A Multimodal Electronic Travel Aid Device

A. Fusiello, A. Panuccio, V. Murino, F. Fontana, D. Rocchesso
Dipartimento di Informatica - Università degli Studi di Verona
Strada Le Grazie, 15 - 37134 Verona, Italy
{fusiello,panuccio,murino,fontana,rocchesso}@sci.univr.it

Abstract

This paper describes an Electronic Travel Aid device, that may enable blind individuals to "see the world with their ears". A wearable prototype will be assembled using low-cost hardware: earphones, sunglasses fitted with two micro cameras, and a palmtop computer. The system, which currently runs on a desktop computer, is able to detect the light spot produced by a laser pointer, compute its angular position and depth, and generate a corresponding sound providing the auditory cues for the perception of the position and distance of the pointed surface patch. It permits different sonification modes that can be chosen by drawing, with the laser pointer, a predefined stroke which will be recognized by a Hidden Markov Model. In this way the blind person can use a common pointer as a replacement of the cane and will interact with the device by using a flexible and natural sketch based interface.

1. Introduction

Electronic Travel Aids (ETA) devices aim at conveying information (typically via the haptic and auditory channels) about the environment to visually impaired individuals, so that they can exploit (at least) part of the information that sighted people normally use to experience the world and navigate it.

Since the mid-seventies, a number of vision substitution devices have been proposed which convert visual into auditory or haptic information with the aim of providing mobility aids for blind people [12, 7, 10]. Early examples of this approach are the Laser-Cane [3] and the Ultra Sonic Torch [15]. These devices are all based on producing beams of ultrasonic sound or laser light: the device receives reflected waves, and produces either an audio or tactile stimulus in response to nearby objects. The intensity of the sound or tactile vibration is proportional to the distance of the pointed object. In the VIDET project [28] a device converting the information of object distance into haptic cues has been developed.

As compared to visual information, sound does not rely on focused attention of the human subject: it is pervasive in a 3-D environment, less sensible to occlusion, and it allows the simultaneous transmission of multiple streams that can be easily segregated or aggregated by humans. However, it is difficult to design sounds in such a way that they effectively become information carriers. When one aims at using sounds to display spatial features of objects, it should be considered that, while vision is best suited for perceiving attributes of light modifiers (surfaces, fluids), audition is best suited for perceiving attributes of sound sources, regardless of where they are located [16].

A significant advance in the design of auditory displays for ETAs has been made by Meijer [19], who proposed a real-time device, called The vOICe, which scans and digitizes an image captured by a video camera. This image is then divided into vertical strips. Pixels accounting for recognized objects are *sonified* using simple tones, higher pitches being associated to the pixels closer to the image top. This enables the user to recognize the elevation of objects in a scene, via an association between height and frequency of the tones. Finally, the horizontal position of objects (given by the horizontal position of the corresponding strip) is resolved using stereo panning, and the intensity of the light captured by the video camera is translated into loudness of the respective tones. In a recent work on data sonification Walker and Lane [29] determine the preferred data-to-display mappings, polarities, and psychophysical scaling functions relating data values to underlying acoustic parameters for blind and visually impaired listeners.

In [23] the authors evaluated three novel orientation interfaces, namely: a virtual sound beacon, digitized speech, and a tapping interface. Significant results were: 1) the tapping interface was usable by all under all conditions, 2) speech was sometimes confusing and not always usable, and 3) the virtual beacons were preferred by many for many situations, but were not usable in very noisy environments or by people with hearing impairments in one or both ears.

Although sound synthesis relies on well-established

techniques, yet the definition of auditory displays, which are optimized both in the perceptual and the resources consumption aspect, is still under investigation. Moreover, there is relatively little research on the integration of audio and visual cues for environment perception, although the potential benefits can be very large in the field of the human-computer interaction (HCI). In this respect, sounds that are generated as a consequence of a visual understanding process can be an important vehicle for exchanging structured and unstructured information among objects and humans, and may help and facilitate human perception in many ways.

In our research, we are exploring the idea of augmenting human visual perception using acoustic stimuli. This should lead to novel multimodal interfaces, but it also applies to the design of ETAs and other aids for blind people. On the visual side, we have to deal with methods that allow segmentation and three-dimensional (3D) reconstruction of the scene. On the auditory side, we need techniques that allow blind people to experience the 3D scene through synthetic sounds.

