

Real-time auditory-visual distance rendering for a virtual reaching task

Luca Mion*

Federico Avanzini†

Dept. of Information Engineering
University of Padova

Bruno Mantel‡

Benoit Bardy§

Motor Efficiency and Deficiency Lab.
Montpellier-1 University

Thomas A. Stoffregen¶

School of Kinesiology
University of Minnesota

Abstract

This paper reports on a study on the perception and rendering of distance in multimodal virtual environments. A model for binaural sound synthesis is discussed, and its integration in a real-time system with motion tracking and visual rendering is presented. Results from a validation experiment show that the model effectively simulates relevant auditory cues for distance perception in dynamic conditions. The model is then used in a subsequent experiment on the perception of egocentric distance. The design and preliminary result from this experiment are discussed.

CR Categories: H.5.1 [Information Interfaces and Presentation (e.g., HCI)]: Multimedia Information Systems—Artificial, augmented, and virtual realities; H.5.2 [Information Interfaces and Presentation (e.g., HCI)]: User Interfaces—Auditory (non-speech) feedback

Keywords: Multimodal interaction, 3-D sound, virtual auditory space, egocentric distance

1 Introduction

Research on multimodal perception has focused mainly on modal information and integration at some levels of the central nervous system. Nevertheless, animal-environment interactions have simultaneous consequences on several energies that stimulate our perceptual systems. These interactions provide structure not solely to individual energies but also to the way each energy varies relatively to the others [Stoffregen and Bardy 2001]. This relation across energies contains important intermodal information that is available as is any other within-modal pattern investigated so far.

This paper is part of an ongoing work finalized at investigating the effect of multisensory information in egocentric distance perception. The goal is to formalize and manipulate how distance is specified across optics, acoustics and inertia, and to test how it is perceived in a virtual environment that accurately simulates the intermodal relations between these energies. Previous studies revealed that subjects perceive distance more accurately when the intermodal

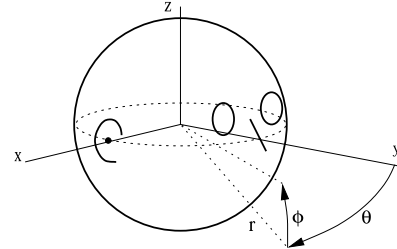


Figure 1: Polar coordinate system.

relation between optics and inertia is preserved [Mantel et al. 2005]. We are interested in extending the study to include acoustics.

Therefore the paper focuses mostly on auditory distance perception and rendering. Section 2 summarizes the static and dynamic cues that are most relevant for auditory distance perception. Section 3 presents a structural model for binaural sound synthesis. This is integrated into a real-time system as described in Sec. 4, where we also report on a pre-experiment used to validate the model, and discuss preliminary findings from the virtual reaching task experiment.

2 Cues for auditory distance rendering

Designers of 3-D audio systems know that auditory estimation of azimuth (θ) is more accurate than elevation (ϕ) estimation, and that distance (r) estimation is the most difficult task since it involves integration of multiple cues (see Fig. 1 for the meaning of these coordinates). A review is provided in [Zahorik et al. 2005].

2.1 Static cues

In the absence of other information, the *intensity* of a sound source is the primary distance cue used by listeners. Given a reference intensity and distance, the inverse square law predicts that an omnidirectional sound source's intensity will fall by 6 dB for each distance doubling. In general however it is not clear what the best model is for distance-dependent intensity scaling. As an example, it has been shown that the preferred scaling depends on the stimulus type [Begault 1991]. When *reverberation* is present the overall intensity at the ear is less dependent on distance, since intensity scaling applies only to the direct sound whereas the reflected energy remains approximately constant. The proportion of reflected to direct energy, the so-called *R/D ratio*, functions as a stronger cue for distance than intensity. A sensation of changing distance can occur if the overall intensity is constant but the R/D ratio is altered, and the apparent distance of a sound source is typically underestimated in an anechoic environment (i.e. in the absence of reverberation) [Mershon and Bowers 1979]. One can say that reverberation provides the "spatiality" that allows listeners to move from the domain of loudness inferences to that of distance inferences, i.e. from an analytic listening attitude to an *everyday listening* attitude.

