A PHYSICS-BASED APPROACH TO THE PRESENTATION OF ACOUSTIC DEPTH

Federico Fontana and Davide Rocchesso

University of Verona
Dipartimento di Informatica
Strada Le Grazie 14
37134 Verona - Italy
fontana@sci.univr.it,davide.rocchesso@univr.it

ABSTRACT

A virtual listening environment providing localization cues is proposed for the reproduction of acoustic depth. By simulating the propagation of acoustic waves inside a tube it allows to change the source/listening point positions interactively, in a way that listeners experience various spatial configurations depending on the source/listener mutual position, and, correspondingly, perceive different auditory cues of depth. The quantitative relationship existing between physical and auditory distance assessments suggests to represent the tube using a model which allows direct control of the depth parameter. Simulations and experiments demonstrate the effectiveness of the model and its relative robustness in applications contexts where state of the art equipment and ideal listening conditions cannot be guaranteed.

1. INTRODUCTION

Researchers in sound spatialization must deal with several issues of acoustics while working on the synthesis of auditory spatial cues. Those cues can be added to a sound in the form of signal attributes, and effectively conveyed to single listeners or listening groups in proper listening conditions [1]. Indeed, the synthesis and presentation of spatial cues can be effective only once we can predict which 'auditory scene will be figured out by listeners as long as they hear those cues, either mixed together with or presented without alternative information about the scene, usually coming from vision.

In real environments auditory spatial cues add to sound during its journey from the source emission point to the ear canal entrance. In this case the spatialization process is clear: we can consider source sound as non-spatialized, and sound entering the ear canal as spatialized. Though, listeners can hear only the latter, however modified by the environment: in that sound listeners try to identify the source, segregating its cues from the rest of the information contained in the signal. The remaining cues are then reconducted to corresponding environmental characteristics, including source positions and approximate aspect of the listening environment.

The artificial manipulation and accurate synthesis of those "residual" cues is a primary concern in sound spatialization. Furthermore, if we are able to remove existing spatialization cues in the source sound and we are able to display our artificial cues correctly and transparently, then we are setting up an audio Virtual Reality (VR) installation where users will experience artificial instead of real spatialization.

The journey of a sound from the source to the listener is affected by the environment characteristics and, as long as the acoustic waves reach the listener's proximity, by the listener's own body. Environmental modifications result in the birth of echoes, whose characteristics depend on the reflection properties of the surfaces originating those echoes; also, the environment changes the sound depending on the medium where the acoustic waves propagate [2]. Subjective modifications appear in the acoustic signal entering the ear canal in the form of torso, shoulder and pinna reflections, and head diffraction [3, 4].

In this work we need to classify auditory spatial cues in those which depend on the listener's body, and those which depend on the environment: we will call the former *subjective*, the latter *objective*. Researchers have demonstrated that lateralization and elevation cues, necessary for localizing the relative angular position of a sound source, are subjective [5]. Subjective cues have been also shown to contribute significantly to the distance evaluation of nearby sources [6, 7, 5]. On the other hand objective cues mainly convey auditory spatial impressions, which have been classified with various names in the literature: *spaciousness*, *warmth*, *apparent source width*, *envelopment* and so on [8].

The differences between those two families are significant: subjective cues assessments require specific recording strategies which must take the subject's characteristics into consideration [9]; environmental cues can be just captured during a conventional recording session [10].

Distance cues are substantially environmental, once we avoid dealing with sound sources located in the listener's near-field. A significant amount of psychophysical experiments, the earliest ones dating back to the 19th century, has been conducted on distance evaluation—a comprehensive set of citations cannot be given here [5, 11]. Most of those experiments aimed at finding reliable and generalized psychophysical scales mapping the perceived distance of the auditory event in the actual source/listener distance. Today, most researchers agree in that a unique configuration for those scales does not exist [12]. In fact too many factors affect distance evaluation: those factors cannot be handled altogether at the same time by one single experiment. On the other hand, experiments which investigated the existence of low- or mono-dimensional scales, focusing on peculiar listening conditions, often came to conclusions that cannot be straightforwardly generalized to the everyday-listening experience.

