The Space Between Your Ears:
Auditory Spatial Perception and Virtual Reality

Stephen Garofano
Student Number: 594466
Berlin School of Mind & Brain
Humboldt Universität zu Berlin

Course: Spatial Navigation
Instructor: Sophia Rekers

25 March, 2019

Abstract

The auditory system is a key sensory modality for human spatial perception. By collecting and decoding sound waves, the human brain can deduce spatial relationships among objects, locations and other biological actors. How does the auditory system perform spatial discernment and how does the brain extract valuable spatial information from the chaotic sonic environment? This paper will provide a broad survey of the physiology and cognition of spatial auditory perception. Further, it will apply this information to the practice of creating auditory spatial environments in virtual reality. Directions for future research and development will be touched upon, with attention to therapeutic virtual spatial audio paradigms for the visually impaired.

At a casual thought, many people think of their phenomenological experience as a human in visual terms. However, the auditory system contributes significantly to a human's ability to perceive and react to exogenous stimuli while navigating through the spatial environment. In fact, the human auditory system is extremely sensitive. Receptors in the auditory system respond on a timescale 1000 times faster than visual photoreceptors and detect vibrations in the air with wavelengths no more than the diameter of an atom (Purves et al., 2012). However, much of this sensitivity occurs beneath the level of our conscious awareness.

How do the human ear and brain collaborate to process auditory stimuli and how does this collaboration contribute to humans' ability to navigate their spatial environment? This paper will provide a broad survey of human auditory spatial processing in terms of physiology and cognition. Additionally, the means by which the physiological and cognitive processes of audition can be harnessed to create realistic virtual reality paradigms for the purpose of neuroscientifc research will be examined. Finally, implications for future research will be discussed.

Human Spatial Sound Perception: Waves in the Air

The system of human hearing involves the detection and decoding of vibrations traveling through the air. The phenomena we call 'sound' refers to the brain's perception of pressure waves generated by vibrating objects in our surroundings (Begault, 2000). These pressure waves are comprised of alternating fields of compression and rarefaction of air molecules propagating in three dimensions from a leaf buffeted by wind on a tree branch, the exhaust forced out of an automobile tailpipe or unfortunately, the vocal chords of your neighbor's infant. From this pattern of pressure waves, humans are able to detect an astonishing array of information about their environment.

In the case of the squalling infant, air forced from its lungs moves over the tight bands of tissue in its throat causing them to stretch until they snap back. This stretching and contracting happens many times per second and in so doing, creates moments where air molecules are compressed, followed by moments of reduced air density. This back and forth of increased and decreased air compression can be described as an oscillation or waveform. Physics quantifies wave phenomena in many ways, but most salient for a discussion of human hearing are the

characteristics of amplitude, frequency and phase. In order to determine the source of a sound or create a conception of the size of a room based on its echoic reverberations, the human auditory system must first decompose the complex jumble of sonic stimuli in our acoustic environment into our sensory percepts, in a process called psychoacoustics (Begault, 2000).

Relevant to hearing, amplitude describes the intensity of sonic air pressure and corresponds to our experience of loudness. The greater the intensity of air pressure as the sonic waveform crests, the louder our perception of the sound. Human hearing is extremely sensitive to changes in pressure and at the lower threshold of hearing, air molecules vibrate on the minute scale of picometers (10^{-11} m; Purves et al., 2012) and the range of detection extends from .00002 Pascals to over 100 (Ward, 2010). That's an impressive 7 orders of magnitude.

Frequency describes the time required for the entirety of one pressure wave to pass a fixed point in space. Humans experience the phenomenon of waveform frequency as pitch; the faster an object vibrates, the more pressure waves it produces per second, creating the perception of a higher toned sound. Pitch is measured in Hertz or vibrations per second and the range of human hearing extends from approximately 20 Hz to 20,000 Hz (or 20 kHz) though sound sources both above and beneath that range can be detected by vibration in the body (Begault, 2000). The sound of the crying baby is perceived as a higher pitch than the sound of a crying adult male neuroscience student because the baby's smaller vocal cords are vibrating at a faster rate.