Distance perception is affected by expectation or *familiarity*. If the sound source is cognitively associated with a typical distance range,

*e-mail: luca.mion@dei.unipd.it

†e-mail: avanzini@dei.unipd.it

‡e-mail: bruno.mantel@univ-montp1.fr

§e-mail: benoit.bardy@univ-montp1.fr

¶e-mail: tas@umn.edu

that range will be more easily perceived than unexpected or unfamiliar distances. This is especially true for speech [Gardner 1969]. Distance-dependent *spectral effects* also have a role. With increasing distance, higher frequencies are increasingly attenuated due to air absorption. Spectral modifications also occur in the *near field* (i.e. for distances less than ~ 1 m), where the effects of sound wave curvature must be taken into account [Brungart 2002]. As the source approaches a listener’s head, emphasis is added to lower frequencies, providing a “darkening” of tone color. An open question is whether static binaural listening improves distance perception (those discussed above are all monaural cues). The two main binaural cues are the *interaural time difference (ITD)*, which is due to the extra distance traveled by a sound wave in order to reach the farthest ear, and the *interaural level difference (ILD)*, which is due to the acoustic shadow of the head on the farthest ear. In the near field limit both the ILD and the ITD at low frequencies are emphasized, especially for very lateralized sound sources ($\theta \sim \pm\pi$). This effect is sometimes termed *auditory parallax* (although it is not a dynamic cue), and has been interpreted by some to mean that the accuracy of estimation of a sound from the side should be improved when compared to distance perception on the median plane.

2.2 Dynamic and multisensory cues

In everyday perception humans use *dynamic* cues in addition to static ones to reinforce sound localization. These arise from active, sometimes unconscious, motions of listeners relative to the source, e.g. to minimize interaural differences and estimate the direction of an incoming sound, using the head as a “pointer”. Movable pinnae in animals have the same purposes. Studies on dynamic cues date back to [Wallach 1940] and have shown that active motion can improve localization abilities. As an example, front/back confusions are very common in static listening tests: a sound source located in front of the listener at a certain θ , and a second one located at the rear, at $\pi - \theta$, provide similar static cues. Thus listeners often operate *reversals* in azimuth judgments, erroneously locating sources at the rear instead of at the front, or viceversa. Reversal essentially disappear when listeners are allowed to turn their heads during the task [Wightman and Kistler 1999]. Active motion helps especially in azimuth estimation and to a lesser extent in elevation estimation [Thurlow and Runge 1967; Perrett and Noble 1997].

Active motion also improves distance perception [Speigle and Loomis 1993] by means of two main cues: one is the motion-induced rate of change in intensity (the *acoustic τ* , the acoustic analog of the optical τ which specifies time to contact). The second one is the so-called *motion parallax*, which indicates the rate of change in angular direction resulting from listener translation: for a very close source, a small shift causes a large change in angular direction, while for a very distant source the change is almost null irrespective of the amount of shift. The rate of change of ITD, ILD (and that of spectral notches and peaks in the case of vertical motion) will therefore be affected by the distance.

There are few studies on the effect of joint auditory and visual information on distance perception, e.g. in a fully immersive virtual environment. Vision is known to be more reliable than audition in spatial location judgments, and “visual capture” is observed in many tasks (e.g., the ventriloquist effect, in which the perceived location of a sound shifts towards a visual stimulus presented at a different position). For the specific case of distance judgments, Gardner [1969] coined the term *proximity effect* after observing that in an experimental set-up with five loudspeakers at increasing distance listeners tended to perceive auditory stimuli as delivered from the closest loudspeaker, irrespective of the actual activated loudspeaker. This suggests that visual capture in the distance dimension has similarities to the angular direction capture observed in the ventriloquist

effect. However more recent studies have not confirmed these findings. No proximity effect was observed in the experiments reported in [Zahorik 2001]. With a set-up similar to the one used in [Gardner 1969], almost opposite results were found: accuracy in distance judgments increased when the loudspeaker array was visible to the listener, and visual information was found to lower judgment variability compared with the auditory-only baseline condition. These recent results provide evidence that visual capture effects in distance are not as general as supposed by previous studies.

