Simulating a virtual target in depth through the invariant relation between optics, acoustics and inertias

B. Mantel[1], L. Mion[2], B. Bardy[1], F. Avanzini[2] & T. Stoffregen[3]
[1]Motor Efficiency & Deficiency Lab, Montpellier 1 University, France
[2]Department of Information Engineering, University of Padova, Italy
[3]Human Factor Research Lab, University of Minnesota, U.S.A.

Research on multimodal perception (e.g., of objects in depth) has focused mainly on modal information (e.g., cues) and their integration at some levels of the central nervous system. Nevertheless, animal-environment interactions have simultaneous consequences on several energies (not to say all) that stimulate our perceptual systems. Thus these interactions provide structure not solely to individual energies but also to the way each energy varies relatively to the others (Stoffregen & Bardy, 2001). This relation *across energies* contains important intermodal information that is available as is any other within-modal pattern investigated so far.

We decided to test that intermodal specification hypothesis in the context of egocentric distance perception. The idea is (i) to formalize and manipulate how distance is specified across optics, acoustics and inertias and (ii) to test whether it is perceived and used by humans.

Optical, acoustical and inertial dimensions

In monocular viewing, the perceiver has no access to intermodal information such as stereopsy or convergence. If he/she is looking at a virtual object through a head mounted display (HMD), he/she cannot rely on accommodation. If the object is unknown and does not look familiar, there is no previous knowledge of its size. If the object is shown against a black background, there is no distance information provided by the line of the horizon. Nor there is aerial perspective if the 3D display model does not include lightning. Even motion parallax (or more exactly the relative apparent motion of the static target generated by the perceiver's movements) does not provide, on its own, any information about egocentric distance, as it yields only angles (and angular velocity and acceleration). On the other hand, there is an invariant relation across optical and inertial energies that specify egocentric distance. In short, the optical apparent angular movements of the target are scaled in terms of mechanical and inertial consequences of the movement of the head (and vice versa).

Previous studies with such a design have revealed that perceivers do perceive accurately the distance only when the intermodal relation between energies is preserved (Mantel, Bardy & Stoffregen, 2005). Moreover, introducing a gain in the intermodal relation shifts the perceived distance accordingly. When subjects were passively shown optical consequences of their movements played-back, they were unable to perform the task, confirming that parallax is not sufficient on its own.

In everyday life, when humans interact with the environment, important intermodal information is supplied by the auditory modality. Head motion is exploited to generate the aural information for the evaluation of the spatial position of sound sources (Wallach, 1940). The task of evaluating the sound direction is accomplished by integrating cues for the perception of azimuth (i.e., angular direction in the horizontal plane) with the spectral changes that occur with head movements creating the perception of elevation (i.e., angular direction in the vertical plane). All these auditory cues are produced by physical effects of the diffraction of sound waves by the human torso, by the shoulders, the head and outer ears (pinnae), which modify the spectrum of the sound that reaches the ear drums (Blauert, 1997). In particular, in the case of sources within 1 m of the head, distance perception is affected by additional range-dependent effects and binaural distance cues arise which are not present for more distant sources. Also, an auditory motion parallax effect results from range-dependent azimuth changes, so that for close sources a small shift causes a large azimuth change, while for distant sources there is almost no azimuth change. All these cues can be captured by the head-related transfer function (HRTF) that describes how a given sound wave is filtered by the physical elements of the human body before the sound reaches the eardrum. Finally, effects of a reverberant environment can also play a relevant role: early reflections and the change in proportion of reflected to direct energy are important cues especially for distance estimation.

Virtual set-up

To simulate a static virtual target at a particular distance in front of the observer (that is, beyond the two LCD screens of the HMD), head position and orientation are captured with a 6 degrees of freedom electromagnetic sensor (Ascension Technology's Flock of Bird) and used to drive in real time the display of the target as well as its resulting sound. With such a design, the target can be virtually located (both optically and acoustically) at any distance and along any direction. The Flock of Bird is running at 100 Hz and has an intrinsic latency of 60 to 70 ms (in part because of its built-in filters). The 3D display of the target is achieved using OpenGL (under C++) by applying the recorded head motion to the virtual camera.

To produce convincing auditory azimuth and elevation effects, HRTF should be measured and tuned separately for each single listener, which is both time consuming and expensive to implement. For the above reasons, we used a simplified model (Brown, 1998) which assumes an approximate spherical head for synthesizing binaural sound from a monaural source. The components of the model have a one-to-one correspondence with the shoulders, head, and pinnae, with each component accounting for a different temporal feature of the impulse response. The model is parametrized to allow for individual variations in size of the head, and a reverberation section is implemented to simulate the reverberant characteristics of a real room acoustically. This model allows a flexible implementation using the real-time synthesis environment PD (Pure Data)1, so that the sound generation is performed externally of the visual simulation. Open Sound Control (OSC) formatted messages are sent to PD via UDP sockets, packing six coordinates of the head (3 translational and 3 angular) in the egocentric reference system. The latency of the system is negligible since the network communication runs with about 0.1 ms delay, while the latency of the audio engine is of about 1.45 ms which is well below the latency of the motion tracking system.

We will use this set-up to test whether humans can judge the *reachability* of a static virtual target that can be seen, that can be heard or that can be both seen and heard. We expect that only perceivers that are allowed to move relative to the target will be able to perform the task, because egocentric distance will only be specified in the invariant relation between optics and inertias or acoustics and inertias (and of course between optics, acoustics and inertias).

Mantel, B., Bardy, B.G. & Stoffregen, T.A. (2005). Intermodal specification of egocentric distance in a target reaching task. In H. Heft & K. L. Marsh (Eds.), Studies in Perception and Action VIII, (pp. 173-176). Mahwah, NJ: Lawrence Erlbaum Associates, Inc., 2005.

Stoffregen, T.A. & Bardy, B.G. (2001). On specification and the senses. Behavioral and Brain Sciences, 24(2), 195-261.

Wallach, H. (1940). The role of head movements and vestibular and visual cues in sound localization. Journal of Experimental Psychology, 27, 339–368.

Blauert, J. P. (1997). Spatial Hearing, rev. ed. Cambridge, MA: MIT Press.

Brown, C.P. & Duda, R.O. (1998). A structural model for binaural sound synthesis. IEEE Trans. Speech and Audio Processing, Vol. 6, No. 5, pp. 476-488.

[1] http://puredata.info/