The phenomenon of phase in sound is somewhat less intuitive. If two identical waveforms encounter each other with such timing that the peak of one occurs while the other is at its trough, the waves are said to be 'out of phase' and will completely nullify each other. If these waves are exactly 'in phase' such that their peaks and troughs occur simultaneously, they will constructively interact, resulting in a wave of identical frequency but doubled amplitude. However, phase relationships are not binary and can involve very minute differences in the timing of waveforms. The human brain has an incredible capacity to perceive such phase differences in sonic waveforms and use this information to determine, among other things, the directionality of a sound source (Begault, 2000).

Human Spatial Sound Perception: In Your Ear

Prior to discerning minute differences in sound waves, the brain must first collect these vibrations. This is accomplished by the external ear, consisting of the familiar visible structures of the outer ear, the concha and pinna, as well as the external auditory meatus, more familiarly known as the ear canal. The quizzical shape of these structures serves an important purpose as they selectively boost frequencies in the range of 2 to 5 kHz (Begault, 2000). Interestingly, human speech contains important components in this frequency band and the amplification of this spectrum by the external ear makes damage from loud, broad-band noise especially impactful to the discernment of speech (Purves et al., 2012).

An additional function of the shape of the external ear is that its vertical asymmetry provides a means of discerning the elevation of sound sources. The pinna and concha conduct more high frequency vibration from sounds emanating above ear-level than from below (Purves et al., 2012). Further, the front-to-back asymmetry of the external ear provides a means of determining if a sound source is in front or behind an individual. The pinnae cast a sonic 'shadow' such that sounds emanating from behind will be obstructed and thus quieter ("Oculus/Developers," n.d.).

Once within the ear canal, sound pressure waves impinge upon the tympanic membrane, (eardrum) whose vibration is passed along to the ossicles or small bones of the middle ear, the malleus, incus and stapes. The ossicles then transfer the vibration to the cochlea, the spiraling, fluid-filled structure in which pressure waves are converted to neural electrical impulse. At each phase of this vibrational transference, from the air around the head to the hair cells within the cochlea, a process of amplification and tuning is performed (Purves et al., 2012). In this we see an interesting example of the concept of embodied cognition, wherein the curious shape of one's outer ear, seemingly so unintelligent that one might pierce it with metal ornamentation, is actually performing calculations which inform one's awareness of the external environment.

Within the cochlea, further amplification and frequency separation is performed. The spiraling geometry of the cochlea transfers acoustic energy in such a way that various frequencies are focused at particular points along the basilar membrane within. The apex or innermost end of the basilar membrane inside the coiled cochlea is mechanically 'tuned' to maximize low frequencies, whereas its base, near the cochlear opening, maximizes high frequency vibration (Purves et al., 2012). Through this 'tonotopic' organization of the cochlea, a

complex sound consisting of many spectral components is disassembled, much as a prism does with light, into its compositional frequencies before even reaching a neuron. Such a feat, if performed mathematically, requires challenging computations like the Fourier transform (Begault, 2000). Thus, one might consider their ears to be rather mathematically gifted.

Inside the cochlea and along the basilar membrane, hair cells convert vibration into neuroelectric signals. Vibrational movement of the basilar membrane bends the hair cell, causing it to depolarize and release neurotransmitters. This process is remarkably fast, occurring in as little as 10 microseconds, which is necessary if the brain is to compute sound source localization. All hair cells are able to transduce sounds by moving in-phase with the pressure waveform up to frequencies of 3 kHz. This phase-locked action provides temporal information which can be used in higher cortical areas to localize sounds in auditory space. For frequencies above 3 kHz, information is conveyed through the tonotopic organization of the basilar membrane such that hair cells at the base of the cochlea are 'tuned' or preferentially responding to higher frequencies, though without phase-locking (Purves et al., 2012).