This paper describes a system prototype where such issues are explored. The actual system is composed by earphones, two CMOS micro cameras (a stereo head), and a desktop computer (see Fig. 1). A wearable prototype will be assembled using a palmtop computer and fitting the two micro cameras onto a pair of sunglasses. The blind person uses a laser pointer as a cane. The system detects the light spot produced by a laser pointer, computes its depth (Sec. 2), and provides suitable spatialized sounds to a blind person (Sec. 3). Interfaces (for sighted users) making use of a laser pointer have been already proposed in [20, 9].

With the increasing power of wearable computers, it will be possible to include some image analysis in the visual substitution process. The system determines the three-dimensional structure of the scene, then it segments the objects (or surface patches) contained therein based on depth (and possibly color and texture). From that information, it synthesizes sounds that convey information about the surrounding environment to the visually impaired. We foresee the following functional modes for our device:

pointer: sonification of the 3D position of the laser spot produced by a laser pointer.

global sonification: the entire environment in the cameras' field of view is sonified (the laser pointer is not used). The disparity map obtained from stereopsys is segmented in homogeneous region, and the position and area of each distinct region is sonified.

max/min depth: as the previous one, but only the nearest and farthest regions are sonified.

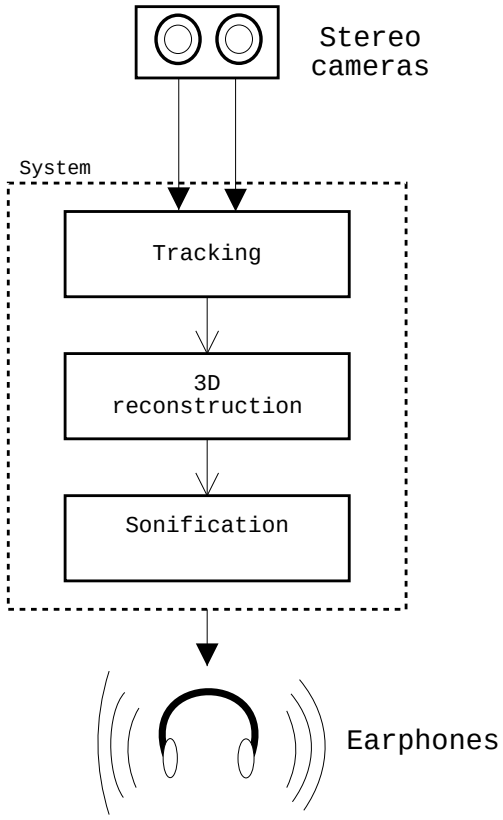

Figure 1. Overall system architecture. The laser spot is tracked, then its 3D position is computed, and this datum is sonified.

The user can select the mode by drawing a predefined stroke with the laser pointer (Sec.4).

2. Visual analysis

One of the cues used by the human visual system to acquire depth information is the disparity between the images formed on the left and right retina. This process, called stereopsys [14, 24], can be replicated by a computer, at a rate of some frames per second, using a camera pair. A well-known problem in computational stereopsys is finding *correspondences*, i.e., finding which points in the left and right images are projections of the same scene point (a *conjugate pair*). Using a laser pointer makes this task trivial, provided that we are able to locate the laser spot in the two images.

In the following, the processing applied to each image (left and right) in order to locate the target (i.e. the centroid of the blob produced by the laser pointer) in each image will

be briefly described.

The visual analysis starts by applying a red band-pass optical filter on both cameras, in order to enhance the image of the red laser spot. Then the brightness of the two images is normalized by means of a simple algorithm [8] which computes the parameters α, β of the gray-level *global* transformation

$$I_l(x,y) = \alpha I_r(x,y) + \beta \qquad \forall (x,y) \tag{1}$$

by fitting a straight line to the plot of the left cumulative histogram versus the right cumulative histogram.

In order to reduce jitters, images are smoothed in time, by averaging pixels at the same position in the preceding m frames (we used $m = 2$). Only pixels whose difference is less than a given threshold are averaged, thereby obtaining a remarkable increment of the stability of the image, without introducing motion blur:

$$\hat{I}_n(x,y) = \begin{cases} \dfrac{1}{m} \displaystyle\sum_{k=n-m}^{n} I_k(x,y) & \text{if } \forall k \ |I_k[x,y] - I_n[x,y]| < M \\ I_n(x,y) & \text{otherwise} \end{cases} \tag{2}$$

Smoothed images are binarized with an automatic threshold T:

$$T = \max\{I(x,y)\} - k. \tag{3}$$

The value of k is computed adaptively (discussed later) starting from default value of $k = 40$ (which usually gives already good results).