It must be remembered that judgments on distance are influenced by visual cues as well [13]. Finally, subjective factors (resulting in the so-called distance *localization blur* [5]) always appear during experiments on distance evaluation.

A general conclusion which arises from most of those experiments is that three cues are particularly significant for distance perception: *loudness*, *direct-to-reverberant energy ratio* and *spectrum* [12]. This conclusion confirms that distance cues have an environmental origin and, hence, a monaural character. From this, it descends that the virtual reproduction of distance must rely on strategies which differ from modeling techniques commonly used in the synthesis of subjective cues. Subjective cue reproduction, in fact, needs models whose parameters must be individually tuned on the subject's anthropometric characteristics [3, 4].

Rather, a possible modeling strategy for the virtual localization of a sound source along *depth* seems to belong to the area of artificial reverberation [14]. It is nevertheless true that a "distance" knob is usually not found in the control panel of a commercial artificial reverberator. This follows mainly from the fact that artificial reverberators are specifically designed for musical, and not display purposes. In other words they are designed to synthesize spatial cues which are aesthetically convincing, but not necessarily informative. Besides that, spatial cues produced by artificial reverberators have a counterpart in the auditory spatial impressions which are experienced in concert halls and other contexts devoted to musical listening. Less attention has been deserved to everyday-listening contexts, where people make use of spatial cues mainly for evaluation and recognition purposes.

In modern multimodal displays, adding auditory distance as an informative parameter to the bunch of information presented to the user, in the different modalities, seems to make sense. In particular, auditory spatial cues can be naturally superimposed to audio messages that are synthesized using sonification methods [15]. In fact, sonification defines cues which are naturally associated by humans to source sounds, whereas distance information comes from (objective) spatial cues. The display, then, should be conveniently completed with (subjective) angular localization cues. Though, the third step can be more difficult to achieve due to the sensitivity of the angular (especially vertical) localization rendering models on the listener's body characteristics and on the type and quality of the reproduction equipment.

In this paper we present a spatialization model which is devoted to render the source depth, i.e., the source/listener distance. Its design strategy follows a physics-based approach. As we will see, this strategy has an important advantage: the communication between the model and the human interface is *direct*, both from the user to the machine and from the machine to the user.

After a brief explanation on how the physically-based model works, we will show that a particular realization of this model succeeds in rendering the sound source distance, with a localization error which is comparable with figures obtained during experiments on distance evaluation conducted in real environments. We will see that the environmental nature of distance cues allows a design which makes the model robust in front of non-optimal sound presentation: for example, in an office room using conventional audio reproduction equipment.

2. PHYSICS-BASED MODELING

The *perceptual* approach to artificial reverberation has been winning in the past over the *structural* one, for reasons which are mainly related with the real-time constraint [14]. On the one hand this constraint does not limit the quality of the reverberators. On the other hand, perceptually-based signal processors must be carefully tuned prior to their functioning. For this reason, they are provided with presets that limit the visibility of the control space to a set of pre-stored configurations.

Such presets act like maps between the user and the machine. They are an intermediate layer lying in between those two levels. In other words, the communication between the user and the machine is not direct, in the sense that the user cannot set the machine driving parameters (in this case filter coefficients) directly.

The existence of an intermediate layer between the user and the machine cannot be avoided in a reverberator for musical purposes. In fact, as long as it has to render musical auditory impressions such as the ones we have seen in the introduction, then proper rules mapping those impressions into filter coefficients are needed in any case.

The same layer turns out to be an undesired additional step to deal with as long as we move to the rendering of a quantitative impression such as auditory depth. With that layer we should perceptually tune the depth-to-coefficient maps based, for instance, on the evaluation by a selected group of listeners. Best would be if we had a depth knob which sets those coefficients directly, prior to the evaluation process. We can do this if we hypothesize that convenient depth cues can conveyed using a system which reproduces realistic contexts in which a source and a listener are present, one a precise distance far from from the other. At that point we have a model which reproduces a virtual scenario where acoustic depth is simulated and quantified directly.