Human Spatial Sound Perception: In Your Brain

Once sound is transduced to neuroelectrical signal and transmitted along the auditory nerve, the tasks of spectral decomposition, feature extraction and source localization fall to the brain alone. An impressive amount of processing has been performed on the sound before it becomes neural impulse, but it is within the brainstem, the midbrain, the auditory thalamus and the auditory cortex that the most complex processes of sonic decoding occur. In the four to five synapses between the cochlea and the primary auditory cortex, the neural pathways are complex. Tonotopic organization is preserved such that individual frequency components are carried along their own dedicated nerve fibers. Additionally, the ascending auditory pathway has a high degree of bilateral connectivity. As such, damage to central auditory cortices rarely manifests as monaural hearing loss (Purves et al., 2012).

The cochlear nuclei of the brainstem begin the process of computing sound localization based on differences in signals communicated by the left and right ears. These interaural differences consist of variation in phase (timing) and amplitude (loudness) of the same sound detected at each ear. Phase-locked signals can be produced in the cochlea for frequencies below 3 kHz and so interaural time differences are used to compute localization in this band. For

localizing frequencies above 3 kHz, the interaural amplitude difference is calculated along a parallel pathway. The sensitivity of the brainstem to interaural time difference is significant, with a resolution as fine as 10 microseconds, providing a resolution of localization within 1 degree on the horizontal plane (Purves et al., 2012).

Given the 3 kHz threshold above which phase-locked signals become impossible, it is interesting to note that around 2 kHz, the human head begins to cast an acoustic 'shadow' as wavelengths become too short to bend around it ("Oculus/Developers," n.d.). Thus, a sound source on the right side of the body will be detected as less intense by the left ear. Also in the brainstem, nuclei responding preferentially to frequency spectra created by the vertically asymmetric shape of the external ear compute the vertical elevation of sound sources (Begault, 2000).

As the auditory pathway ascends to the midbrain, integrative processing of signals occurs such that binaural inputs are combined to create a topographical representation of auditory space. Given that frequency, not space is onto auditory receptors, the brain must synthesize this construct beginning in the inferior colliculus of the midbrain. Also in the midbrain, neurons which fire preferentially for sounds of specific duration or frequency comprise the beginnings of feature extraction from the sound stream. This selectivity is continued in the auditory thalamus where a convergence of signal streams from lower nuclei are further filtered based on spectral and temporal characteristics (Purves et al., 2012).

The final destination for signals in the auditory system is the primary auditory cortex located in Heschel's gyrus in the temporal lobes. Signals reaching the auditory cortex have been processed in a number of ways in their ascent but upon reaching A1, tonotopic organization has been preserved. As with the visual system, some evidence exists for hierarchical processing within the auditory cortex where basic auditory features are extracted in early cortices with later regions processing more complex information (Ward, 2010).

Though functional activity of auditory cortical areas is less understood than that of nuclei lower in the auditory pathway, extraction of higher order features, like those comprising speech and music, are performed in the auditory cortex. Neurons in the core and belt regions of the auditory cortex are frequency-selective and respond preferentially based on interaural phase and intensity differences, suggesting that sound source localization is completed in cortical regions. Further, there is support for a two-stream processing scheme beginning in auditory cortices in

which a dorsal stream leading to structures in the parietal lobes processes spatial localization and a ventral stream within the temporal lobe uses spectral characteristics to identify the nature of a sound's origin (Ward, 2010).

When considering cortical functions in the human auditory system, it may be of particular interest to enthusiasts of spatial navigation, whose hearts hold the hippocampus dear for its uniquely specialized components such as the place cell, the grid cell or the border cell, that the hippocampus is active in hearing as well. Given exciting recent hypotheses that hippocampal place and grid cells may create a flexible coding for spatial representations of cognitive constructs (Bellmund et al., 2018) and well-known research on hippocampal volume increase as taxi drivers learn the streets of London (Maguire et al., 2000), it may (or may not) come as a surprise that piano tuners' hippocampi are similarly swelling.

