kind of devices (SVETA, 3D-support, cross-modal ETA, vOICe, EAV, NAVI, etc.). Researchers should avoid blocking the ears to blind people, since they are their main input path of information. This "ears-open" design has been discussed in relation to the SonicGuide (using small tubes) and the SWAN system (using bone conduction headphones) [60,69].

A proper taxonomy of algorithms helps in understanding and structuring the state of the art database, and guiding further research in the most appropriate direction. However, nowadays this is difficult given that there are still no quality and accuracy metrics in this field. This is due to the variety of approaches, some of them completely different (such as the psychoacoustic and the verbal representation), which prevents the proposal of clear and objective metrics. This problem could be reduced if standard testbeds were proposed. These testbeds should be able, at least, to measure the variables proposed in the previous list.

REFERENCES

[1] M. Capp, Ph. Picton: The Optophone: An Electronic Blind Aid. *Engineering Science and Education Journal*. 9 (2), 137–143 (2000)

[2] W. Starkiewicz, T. Kuliszewski: The 80-channel elektroftalm. Proceedings of the lnternational Congress Technology Blindness, *Am.Found.Blindness.* 1, 157 (1963)

[3] L.W. Farmer: Mobility devices, *Bull Prosthet Res.* 47–118 (1978)

[4] F. d'Albe: *The moon element*, T. Fisher Unwin, Ltd., London (1924)

[5] R.L. Beurle: *Summary of suggestions on sensory devices*, San Dunstan's, London (1947)

[6] E. Milios, B. Kapralos, A. Kopinska, S. Stergiopoulos: Sonification of range information for 3-D space perception. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 11 (4), 416–421 (2003)

[7] E.H. Weber: *De Pulsu, Resorpitone, Auditu et Tactu: Annotationes Anatomicae et Physiologicae.*, Koehlor, Leipzig, Germany (1834)

[8] X Jie, W. Xiaochi, F. Zhigang: Research and Implementation of Blind Sidewalk Detection in Portable ETA System. *International Forum on Information Technology and Applications*. 431–434 (2010)

[9] S. Shoval, J. Borestein, Y. Koren: Auditory Guidance with the Navbelt-A Computerized Travel Aid for the Blind.

IEEE Transactions on Systems, Man, and Cybernetics. 28 (3), 459–467 (1998)

[10] A. Fusiello, A. Panuccio, V. Murino, F. Fontana, D. Rocchesso: A Multimodal Electronic Travel Aid Device. Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces. 39–44 (2002)

[11] F. Fontana, A. Fusiello, M. Gobbi, V. Murino, D Rocchesso, L. Sartor, A. Panuccio: A Cross-Modal Electronic Travel Aid Device. Mobile HCI 2002, *Lecture Notes on Computer Science*. 2411, 393–397 (2002)

[12] T. Ifukube, T. Sasaki, C. Peng: A Blind Mobility Aid Modeled After Echolocation of Bats, *IEEE Transactions on Biomedical Engineering*. 38, 461–465 (1991)

[13] J. Gonzalez-Mora, A. Rodriguez-Hernandez, E. Burunat, F. Martin, M. Castellano: Seeing the world by hearing: Virtual Acoustic Space (VAS) a new space perception system for blind people, International Conference on Information & Communication Technologies: from Theory to Applications (IEEE Cat.No.06EX1220C), 6-ROM (2006)

[14] Y. Kawai, F. Tomita: A Support System for Visually Impaired Persons Using Acoustic Interface - Recognition of 3-D Spatial Information. *HCI International.* 1, 203–207 (2001)

[15] Y. Kawai, F. Tomita: A Visual Support System for Visually Impaired Persons Using Acoustic Interface. IAPR Workshop on Machine Vision Applications (MVA 2000). 379–382 (2000)

[16] Y. Kawai, F. Tomita: A Support System for Visually Impaired Persons Using Three-Dimensional Virtual Sound. International Conference on Computers Helping People with Special Needs (ICCHP 2000). 327–334 (2000)

[17] M.M. Fernández Tomás, G. Peris-Fajarnés, L. Dunai, J. Redondo: Convolution application in environment sonification for Blind people. VIII Jornadas de Matemática Aplicada, UPV. (2007)

[18] N. Ortigosa Araque, L. Dunai, F. Rossetti, L. Listi, M. Mirmehdi, J.L. González Mora, A. Rodriguez Hernández, A. Meloni, S. Morillas Gómez, A. Schick, L. Scalise, V. Santiago Praderas, G. Peris-Fajarnés, I. Dunai: Sound Map Generation for a Prototype Blind Mobility System Using Multiple Sensors. ABLETECH 08 Conference. 10 (2008)