More in general, we can think to *represent* depth instead of exactly simulating it. In fact, in principle we are not sure that the best way to convey depth in a virtual environment consists on simulating the source/listener distance as carefully as possible. This conclusion is supported by the discrepancies subjects exhibited in evaluating distance during various types of experiments conducted in real environments.

The idea of representing instead of simulating sounds is already familiar in the sonification area, and to researchers working with physically-informed models for the synthesis of ecological sounds [16, 17].

Physics-based modeling of spatial cues, so, requires to set up models that represent space. We chose to simulate 3-D pressure wave propagation along a multidimensional uniform transmission medium using a numerical scheme which turns out to be particularly suitable for this purpose: the 3-D rectangular Waveguide Mesh (WM) [18]. In fact, this scheme allows to set up a uniform pressure wave propagation domain quite easily. This domain can be seen, in first approximation, like a composition of small cubes assembled together and centered around *nodes*, each one simulating ideal and uniform pressure wave propagation as the one described by physics [2].

Wave diffraction is naturally modeled by the WM. Enclosure size can be converted into a corresponding number of mesh nodes once the propagation speed of sound and the sampling frequency of the simulation have been set. Finally, object and wall reflections are modeled by interfacing boundary nodes with so-called *Digital Waveguide Filters* (DWF) [19].

3. INTERACTION ISSUES

Before introducing the model, we want to make some considerations about the potential effectiveness of an interface providing acoustic depth as an additional information to the user.

How to present depth cues optimally? Our optimization criterion is threefold. We in fact want to consider the following aspects:

- quality of the cues conveyed by the auditory display;
- cost of the presentation;
- usability of the interface in everyday-listening contexts.

It is our opinion that the potential technological impact of a spatialization model underlies those aspects. As we briefly told in the introduction, we can never assess the performance of a sound spatialization model completely unless we are not fully aware of the way and the place where sound is presented to the listener.

In particular, the third aspect is not so straightforward to deal with when the auditory modality is added to a machine interface. For instance, the presentation of sophisticate audio VR cues usually overlaps with the user's possible need to communicate with other people, especially in cooperative contexts such as those normally experienced by groups working in the same physical environment. In that case, audio messages coming from different systems and mixing together in the same environment can degrade the performance of a work group instead of improving it.

If we consider today's available reproduction systems, we come up with two basic solutions for the presentation of sounds: loudspeakers and headphones. We have conducted experiments with our spatialization model using both presentation methods. In both cases we avoided to use specific equipment and subjective binauralization strategies. Both the loudspeaker and headphone presentations nevertheless convey convincing distance cues to listeners. In the following, we will understand why.

First, we must specify which kind of *presence* the user expects to experience [20]. To find an analogy with the visual mode, in that case it is true that a subject, put in front of a conventional computer visual interface (say, a screen), expects to discriminate nearby from far displayed objects in a context where distance is represented rather than reproduced. Despite the limits of the screen interface, this discrimination can be even accurate if proper display strategies are put into action for representing distance. On the other hand the same subject, if experiencing a more immersive virtual environment, will expect to perceive distance with more realism, perhaps the same realism he experiences in the everyday life, i.e., when stereoscopic vision is enabled. Otherwise he will rate the virtual experience as fair.

The same expectation probably have those subjects who, in front of a machine-to-user interface using conventional equipment for audio presentation, are asked to rate distance cues: as long as binaural listening is not enabled, they cannot hear but a *representation* of the auditory distance.

Once this distinction has been made clear, we still wonder if represented auditory distance works as accurately as real or realistic (i.e., using audio VR) distance perception. The answer is positive. This conclusion is confirmed by accurate psychophysical experiments, in which subjects evaluated relative distance by sound sources represented using loudspeakers [21]. In that case subjects demonstrated capabilities of evaluating auditory distance comparable with the performances they exhibited during experiments conducted in real listening contexts, that is, using real distant sound sources.

Our experiments, in this sense, move a step further: since subjects wore headphones or were put in front of loudspeakers, they were aware of listening to sound source representations; in the meantime they were given no information about the displayed scene previously to the test. Hence, the results we have obtained so far would confirm that satisfactory distance evaluation holds also in the case when the representation of a real scene is substituted with the representation of a virtual scene using our model.