Teki and colleagues (2012) conducted structural scans of the brains of professional piano tuners versus controls, taking into account individuals' age and time spent in the tuning profession. Piano tuning requires extremely fine discernment of phase relationships between sonic frequencies as notes of the piano are adjusted relative to each other. The researchers found that grey matter volume in the anterior hippocampus and parahippocampal gyrus and white matter volume in the posterior hippocampus strikingly increased with greater piano tuning experience (Teki et al, 2012).

The authors posit a spatial construct in the act of piano tuning as the tuner must navigate (in the auditory domain) a spectral topography. In this conception, the pitch of the tuning fork is the first 'landmark' and from this frequency, the tuner navigates among tuned and un-tuned keys, using their spectral relationships as waypoints to be referenced and moved until all keys are in their proper frequency positions (Teki et al, 2012). These findings, taken in the context of the wide repertoire of hippocampal function and the Bellmund (2018) hypothesis of variable representation of cognitive spaces by hippocampal cells, point to a remarkable flexibility and power of processing within this structure.

Another cortical structure with important contribution to human hearing and spatial localization of sound is the planum temporale. Located posterior to the primary auditory cortex this structure is cited as the substrate for a construct referred to as the head-related transfer function (HRTF). The HRTF is a model created by the brain to characterize the individual shapes and contours of the head and ears and the distortions those shapes cause in impinging

sound waves. The planum temporale uses this model to interpret incoming sounds and infer the directionality of the source (Ward, 2010).

A NASA-funded study by Wenzel, Arruda, Kistler, & Wightman (1993) indicated that the brain's HRTF model of head and ear shape is so specific to the individual that subjects wearing in-ear headphones (nullifying the direction-finding of their pinnae) and exposed to recordings made with microphones placed inside the ear canals of a generic dummy head experienced confusion in determining if sounds originated from in front or behind them. Begault (2000) compares the individual uniqueness involved in this phenomenon to an acoustic fingerprint.

The planum temporale integrates a variety of sensory modalities in combination with the HRTF, including egocentric and allocentric codes of space. Vestibular information must be integrated if the brain is to correctly judge the spatial origins of sound (Ward, 2010). While this poses a computational challenge for the brain, it can also be used to refine spatial localizations. By turning the head or the body slightly in the vertical or horizontal plane while tracking a sound source, the brain can evaluate the changing sonic stimuli and compare them with the HRTF and vestibular information to pinpoint sound origins (Begault, 2000).

Spatial Sound in Virtual Reality

Given the sophisticated and varied ways that the human auditory system determines the spatial origin of sound sources, designers of virtual reality environments are challenged to devise methods for tricking the ears and brain. However, in this pursuit, VR developers are aided by powerful cognitive inclinations to deduce spatial relationships from auditory information. In an early review of NASA's work on spatial audio applications, Begault (2000) noted that humans easily bind the inaccurate sound emanating from a small television speaker to the image of a person's mouth moving on the screen. In many ways, our ears follow our eyes.

In order to deliver convincing spatial sound, a VR system must combine both head-phones and head tracking in 3 dimensions for both position and orientation. In this way the precise orientation of the ears is known to the system and correct combination of sonic stimuli can be delivered ("Oculus/Developers," n.d.). Given the sensitivity and accuracy of human hearing, effective virtual sound requires both high accuracy in head tracking and clever sound design in order to avoid smearing and spatial confusion.

As we have seen, interaural differences in timing and intensity of sound are computed by the brain for lateral and vertical localization of sound. Much of this is accomplished by the pinnae, which are inactive when one is wearing headphones. For this reason, successful sound design in virtual reality requires that developers compute a head related transfer function. Creating an ideal mathematical model of the way individual head and ear shape systematically distort incoming sounds requires placing an individual in an anechoic chamber (a space without reverberation), setting small microphones within the external ear and recording broadband sound from every direction. By subtracting the source sound from the recording at the ear canal, the HRTF can be computed (Begault, 2000; "Oculus/Developers," n.d.).