[19] P.B.L. Meijer: An Experimental System for Auditory Image Representations, *IEEE Transactions on Biomedical Engineering*. 39, 112–121 (1992)

[20] G. Sainarayanan, R. Nagarajan, S. Yaacob: Fuzzy image processing scheme for autonomous navigation of human blind. *Applied Soft Computing.* 7 (1), 257–264 (2007)

[21] A.D. Heyes: The use of musical scales to represent distance to object in an electronic travel aid for the blind. *Perceptual and Motor Skills*. 51 (2), 68–75 (1981)

[22] A.D. Heyes: Human Navigation by Sound. *Physics in Technology*. 14 (2), 68–75 (1983)

[23] A.D. Heyes: The Sonic Pathfinder - A new travel aid for the blind. in: W.J. Perk and Ed. s (Eds.), *In Technology aids for the disabled*, Butterworth, 165–171 (1983)

[24] T. Heyes: Sonic Pathfinder: A Programmable Guidance Aid for the Blind, *Electronics & Wireless World*. 90, 26–29 & 62 (1984)

[25] A.D. Heyes, G. Clarcke: The role of training in the use of the Sonic Pathfinder. Proceedings of the American Association for the Education and rehabilitation of the Blind and Visually Impaired, Southwest Regional Conference, Hawaii (1991)

[26] L. Kay: Auditory perception of objects by blind persons, using a bioacoustic high resolution air sonar, *Journal of the Acoustical Society of America*. 107, 3266–3275 (2000)

[27] N.C. Darling, G.L. Goodrich, J.K. Wiley: A preliminary followup study of electronic travel aid users, *Bull Prosthet Res.* 10, 82–91 (1977)

[28] C. Morris Signs: Language and Behaior. *Writtings on a General Theory of Signs The Hague*: Mouton, 79–397 (1971)

[29] G. Balakrishnan, G. Sainarayanan, R. Nagarajan, S. Yaacob: Fuzzy matching scheme for stereo vision based electronic travel aid, Tencon 2005 - 2005 Ieee Region 10 Conference, Vols 1-5, 1142–1145 (2006)

[30] R.F.A. Rossi, M.K. Zuffo, J.A. Zuffo: Improving spatial perception through sound field simulation in VR, 2005 IEEE International Conference on Virtual Environments, Human-Computer Interfaces and Measurement Systems (IEEE Cat.No.05EX1045C), 103–108 (2005)

[31] M. Bujacz, P. Skulimowski, P. Strumillo: Naviton - a prototype mobility aid for auditory presentation of 3D scenes, *Journal of Audio Engineering Society*. 60, 696–708 (2012)

[32] D. Bavelier, H.J. Neville: Cross-modal plasticity: where and how? *Nature Reviews Neuroscience*. 3, 443- 452 (2002)

[33] J.A. Ramírez Rábago: Generación de fuentes virtuales de sonido en audífonos (2005)

[34] *MIDI Manufacturers Association MMA*: General MIDI 1, 2 and Lite Specifications. http://www.midi.org/techspecs/gm.php (2012)

[35] D.A. Ross, B.B. Blasch: Wearable Interfaces for Orientation and Wayfinding. ASSETS'00. 193–200 (2000)

[36] M. Pec, M. Bujacz, P. Strumillo, A. Materka: Individual HRTF Measurements for Accurate Obstacle Sonification in an Electronic Travel Aid for The Blind. International Conference on Signals and Electronic Systems (ICSES 2008). 235–238 (2008)

[37] G. Wersényi: Effect of Emulated Head-Tracking for Reducing Localization Errors in Virtual Audio Simulation, IEEE Transactions on Audio, Speech, and Language Processing. 17, 247–252 (2009)

[38] G. Peris-Fajarnés, I. Dunai, V. Santiago Praderas: Detección de obstáculos mediante sonidos acústicos virtuales. *DRT4ALL* 2011. 133–139 (2011)

[39] D. Castro Toledo, S. Morillas, T. Magal, G. Peris-Fajarnés: 3D Environment Representation through Acoustic Images. Auditory Learning in Multimedia Systems. Proceedings of Concurrent Developments in Technology-Assisted Education. 735–740 (2006)

[40] D.S. Brungart: Auditory localization of nearby sources III. Stimulus effects. *Journal of the Acoustical Society of America*. 106 (6), 3589–3602 (1999)