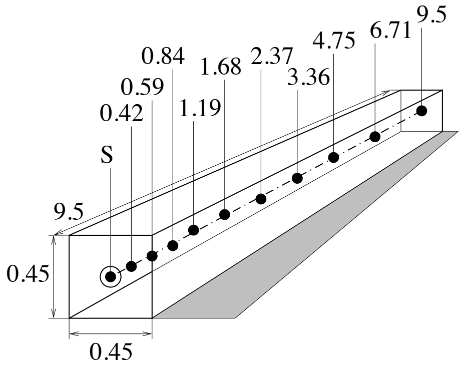

Figure 1: *Virtual tube. All sizes in meters.*

4. THE LISTENING ENVIRONMENT

We have tried to capture the essential aspects of distance perception, figuring them out in a physics-based model. The result is the tubular domain appearing in Figure 1 [22]. The tube is sized 9.5×0.45×0.45 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, to avoid the generation of audible repetitions of sounds. Sources can be located in correspondence of any node forming the WM.

The sampling frequency has important effects on the computational requirement of the model. We have set it to 8 kHz. Our choice is oriented to efficiency rather than sound realism: reliable distance cues should be conveyed also using low sampling frequencies.

Prior to any experiment, an informal listening to sounds spatialized with the virtual tube shows that such sounds will hardly have musical applications, since the tube exaggerates reverberation similarly to what happens when somebody talks to the inside of a tube. On the other hand, the same sounds seem to be quite informative about the distance of the sound source. This is especially true if we represent ecological events such as the rolling of a ball inside the tube.

Furthermore, the exaggeration of reverberant cues makes the model more robust when depth is displayed in physical environments where groups of people work together. In fact, reverberation compresses the loudness range as long as distance varies. Although loudness cues can be effective in rendering distance especially with familiar sound sources [5, 12], nevertheless excessive loudness changes at the interface output decrease the user's performance when sound is too soft, and are unpleasant for the working group when sound becomes too loud.

5. EXPERIMENT

Here only the results coming from the experiment using loudspeakers are summarized (Figure 2) [22, 23]. Our listening test was based on the magnitude estimation method without modulus, by means of which we investigated how 4 female and 6 male volunteers scaled the perceived distance. Subjects joined a normally echoic, quite but not silent office room. They were blindfolded, and were listening to an unfamiliar but not unrealistic source sound (a cowbell), which was moved along the points indicated in Figure

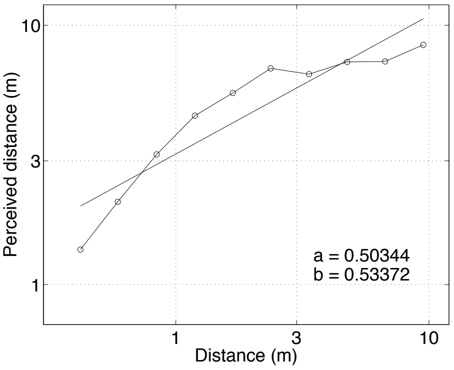

Figure 2: *Loudspeaker listening: Individual distance evaluations together with individual linear regression lines. a: intercept. b: slope (left). Average distance evaluation together with linear regression line. a: intercept. b: slope (right).*

1 (S is the sound source position). Subjects received 30 uniformly distributed stimuli, and had to rate distance after each stimulus without any prior training.

6. CONCLUSION AND FUTURE RESEARCH

Figure 2 shows that the subjects' performance in evaluating acoustic depth without any prior information about the auditory scene is good, not more blurred than evaluations coming from experiments conducted in real environments.

The virtual tube performs particularly well when ecological sounds are used which are consistent with the scenario. Its multidimensional structure even allows to render dynamic variations of the sound source position without audible artifacts. The WM is versatile enough for straightforward changes of the virtual scenario. This would enable to test other auditory scenes, where depth can be conveyed using alternative spatial structures. Audio examples are available at http://www.sci.univr.it/∼fontana.