However, this is obviously impossible for every individual who would like to experience virtual reality. As a substitute, sound designers use an averaged model of the human head with microphones mounted in its ear canals and perform the process described above. This method is imperfect but functional, and a variety of publicly available HRTF data sets allow for VR sound design by those without access to an anechoic chamber. For the purpose of designing virtual environments for neuroscience research, it is important that designers are aware of the resolution of their HRTF, as a sparsely sampled data set results in unconvincing sound ("Oculus/Developers," n.d.).

Given a high resolution HRTF data set and accurate tracking of head position and orientation, a VR system can adjust the intensity and timing of sounds played through the headphones to mimic the sonic shadow cast by the head and ears and creating the sensation of lateral, and to some extent, vertical orientation of sound sources. In order to create the perception of distance from a sound source, VR sound designers manipulate the sonic features the brain naturally uses to determine this spatial relationship: loudness, reverberation and high frequency attenuation ("Oculus/Developers," n.d.).

That sounds emanating from nearby are louder than distant sounds seems obvious. However, the brain interprets loudness in a relative way. For familiar sounds such as a human voice, the brain has a frame of reference and VR sound designers must pay attention to both the type of sound being represented and how it fits against other sounds in the virtual ambient space (Begault, 2000). If a sound is moving from distant to near, clearly it must increase in volume to be convincing and decrease in intensity as it moves away. Depending on the processing power

of the VR system, it is also possible to recreate Doppler effect, wherein the frequency spectra of a sound get higher as it approaches and decrease as it becomes more distant (Begault, 2000).

Reverberation is the perceptual phenomena resulting from sound waves bouncing off solid objects and ending up at the ear some time later than sound waves which travelled directly from source to listener. An echo in a cathedral is a very obvious type of reverberation where the initial sound and its reflection off distant walls are widely separated in time. However, given the auditory system's high sensitivity to timing, humans are able to detect very fast reflections in small spaces or bouncing off objects in peripersonal space (Begault, 2000).

Given that one's experience of architectural spaces is influenced by the echoic quality of a large room or the hushed sensation of a small room, VR sound designers make use of reverberation to create the impression of interior spaces. The physical dimensions of a virtual room may be modeled mathematically to accurately simulate this effect, however, creating realistic reverberation is computationally costly and given finite processing resources, virtual acoustic environments often sacrifice detail in this domain (Begault, 2000).

Similarly difficult to reproduce given current computational capacities, are the changes in reverberation one experiences as one moves throughout a room and the transition one experiences moving from one reverberant space to another of different size ("Oculus/Developers," n.d.). However, by changing the proportions of a mixture of unreflected sound and echo, sound designers can achieve some of the acoustic experience of architectural spaces. More reverberation and less direct sound produces the experience of a large space (or one built of hard, sonically reflective surfaces) and less echo and more unreflected sound gives the sensation of a smaller (or soft surfaced) space (Begault, 2000).

Finally, in creating the sensation of distance to a sound source, VR sound designers can mimic the fact that high frequency sounds attenuate more rapidly in air than low frequency sounds ("Oculus/Developers," n.d.). In a natural environment, the low frequency components of a sound will travel longer distances and thus, we hear only the rumble of a distant passing truck and a conversation held across the room is a murmur. Though this phenomenon happens naturally only over long distances, removing high frequency components of a sound can shift them farther away in our perceptual space (Begault, 2000).

The limits of computational processing power place boundaries on the sound designer's ability to create fully realistic auditory environments in the virtual space. The more complex a

virtual acoustic environment, the more processing is required of the system. If the system is overburdened computationally, latencies or time-lag will result causing unpleasant disorientation for the VR participant. As with the visual components of virtual reality, a choice must be made between detail and the smoothness of the VR experience ("Oculus/Developers," n.d.).

Future Directions…In Space

Current neuroscientific applications of virtual reality include social neuroscience paradigms, investigations of visuo-tactile multisensory integration, neurotherapeutic implementations and the more visually based aspects of spatial navigation (Bohil, Alicea, & Biocca, 2011). But as the technology develops, it will become a powerful tool for investigating the spatial perceptual abilities of human hearing. Though seemingly contradictory, given the dominance of the visual component of VR, virtual environments also have a great deal to offer the visually impaired.