3 A model for binaural sound synthesis

Although realistic auditory distance rendering is in general hard to achieve, the specific reaching task scenario considered in this work restricts the problem in many respects (see Sec. 4.3 for a description of the experimental design). First, subjects evaluate egocentric distance from a virtual object to reach (the sound source) which is known to be always located at the front: this eliminates the problem of front-back reversals. Second, the range of considered distances is relatively narrow, and is at the threshold of near-field conditions (~ 1 m): therefore variations in the R/D ratio also are relatively small, and familiarity effects are not as relevant as in generic conditions. Third, users move mainly within the horizontal plane while there is little movement in the vertical direction: we can therefore speculate that monaural spectral effects of pinnae and torso are not as relevant as in generic conditions. Based on these considerations we expect the most salient auditory cues in the virtual reaching task to be dynamic ones (acoustic τ and motion parallax). Accordingly, we will use a relatively simple binaural sound synthesis model, which nonetheless simulates these cues accurately.

The model includes three components: distance-dependent intensity scaling, delay/shadow head effects, and spectral effects of pinnae and torso. Assuming an ideal omnidirectional point source, and given a reference sound pressure level at 1 m, we use the inverse square law and simulate a decrease in intensity of 6 dB for each distance doubling. Head effects are simulated using a spherical head model. Consider a sphere of radius a , a point sound source at azimuth θ and distance $r > a$ from the sphere center, and a point (the ear) located at θ_{ear} on the sphere. Then the diffraction of an acoustic wave by the sphere seen on the chosen point is expressed with the following transfer function [Duda and Martens 1998]

$$H(\rho, \mu, \theta) = -\frac{\rho e^{-j\mu\rho}}{\mu} \sum_{m=0}^{+\infty} (2m+1) P_m(\cos(\theta - \theta_{ear})) \frac{h_m(\mu\rho)}{h'_m(\mu)}, \quad (1)$$

where $\rho = r/a$ is the normalized distance and $\mu = \omega a/c$ is the normalized frequency (c is the speed of sound), P_m and h_m are the m^{th} order Legendre polynomial and spherical Hankel function, respectively. The angle $\theta - \theta_{ear}$ is the angle of incidence on the ear. It has been shown [Brown and Duda 1998] that in the limit of large relative distances the transfer function $H(\infty, \mu, \alpha)$ can be approximated with a first-order head-shadow filter H_{HS} of the form

$$H_{HS}(\mu, \theta) = \frac{1 + \frac{1}{2}j\mu \cdot \alpha(\theta - \theta_{ear})}{1 + \frac{1}{2}j\mu}, \quad (2)$$

where the coefficient $\alpha(\theta - \theta_{ear})$ controls the location of the zero. With an appropriate parametrization of α , the filter H_{HS} produces a reasonable approximation of the theoretical response (1). In order to simulate near-field effects, Eq. (1) may be evaluated directly without resorting to the approximation (2). Note however that numerical evaluation of the spherical Hankel functions increases the computational costs considerably [Duda and Martens 1998].