A size filter is then applied, which discards the connected components (or blobs) whose size is outside a given range, which have been obtained empirically by measuring the maximum and minimum laser spot size in the typical distance operating range.

In order to make the spot detection more robust, we impose the epipolar constraint [11] on the surviving blobs; only those satisfying the constraint both left to right and right to left are kept as candidate targets.

The number of candidate targets is used as a feedback to set the binarization threshold: the value of k in Eq.3 is varied using a dichotomic search until the candidate targets are more than zero and less than few units (typically 5). If k reaches a minimum value no targets are selected, presumably because the laser spot is not visible.

To increase the precision of tracking, a forecast of the position of the pointer in the image is used. Let $X(n)$ be the state matrix describing position and speed of the target at discrete time n. Let $\Phi(n, n+1)$ be the transition matrix describing the evolution of the system from time n to $n+1$ according to a constant speed model [1]. We can predict the state at the time $n+1$:

$$\hat{X}(n+1) = \Phi(n, n+1) X(n) = \begin{pmatrix} 1 & \Delta t \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_n & y_n \\ x'_n & y'_n \end{pmatrix}. \tag{4}$$

The predicted position $\hat{X}(n+1)$ will be used to match the current target to one of the target candidates found by the preceding elaborations. In the one-target case the most common choice is to take the closest candidate, as stated by the *Nearest Neighbor Data Association* algorithm [1].

Having identified the target in the current frame at position $p = (x_{n+1}; y_{n+1})$, the state is updated with:

$$X(n+1) = \begin{pmatrix} x_{n+1} & y_{n+1} \\ \dfrac{x_{n+1} - x_n}{\Delta t} & \dfrac{y_{n+1} - y_n}{\Delta t} \end{pmatrix}. \tag{5}$$

Finally, triangulation recovers the full 3-D coordinates of the laser spot, using the disparity of the targets in the two stereo images, the internal parameters of the two cameras, and the geometry of the stereo setting (obtained from calibration). Each conjugate pair ideally defines two rays that intersect in a point of the space. In real world, for errors in measurement of point positions and in calibration of the stereo head, the rays don't always intersect, so the triangulation computes the coordinates of a 3-D point nearest to both rays [11].

The stereo head is the STH-V3 model produced by Videre Design. It has a rigid aluminum body with 85 mm baseline and 3.6 mm lenses. Ignoring thermal effects, parameters can be considered constant, therefore calibration is performed once for all, using Zhang's technique [30].

Stereopsys can be also applied to build a dense disparity map, where connected regions above a predefined threshold are identified. This feature we'll be exploited by the "global sonification" operating mode described in the Introduction.

3. Sonification

Sounds that give directional cues are said to be "spatialized" [2]. In other words, they provide distance, azimuth and elevation cues about an object, assumed to be a sound source.

Despite all research done (see [5]), we have only partial knowledge about how we determine the position of a sound source, and even less about how to recreate this effect. In everyday listening, hearing a monophonic source gives not only information about the source position, but also characteristics about the environment, such as source dimension and shape of the room. The characteristics that a sound acquires during its path from source to the listener's ear canal determine the spatial cues that are conveyed to the listener by the sound.

Auditory scenes displayed using headphones need the creation, from a pure sound, of a couple of outputs that enable the user to localize the sound source. The primary cues for localizing sounds in the horizontal plane are binaural, and result from inter-aural intensity and time differences of the sounds arriving at the ears. Cues for localizing sound

in the vertical plane are mainly monaural. Finally, distance cues can be provided by properly tuning the sound loudness, and adding reverberation [5].

A *structural* model, devoted to spatialize sounds, is driven directly in its structural parameters, i.e., azimuth, elevation and distance. These parameters are mapped by the model onto sound features in order to convey the appropriate spatial cues [2, 21].

Our model adopts a versatile structural model for the rendering of the azimuth [6]. A model providing distance cues through loudness control and reverberation effects has been developed and tested. Further studies are needed to convey reliable elevation cues to the listener. This system for blind people is both an inspiration and an end application that motivates the advances of our studies on the rendering of distance and elevation.