Despite a common conception that loss of vision corresponds with an improvement of hearing, Voss (2016) found evidence that blind people often suffer spatial hearing deficits. A recent study by Kolarik, Pardhan, Cirstea and Moore (2017) investigated auditory spatial perception in blind people using a virtual acoustic environment. Systematic distortions in blind people's distance estimation were identified relative to sighted controls. Such findings have great value, as identifying common deficits in the auditory spatial perception of blind people can allow for the creation of training paradigms to counteract these deficits.

Passamonti, Frissen and Ládavas (2009) developed a spatial hearing recalibration paradigm for the visually impaired (partially sighted). But for the profoundly blind, a VR environment that tracks head orientation and position could be used to provide precisely tuned spatial auditory stimuli to the user's ears. This could then be paired with hand, controller, gestural tracking or other VR user-input modality to create a feedback loop allowing blind people to finely calibrate their spatial hearing in a safe, and possibly entertaining way.

As the computational processing power of VR systems increases and research on auditory physiology and cognition deepens, we can expect to simultaneously understand more about how hearing works and experience more refined ways to fool it. In the realm of neuroscience research, these two phenomena will constructively interact, like two waveforms meeting in-phase.

References

Begault, D. R. (2000). *3-D sound for virtual reality and multimedia.* Moffett Field, CA: National Aeronautics and Space Administration.

Bellmund, J. L. S., Gärdenfors, P., Moser, E. I., Doeller, C. F. (2018). Navigating cognition: Spatial codes for human thinking. *Science, 362,* 654. doi: 10.1126/science.aat6766

Bohill, C. J., Alicea, B., & Biocca, F. A. (2011). Virtual reality in neuroscience research and therapy. *Nature Reviews Neuroscience, 12,* 752-762. doi:10.1038/nrn3122

Kolarik, A. J., Pardhan, S., Cirstea, S., & Moore, B. C. J. (2017). Auditory spatial representations of the world are compressed in blind humans. *Exp Brain Res, 235,* 597–606. doi: 10.1007/s00221-016-4823-1

Maguire, E. A., Gadian, D. G., Johnsrude, I. S., Good, C. D., Ashburner, J., Frackowiak, R. S. J., & Frith, C. D. (2000). Navigation-related structural change in the hippocampi of taxi drivers. *Proceedings of the National Academy of Science, USA, 97,* 4398-4403.

Oculus/Developers: 3D Audio Spatialization. Retrieved from https://developer.oculus.com/documentation/audiosdk/latest/concepts/audio-intro-spatialization/

Passamonti, C., Frissen, I., & Làdavas, E. (2009). Visual recalibration of auditory spatial perception: two separate neural circuits for perceptual learning. *European Journal of Neuroscience, 30*, 1141-1150. doi: 10.1111/j.1460-9568.2009.06910.x

Purves, D., Augustine, G. J., Fitzpatrick, D., Hall, W. C., LaManita, A. S., & White, L. (Eds.). (2012). *Neuroscience* (5th Ed.). Sunderland, MA: Sinauer Assoc., Inc.

Teki, S., Sukhbinder, K., von Kriegstein, K., Stewart, L., Lyness, C. R., Moore, B. C. J., Capleton, B., & Griffiths, T. D. (2012). Navigating the auditory scene: An expert role for the hippocampus, *The Journal of Neuroscience, 32,*(35), 12251-12257. doi:10.1523/JNEUROSCI.0082–12.2012

Voss, P. (2016). Auditory Spatial Perception without Vision. *Front. Psychol. 7*. doi: 10.3389/fpsyg.2016.01960

Ward, J. (2010). *The student's guide to cognitive neuroscience* (2nd ed.). New York, NY: Psychology Press.

Wenzel, E. M., Arruda, M., Kistler, D. J., & Wightman, F. L. (1993). Localization using non-individualized head-related transfer functions. *Journal of the Acoustical Society of America, 94,* 111-123.