The effects of pinnae and torso are taken into account by simulating reflections of the direct sound on these elements, through transfer

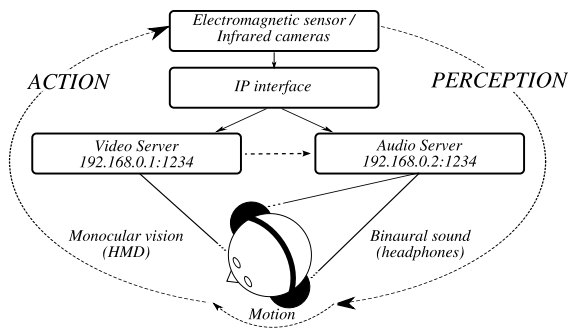


Figure 2: The system.

functions of the form $H_R(\theta, \phi) = b_R e^{-j\mu T_R(\theta - \theta_{ear}, \phi)}$, where b_R are absorption coefficients and the delays T_R of the reflected rays are functions of both azimuth and elevation. Torso reflections operate in parallel with the head shadow filter while pinnae reflections operate in series on both the head-shadowed direct sound and the rays reflected from the torso. The final structure of the model is identical to the one proposed in [Brown and Duda 1998].

4 The system

4.1 Real-time realization

The model has been implemented as a plug-in for the real-time sound synthesis environment PD (Pure Data).¹ To simulate a static virtual target at different locations we implemented a low-latency multimodal rendering system based on network communication (see Fig. 2). Head position and orientation are captured by a real-time motion tracking system that sends coordinates via network interface. Video/audio servers receive the controls via socket to drive the rendering models, displayed by means of a head-mounted display (HMD) and insulated headphones, respectively.

We tested two real-time motion tracking systems, based on an electromagnetic sensor and on infrared cameras, respectively. The first system is the Ascension Technology’s Flock of Birds,² which provides 6-DOF tracking at a sampling rate of 100 Hz, spatial resolution 1 mm, and angular resolution $< 1^\circ$. The second one is the eMotion SMART,³ which uses markers and 6 cameras with IR light strobes plus a synchronization unit, and provides 3-DOF marker tracking at a sampling rate of 120 Hz. The tracking system is connected to a PC through a RS232/serial interface. The main application (developed in C++) retrieves captured position and orientation and converts them to appropriate metrics and reference frame (head centered). Then data are simultaneously sent to the audio server and to the OpenGL graphic rendering application, that applies the recorded head motion to the virtual camera. With such a design, the target can be virtually located (both optically and acoustically) at any distance and along any direction. Subject’s active exploration allows for closing the action-perception loop, since the optical/acoustic flow of information is generated by his/her movement.

The total latency from data capture to optical and auditory display is mostly due to the intrinsic latency of the motion tracking system. For the Flock of Bird this lies between 60 and 70 ms (in part because of its built-in filters), while for the SMART system this is approximately 8-9 ms. The additional latency due to network

¹<http://puredata.info>

²<http://www.ascension-tech.com>

³<http://www.emotion3d.com>

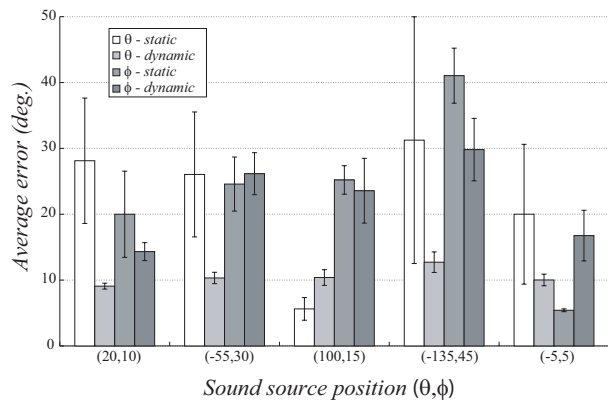


Figure 3: Average (across subjects) absolute error estimation for azimuth and elevation, in static and dynamic conditions.

communication is negligible (~ 0.1 ms in the configurations used). The intermodal latency is in most part due to the latency of the audio rendering engine: with an audio sampling rate of 44.1 kHz and an audio buffer length of 64 samples this latency is ~ 1.45 ms.