3.1 A structural model for distance rendering

Once the model for the azimuth [6] has been adapted and fine-tuned to our application, we need a new, independent model capable of rendering distances. In this way we can reasonably think to implement these two models independently in the system, typically in the form of two filter networks in series.

It is well known that environmental reverberation is one of the main cues for auditory distance perception [5]. After experimentation with virtual acoustic environments, we realized that the shape of the enclosure affects the strength and precision of distance perception, and we decided to embed the virtual sound source into a virtual tube. So, the metaphor is that of the user listening to the laser spot (or, more generally, to the sound source associated to a visual object) by means of a long and narrow tube, a sort of phonoscope.

The listening environment we will consider is the interior of a square cavity having the aspect of a long tube, sized $8\times0.4\times0.4$ meters. The internal surfaces of the tube are modeled to exhibit natural absorption properties against the incident sound pressure waves. The surfaces located at the two edges are modeled to behave as *total* absorbers (see Fig. 2) [17].

An investigation of the physical and perceptual properties of this acoustical system is presented in a companion paper [13]. Rather, it seems reasonable to think that, although quite artificial, this listening context conveys sounds that acquire noticeable spatial cues during their path from the source to the listener. Given the peculiar geometry of the resonator, these cues should mainly account for distance. The edge surfaces have been set to be totally absorbent in order to avoid echoes originating from subsequent reflections of the wavefronts along the main direction of wave propagation. These echoes, while being ineffective for distance recognition in the range specified by the tube size, would be quite annoying for the user.

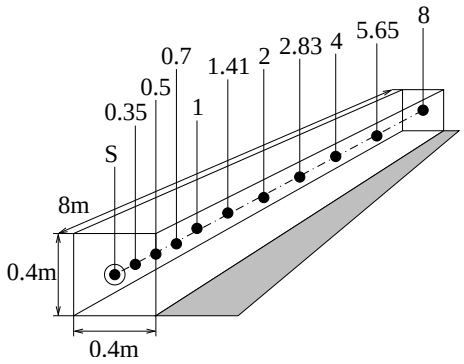

Figure 2. The listening environment.

The resonating environment is structurally simple although capable of dealing with an interesting range of the physical quantity to be rendered. In particular, the trade-off it exhibits between the provided distance range and the resonator volume is a precious feature during the development of a structural model of the resonator, as it allows a detailed physical simulation to run in reasonable time.

The square tube has been modeled by means of finite-difference schemes [25]. In particular, a *wave*-based formulation of the finite-difference scheme has been used, known as the Waveguide Mesh, that makes use of the wave decomposition of a pressure signal into its wave components [26]. This formulation, already used in the simulation of reverberant enclosures [18], has a major advantage in this application case, since it allows to deal with the boundary of the propagation domain quite effectively [27]. This enables to apply boundary *waveguide filters*, that model the reflection properties of physical room walls [17].

The final audio rendering system, still not provided with a sub-system for the elevation cues, has the structure depicted in Fig. 3. First, the source sound is provided with distance cues. Then, horizontal localization cues are added. Finally, the resulting sound is displayed through a pair of consumer headphones.

4. Sketch Based Interaction

The system will include a sketch based interaction which allows the user to select the desired scene sonification mode by drawing a predefined stroke with the laser pointer. The approach is based on the Hidden Markov Models which can absorb spatio-temporal variance of laser pointer sketches. We use four Gaussian HMMs trained on the sequence of curvature coefficients [11] extracted from the predefined strokes. This method was recently used in shape classification [4], due to its attractive intrinsic properties: the repre-

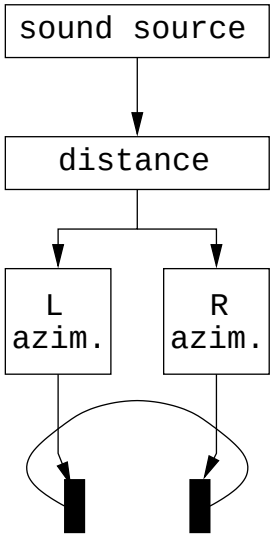

Figure 3. Structure of the system for horizontal location and distance rendering.

sentation is invariant to stroke translation and the curvature is scale invariant, assuming the signal is properly normalized. In this context, stroke recognition is performed by analyzing a sliding window on the trajectory extracted by the visual analysis of the laser pointer: its likelihood will be evaluated by the well known HMM forward procedure and compared with a proper threshold in order to discard garbage strokes. This will produce a sketching interface that feels much more natural and allows the user to easily interact with the device software without having to deal with palmtop's keys and cursors.