[41] D.S. Brungart, B.D. Simpson: Effects of temporal fine structure on the localization of broadband sounds: potential implications for the design of spatial audio displays, Proceedings of the 14th International Conference on Auditory Display (ICAD 2008) (2008)

[42] B. Simpson, N. Iyer, D. Brungart: Aurally Aided Visual Search with Multiple Audio Cues, *Proceedings of the 16th International Conference on Auditory Display (ICAD2010)*. June 9-15 (2010)

[43] D. Navon, D. Gopher: On the Economy of the Human-Processing System, *Psychological Review*. 86, 214–255 (1979)

[44] D.A. Norman, D.G. Bobrow: On Data-Limited and Resource-Limited Processes, *Cognitive Psychology*, 44–64 (1975)

[45] F. Zhigang, L. Ting: Audification-based Electronic Travel Aid System. *IEEE International Conference on Computer Design and Applications (ICCDA 2010)*. 5, 137–141 (2010)

[46] J. Blauert: *Spatial hearing: the psychophysics of human sound localization.*, MIT Press Cambidge (1983)

[47] D. Aguerrevere, M. Choudhury, A. Barreto: Portable 3D sound / sonar navigation system for blind individuals. 2nd LACCEI International Latin American and Caribbean Conference on Engineering Technology. 2–4 (2004)

[48] B.N. Walker, J. Lindsay: Using virtual reality to prototype auditory navigation displays. *Assistive Technology Journal.* 17 (1), 72–81 (2005)

[49] B.N. Walker, G. Kramer: Human factors and the acoustic ecology: Considerations for multimedia audio design. Proceedings of the Audio Engineering Society 101st Convention (1996)

[50] C. Oswald: Non-Speech Audio-Semiotics. A review and Revision of Auditory Icon and Earcon Theory. Proceedings of the 18th International Conference on Auditory Display. 36–43 (2012)

[51] W.W. Gaver: Using and creating auditory icons. in: G. Kramer (Ed.), *Auditory display: sonification, audification, and auditory interfaces*, Addison-Wesley, 417–446 (1994)

[52] M.M. Blattner, D.A. Sumikawa, R.M. Greenberg: Earcons and icons: Their structure and common design principles, *Human-Computer Interaction*. 4, 11–44 (1989)

[53] A.R. O'Hea: *Optophone design: optical-to-auditory substitution for the blind*. The Open University, UK. (1994)

[54] S.A. Dallas, A.L. Erickson: Sound pattern generator representing matrix data formatlhas matrix video converted to parallel form, modulating audio tone, giving video information in terms of time and frequency. (WO8200395-A1; EP55762-A; US4378569-A; CA1165447-A; IL63239-A; EP55762-B; DE3174174-G) (1960)

[55] Bay Advanced Technologies: *'K' SONAR The Handbook*, Auckland 1005 New Zealand (2006)

[56] B.N. Walker, A. Nance, J. Lindsay: SPEARCONS: Speech-based Earcons Improve Navigation Performance in Auditory Menus, Proceedings of the 12th International Conference on Auditory Display (ICAD 2006) (2006).

[57] R. Absar, C. Guastavino: Usability of Non-Speech Sounds in User Interfaces, Proceedings ofthe 14th International Conference on Auditory Display(2008) (2008)

[58] T. Dingler, J. Lindsay, B.N. Walker: Learnability of Sound Cues for Environmental Features: Auditory Icons, Earcons, Spearcons, and Speech, Proceedingsof the 14th International Conference on Auditory Display (2008).(2008).

[59] G. Parseihian, B.F.G. Katz: Morphocons: A New Sonification Concept Based on Morphological Earcons, *Journal of Audio Engineering Society*. 60, 409–418 (2012)

[60] B.F.G. Katz, F. Dramas, G. Parseihian, O. Gutierrez, S. Kammoun, A. Brilhault, L. Brunet, M. Gallay, B. Oriola, M. Auvray, P. Truillet, M. Denis, S. Thorpe, C. Jouffrais: NAVIG: Guidance System for the Visually Impaired Using Virtual Augmented Reality, *Journal of Technology and Disability*. 24, 163–178 (2012)

[61] BESTPLUTON World Cie: The "Mini-Radar", your small precious companion that warns you obstacles in a spoken way, and that helps you to walk straight. http://bestpluton.free.fr/EnglishMiniRadar.htm (14-4-2011)