On the other hand, the computational resources required by the WM model is still beyond (but not too far from) the power available in most actual computers, especially PC's. Hence, we are now planning the development of an efficient realization of the model.

7. REFERENCES

[1] D. R. Begault, *3-D Sound for Virtual Reality and Multimedia*, Academic Press, Boston, MA, 1994.

[2] H. Kuttruff, *Room Acoustics*, Elsevier Science, Essex, England, 1991, Third Ed.; First Ed. 1973.

[3] R. Duda and W. Martens, "Range dependence of the response of a spherical head model," *J. of the Acoustical Society of America*, vol. 104, no. 5, pp. 3048–3058, Nov. 1998.

[4] C. P. Brown and R. O. Duda, "A structural model for binaural sound synthesis," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 5, pp. 476–488, Sept. 1998.

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

[6] D. S. Brungart, "Near-field virtual audio display," *Presence*, vol. 11, no. 1, pp. 93–106, Feb. 2002.

[7] B. Shinn-Cunningham, S. Santarelli, and N. Kopco, "Tori of confusion: Binaural localization cues for sources within reach of a listener," *J. of the Acoustical Society of America*, vol. 107, no. 3, pp. 1627–1636, Mar. 2000.

[8] D. Griesinger, "The psychoacoustics of apparent source width, spaciousness and envelopment in performance spaces," *Acustica*, vol. 83, pp. 721–731, 1997.

[9] W. G. Gardner and K. Martin, "HRTF measurements of a KEMAR," *J. of the Acoustical Society of America*, vol. 97, no. 6, pp. 3907–3908, June 1995.

[10] D. D. Rife and J. Vanderkooy, "Transfer-function measurements using maximum-length sequences," *J. of the Audio Engineering Society*, vol. 37, no. 6, pp. 419–444, June 1989.

[11] D. W. Grantham, "Spatial hearing and related phenomena," in *Hearing*, B. C. J. Moore, Ed., Handbook of Perception and Cognition, chapter 9, pp. 297–345. Academic Press, San Diego, CA, 1995.

[12] P. Zahorik, "Assessing auditory distance perception using virtual acoustics," *J. of the Acoustical Society of America*, vol. 111, no. 4, pp. 1832–1846, Apr. 2002.

[13] B. R. Shelton and C. L. Searle, "The influence of vision on the absolute identification of sound-source position," *Perception & Psychophysics*, vol. 28, pp. 589–596, 1980.

[14] W. G. Gardner, "Reverberation algorithms," in *Applications of Digital Signal Processing to Audio and Acoustics*, M. Kahrs and K. Brandenburg, Eds., pp. 85–131. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.

[15] G. Kramer, *Auditory Display: Sonification, Audification, and Auditory Interfaces*, Addison-Wesley, Reading, MA, 1994.

[16] P. R. Cook, *Real Sound Synthesis for Interactive Applications*, A. K. Peters, L.T.D., 2002.

[17] D. Rocchesso, R. Bresin, and M. Fërnstrom, "Sounding objects," *IEEE Multimedia*, 2003, in press.

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

[19] 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.*, Aarhus, Denmark, Sept. 1994, ICMA, pp. 407–410.

[20] C. Hendrix and W. Barfield, "The sense of presence within auditory virtual environments," *Presence: Teleoperators and Virtual Environments*, vol. 5, no. 3, pp. 290–301, 1996.

[21] P. Zahorik, "Auditory display of sound source distance," in *Proc. Int. Conf. on Auditory Display*, Kyoto, Japan, July 2002.

[22] F. Fontana, D. Rocchesso, and L. Ottaviani, "A structural approach to distance rendering in personal auditory displays," in *Proc. International Conference on Multimodal Interfaces (ICMI'02)*, Pittsburgh, PA, Oct. 2002, pp. 33–38.

[23] L. Ottaviani, F. Fontana, and D. Rocchesso, "Recognition of distance cues from a virtual spatialization model," in *Proc. Conf. on Digital Audio Effects (DAFX-02)*, Hamburg, Germany, Sept. 2002, pp. 187–190.