5. Results

Preliminary tests show that the effectiveness of laser tracking depends mainly on how "visible" the laser is inside the captured image. Therefore, if the scene is (locally) brighter than the laser spot or if the laser points too far away from the camera, then problems arise about the stability of laser tracking, because the signal to noise ratio becomes too low. Better results are obtained indoor, with a 4 mW laser source, and within a depth range of 1 to 6 meters. The power of the laser, the camera gains and the resolution are critical parameters.

Figure 4 shows an example of tracking in proximity of light sources. Thanks to the threshold on the size of the brighter area, the laser pointer is correctly tracked.

We evaluated the subjective effectiveness of the sonification informally, by asking some volunteers to use the system and report their impressions. The overall result have

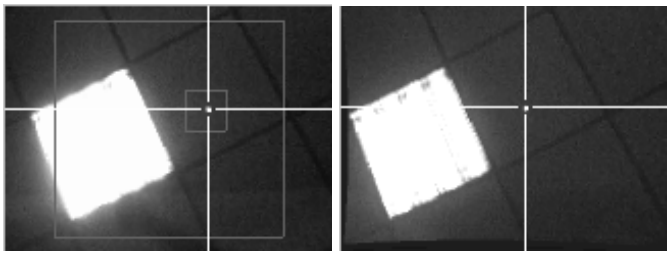

Figure 4. Tracking the laser spot close to a neon lamp.

been satisfactory, with some problems related to the lack of perception of elevation. A more systematic usability test is being planned, with a group of both sighted and visually-impaired subjects. Collins et al.[7] proposed to use the time to navigate in a particular furnished room as an evaluation metric for this kind of systems. Ross and Blash [23] used street crossing as a critical orientation situation for testing.

6. Conclusions

In this work, a new multimodal device based on cooperative use of stereo vision and sonification has been proposed. This prototype system allows to "hear" the surrounding scene acquired by a stereo head that estimate the distance of objects from the observer. The most interesting aspects in the research lie in the simplicity of the visual and the auditory models, and the efficiency of the resulting algorithms. Actually, our system can be used in place of other ETAs developed for blind people, requiring custom sensors or, alternatively, personal aids.

The system is non-intrusive in the environment and, if earplugs or open headphones are used, it does not mask the external sounds coming from the environment. However, some problems arise on the use of such device. Its prolonged use may lead to some strain, as it happens with other devices existing in the market (such as, for example, the Optophone and The vOICe), due to a continuous listening of the same signal at regular time intervals. This sound, although spatialized, produces an unnatural effect and causes a progressive fatigue. The design of a pleasant *soundscape*, possibly using dynamic sound synthesis of everyday sounds [22], will be a key point in this respect. Moreover, the laser pointer in use has a limited supporting range and works better with pretty dark environments, not directly lit by the sun.

Future developments of this prototype will be devoted to the design of alternative sound models in order to provide better distance cues, together with elevation. Another line of improvement lies in the coupling of sonic and visual information coming from objects surfaces, e.g., to acousti-

cally render material or texture, so to enrich the listener's experience about the surrounding environment.

Acknowledgments

This works is part of *"The Sounding Landscape" (SOL)* project (http://vips.sci.univr.it/html/projects.html), supported by HP through the project *HP PHILANTHROPIC*. Partial support has been given by the Project SOb - The Sounding Object (http://www.soundobject.org), as part of the European Commission's Future and Emergent Technologies collaborative R&D programme. A preliminary version appeared in *Mobile HCI 2001*, Pisa (Italy).

References

[1] Y. Bar-Shalom and T. E. Fortmann. *Tracking and data Association*. Academic Press, Orlando, FL, 1988.

[2] D. R. Begault. *3D Sound for virtual reality and multimedia*. AP Professional, Cambridge, MA, 1994.

[3] J. M. Benjamin, N. A. Ali, and A. F. Schepis. A laser cane for the blind. In *Proceedings of the San Diego Biomedical Symposium*, volume 12, pages 53–57, 1973.

[4] M. Bicego and V. Murino. Investigating Hidden Markov Models capabilities in 2D shape classification. Submitted for publication, 2002.

[5] J. Blauert. *Spatial Hearing: the Psychophysics of Human Sound Localization*. MIT Press, Cambridge, MA, 1983.