[62] J.M. Loomis, R.G. Golledge, K.L. Klatzky: Navigation System for the Blind: Auditory Display Modes and Guidance. *Presence.* 7 (2), 193–203 (1998)

[63] G. Wersényi, J. Répás: The Influence of Acoustic Stimuli on "Walking Straight" Navigation by Blindfolded Human Subjects, *Acta Technica Jaurinensis*. 5, 1–18 (2012)

[64] G. Wersényi: Virtual Localization by Blind Persons, *Journal of the Audio Engineering Society*. 60, 568–579 (2012)

[65] R.M. Fish: Audio Display for Blind, *IEEE Transactions on Biomedical Engineering*. 23, 144–154 (1976)

[66] G. Balakrishnan, G. Sainarayanan, R. Nagarajan, S. Yaacob: Stereo Image to Stereo Sound Methods for Vision Based ETA. 1st International Conference on Computers, Communications and Signal Processing with Special Track on Biomedical Engineering, CCSP 2005, Kuala Lumpur. 193–196 (2005)

[67] P. Revuelta Sanz, B. Ruiz Mezcua, J.M. Sánchez Pena: A Sonification Proposal for Safe Travels of Blind People. Proceedings of the 18th International Conference on Auditory Display (ICAD 2012). 233–234 (2012)

[68] L. Kay: KASPA. http://www.batforblind.co.nz (2005)

[69] B.N. Walker, R.M. Stanley, A. Przekwas, X.G. Tan, Z.J. Chen, H.W. Yang, P. Wilkerson, V. Harand, C. Chancey, A.J.M. Houtsna: High Fidelity Modeling and Experimental Evaluation of Binaural Bone Conduction Communication Devices, Proceedings of the 19th International Congress on Acoustics (ICA 2007) (2007)

[70] P.B.L. Meijer: Seeing with Sounds: is it Vision? Invited presentation at VSPA 2001 Conf. on Consciousness, Amsterdam (2001)

[71] P.B.L. Meijer: Stereoscopic Vision for the Blind: Binocular vision support for The vOICe auditory display. http://www.seeingwithsound.com/binocular.htm (2011)

[72] P.B.L. Meijer: Seeing with Sound for the Blind. Is it Vision? Can it be? Invited presentation at Tucson 2002, Tucson, Arizona (2002)

[73] L.H. Riley, G.M. Weil, A.Y. Cohen: Evaluation of the Sonic Mobility Aid. *American Center for Research in Blindness and Rehabilitation*, 125–170 (1966)

[74] T. Heyes: The domain of the sonic pathfinder and an increasing number of other things. From http://www.sonicpathfinder.org (2004)

[75] Y. Kawai, F. Tomita: A Support System for Visually Impaired Persons to Understand Three-dimensional Visual Information using Acoustic Interface, *16Th International Conference on Pattern Recognition, Vol Iii, Proceedings*. 974–977 (2002)

[76] L. Kay: Auditory Perception and its Relation to Ultrasonic Blind Guidance Aids. Presented at the Symposium on "Practical Electronic Aids for the Handicapped" in London on 28th March 1962, Journal of Brit.I.R.E. October 1962, 309–317 (1962)

THE AUTHORS


Pablo Revuelta Sanz


Belén Ruiz Mezcua


José M. Sánchez Pena


Bruce N. Walker

Pablo Revuelta Sanz, obtained his degree in Telecommunication Engineering (2006) at the Carlos III University of Madrid (Spain), and obtained his Master's degree in 2008 from the same university.

He is a Ph.D. student. He has published several conference papers and two book chapters focused on assistive products and image processing.

•

Belén Ruiz Mezcua is Ph.D. in Physics from the Telecommunications School of the UPM (1995).

She is Adjunct Vice-chancellor of research at the Carlos III University and lecturer in the Computer Science department of the Carlos III University. She is the head of the Spanish caption Center "Centro Español de Subtitulado y Audiodescripción".

•

José M. Sánchez Pena received his MS and PhD degrees in Telecommunication Engineering from the Polytechnic University of Madrid in 1988 and 1993, respectively.

He is currently a full professor and responsible for the Displays and Photonic Applications Group in the Electronic Technology Department.

He is a senior member of IEEE.

•

Bruce N. Walker is an Associate Professor at Georgia Tech, in the School of Psychology and Interactive Computing. Particular research interests include sonification and auditory displays. Professor Walker teaches HCI, Sensation & Perception, Auditory Interfaces, and Assistive Technology. In addition to academic research leading to over 100 publications, he has worked and consulted on projects for NASA, state and federal governments, private companies, and the military.