4.2 Experimental validation

Preliminary tests on the sound model were conducted at the Bio-engineering Lab at DEI, using the SMART system. Eight seated subjects were asked to judge the direction of virtual sound sources. The stimuli were sequences of 150 ms pulses of white noise, separated by 350 ms intervals of silence, all located at 1 m from the subject’s head center. A reference source intensity of 60 dB-SPL at 1 m was chosen. Reverberation was also added to simulate the characteristics of a real small-sized room. Stimuli were presented through headphones (AKG-K501, efficiency=94 dB-SPL/mW) with three markers applied (one on the top and two at the sides). The laboratory was kept silent throughout the experiments in order to avoid external acoustic disturbances. No visual feedback was provided.

The virtual sound sources were presented in random order using two conditions: passive playback and active movement. In the first condition, subjects were asked to mark on a grid the perceived direction of the sound source once the sound was stopped. In the second one, they had to move their head to face the virtual source. We recorded the trajectories of the markers during each task. The audio engine was running on a laptop (Pentium M725 @ 1,6 GHz), connected via LAN to a server PC (Pentium 4 @ 2,4 GHz) that processes the position data stream.

In a post-experimental interview subjects confirmed that there was no perceived latency of the sound rendering with respect to motion. Results in Fig. 3 show that active exploration improves localization especially for azimuth, which is in agreement with literature. Error estimation for θ drops in average by 52.7% in dynamic conditions with respect to static conditions (variance decreases by 90%), while for ϕ it drops in average by 4.84% (variance decreases by 5%). Average errors across the trials are 10.51° for θ and 22.14° for ϕ . Participants were more confident on their judgment in dynamic conditions, as confirmed by the low variance for errors (0.99° for θ and 3.61° for ϕ). The only exception is stimulus 3, for which performance in static conditions is very good, probably because the sound is completely lateralized. Finally, reversals occurred in $\sim 27\%$ of judgements in static conditions, while they disappeared in dynamic conditions. This is also in agreement with literature.

These results, and specifically the performance improvements in az-

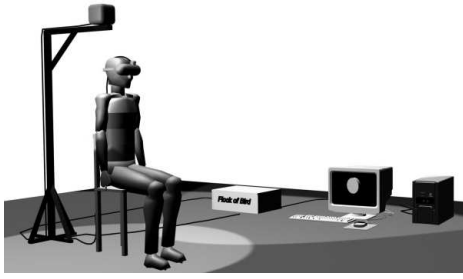


Figure 4: *Experimental setup for the virtual reaching task.*

imuth location, confirm that the relatively simple structural model used in this work is effective in simulating especially ILD and ITD in dynamic conditions. Therefore we expect it to effectively simulate motion parallax effects in the reaching task described next.

4.3 A virtual reaching task

To evaluate more precisely perceptual sensitivity to the model, we are conducting an experiment in which 16 participants have to judge verbally (yes/no) whether a simulated object is within reach. Participants are seated in a dark room, wearing a HMD (monocular viewing) and insulated headphones. The Flock of Birds is used for motion tracking, with the sensor attached at the top of the HMD. Processing of position data stream and visual rendering run on a server PC (Pentium 4 @ 2,8 GHz), while the audio engine is running on a laptop (Pentium M725 @ 1,6 GHz).

The target is displayed at 18 different distances, ranging from .28 to 1.81 in proportion of each participant actual reaching boundary (.09 increments), with 4 trials for each distance. The (optical) angular size of the target and its sound intensity at subject's ears are kept constant among trials. We used 6 different conditions through combination of two factors: feedback ($O/A/O+A$) \times motion (AM/PP), where O,A stand for Optics and Acoustics, and AM,PP stand for Active Motion and Passive Playback. Hence in certain conditions the target can only be seen, while in others it can only be heard or be both seen and heard. At the same time, in some conditions participants are allowed to move and explore the virtual scene while in others they have to remain still while being shown the optical and/or acoustical consequences of their movement, recorded in earlier trials. We expect participants to be precise and accurate in all conditions involving active movements.