[6] C. Brown and R. Duda. A structural model for binaural sound synthesis. *IEEE Transactions on Speech and Audio Processing*, 6(2):476–488, September 1998.

[7] C. Collins, L. Scadden, and A. Alden. Mobility studies with a tactile imaging device. In *Fourth Conf. on Systems and Devices for the disabled*, Seattle, WA, June 1977.

[8] I. J. Cox, S. Hingorani, B. M. Maggs, and S. B. Rao. A maximum likelihood stereo algorithm. *Computer Vision and Image Understanding*, 63(3):542–567, May 1996.

[9] J. D. R. Olsen and T. Nielsen. Laser pointer interaction. In *ACM SIGCHI Conf. on Human Factors in Computing Systems*, pages 17–22, Seattle, Washington, 2001.

[10] J. S. A. Dallas. Sound pattern generator. WIPO Patent Application No. WO82/00395., 1980.

[11] O. Faugeras. *Three-Dimensional Computer Vision: A Geometric Viewpoint*. MIT Press, Cambridge, MA, 1993.

[12] R. M. Fish. An audio display for the blind. *IEEE Trans. Biomed. Eng.*, 23(2):144–154, 1976.

[13] F. Fontana, D. Rocchesso, and L. Ottaviani. A structural approach to distance rendering in personal auditory displays. In *IEEE International Conference on Multimodal Interfaces (ICMI)*, Pittsburgh, PA, 2002.

[14] A. Fusiello, V. Roberto, and E. Trucco. Symmetric stereo with multiple windowing. *International Journal of Pattern Recognition and Artificial Intelligence*, 14(8):1053–1066, December 2000.

[15] L. Kay. Air sonar with acoustical display of spatial information. In R. G. Busnel and J. F. Fish, editors, *Animal Sonar System*, pages 769–816. Plenum Press, New York, 1980.

[16] M. Kubovy and D. Van Valkenburg. Auditory and visual objects. *Cognition*, 80:97–126, 2001.

[17] H. Kuttruff. *Room Acoustics*. Elsevier Science, Essex, England, 1991.

[18] T. Lokki, L. Savioja, R. Väänänen, J. Huopaniemi, and T. Takala. Creating interactive virtual auditory environments. *Computer Graphics and Applications*, 22(4):49–57, July 2002.

[19] M. P. B. L. An experimental system for auditory image representations. *IEEE Trans. Biomed. Eng.*, 39(2):112–121, February 1992.

[20] J. Rekimoto and M. Saitoh. Augmented surfaces: A spatially continuous work space for hybrid computing environments. In *Proceedings of CHI99*, pages 378–385. ACM, 1999.

[21] D. Rocchesso. Spatial effects. In U. Zölzer, editor, *Digital Audio Effects*, pages 137–200. John Wiley and Sons, Ltd., Chichester Sussex, UK, 2002.

[22] D. Rocchesso, M. Fernström, R. Bresin, and B. Moynihan. The Sounding Object. *Computer Graphics and Applications*, 22(4), July 2002. CDROM addendum, see also http://www.soundobject.org.

[23] D. A. Ross and B. B. Blasch. Evaluation of orientation interfaces for wearable computers. In *Proceedings of the Fourth International Symposium on Wearable Computers*, pages 51–68. IEEE Computer Society, October 2000.

[24] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1):7–42, May 2002.

[25] J. Strikwerda. *Finite Difference Schemes and Partial Differential Equations*. Wadsworth & Brooks, Pacific Grove, CA, 1989.

[26] S. A. Van Duyne and J. O. Smith. Physical modeling with the 2-D digital waveguide mesh. In *Proc. Int. Computer Music Conf.*, pages 40–47, Tokyo, Japan, 1993. ICMA.

[27] S. A. Van Duyne and J. O. Smith. A Simplified Approach to Modeling Dispersion Caused by Stiffness in Strings and Plates. In *Proc. Int. Computer Music Conf.*, pages 407–410, Aarhus, Denmark, September 1994. ICMA.

[28] The LAR-DEIS VIDET project. University of Bologna, Italy. Available at URL http://www.lar.deis.unibo.it/activities/videt.

[29] B. N. Walker and D. M. Lane. Psychophysical scaling of sonification mappings: A comparision of visually impaired and sighted listeners. In *Int. Conf. Auditory Display*, pages 90–94, 2001.

[30] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000.