Preliminary results suggest that the performance (i.e., ability to perceive whether the target is reachable or not) of participants is good in all conditions where they are allowed to move. More surprisingly the performance exhibited in A condition is very similar to those obtained in O or $O+A$ conditions. Full results will be published in a forthcoming dedicated paper.

5 Conclusion

We have presented a real-time realization of a structural model for binaural sound synthesis, its integration into a motion tracking system and synchronization with visual rendering. The system is being used in a study on the perception and rendering of distance in multimodal virtual environments. Results from the pre-experiment discussed in Sec. 4.2 have shown that the sound model effectively simulates relevant auditory cues for distance perception in dynamic conditions. We have then presented the design and preliminary result from an experiment on the perception of egocentric distance. These suggest that the performance of participants is good with any

combination of feedback, provided that they are allowed to move.

Ongoing work is devoted to the completion of the experimental sessions and analysis of the results of the virtual reaching task. As for the sound model, work is focusing on developing a real-time realization which includes near-field effects: these are expected to improve the simulation of auditory cues relevant to the task.

Acknowledgments

This research was supported by the EU FP6 NoE "Enactive Interfaces" IST-1-002114. We thank Prof. Claudio Cobelli and the Bioengineering Laboratory at DEI for hosting part of the experiment.

References

- BEGAULT, D. 1991. Preferred sound intensity increase for sensation of half distance. *Perc. and Motor Skills* 72, 1019–1029.
- BROWN, C. P., AND DUDA, R. O. 1998. A structural model for binaural sound synthesis. *IEEE Trans. Speech Audio Process.* 6, 5, 476–488.
- BRUNGART, D. S. 2002. Near-field virtual audio displays. *Presence: Teleoperators and Virtual Environment* 11, 1, 93–106.
- DUDA, R. O., AND MARTENS, W. L. 1998. Range dependence of the response of a spherical head model. *J. Acoust. Soc. Am.* 104, 5, 3048–3058.
- GARDNER, M. B. 1969. Distance estimation of 0° or apparent 0° -oriented speech signals in anechoic space. *J. Acoust. Soc. Am.* 45, 1, 47–53.
- MANTEL, B., BARDY, B. G., AND STOFFREGEN, T. A. 2005. Intermodal specification of egocentric distance in a target reaching task. In *Studies in Perception and Action VIII*, H. Heft and K. L. Marsh, Eds. Erlbaum, Mahwah, 173–176.
- MERSHON, D. H., AND BOWERS, J. N. 1979. Absolute and relative cues for the auditory perception of egocentric distance. *Perception* 8, 3, 311–322.
- PERRETT, S., AND NOBLE, W. 1997. The effect of head rotations on vertical plane sound localization. *J. Acoust. Soc. Am.* 102, 4, 2325–2332.
- SPEIGLE, J. M., AND LOOMIS, J. M. 1993. Auditory distance perception by translating observers. In *Proc. IEEE Symposium on Research Frontiers in Virtual Reality*, 92–99.
- STOFFREGEN, T. A., AND BARDY, B. G. 2001. On specification and the senses. *Behavioral and Brain Sciences* 24, 2, 195–213.
- THURLOW, W. R., AND RUNGE, P. S. 1967. Effect of induced head movements on localization of direction of sounds. *J. Acoust. Soc. Am.* 42, 2, 480–488.
- WALLACH, H. 1940. The role of head movement and vestibular and visual cues in sound localization. *J. Experimental Psychology* 27, 339–368.
- WIGHTMAN, F. L., AND KISTLER, D. J. 1999. Resolution of front-back ambiguity in spatial hearing by listener and source movement. *J. Acoust. Soc. Am.* 105, 5, 2841–2853.
- ZAHORIK, P., BRUNGART, D. S., AND BRONKHORST, A. 2005. Auditory distance perception in humans: a summary of past and present research. *Acta Acustica - Acustica* 91, 3, 409–420.
- ZAHORIK, P. 2001. Estimating sound source distance with and without vision. *Optometry and Vision Science* 78, 5, 270–275.