Psychoacoustics of Spatial Hearing

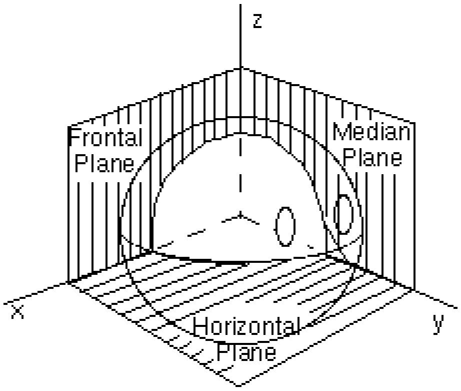
Psychoacoustics of Spatial Hearing

It has been said that “the purpose of the ears is to point the eyes.” While the ability of the auditory system to localize sound sources is just one component of our perceptual systems, it has high survival value, and living organisms have found many ways to extract directional information from sound. Although perceptual mysteries remain, the major cues have been known for a long time, and careful psychological studies have established how accurately we can make localization judgments. Anyone who wants to generate spatial sound for HCI has to know what influences the human auditory system. This section summarizes the major factors that influence spatial hearing. The following topics are covered:

- Coordinate systems
- Azimuth Cues
- Elevation Cues
- Range Cues
- Reverberation and Echoes

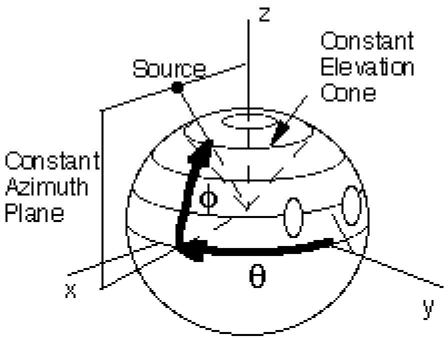
[For an introduction to the psychology of hearing in general, see Handel. The traditional reference for spatial hearing is Blauert. Mills gives an old but insightful overview, but Carlile will probably be the most valuable survey for HCI people.]

Coordinate Systems

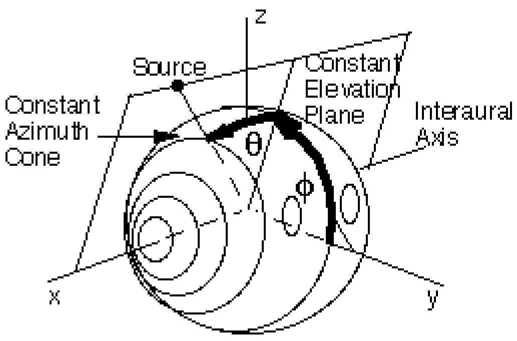


To specify the location of a sound source relative to the listener, we need a coordinate system. One natural choice is the head-centered rectangular-coordinate system shown above. Here the x-axis goes (approximately) through the right ear, the y axis points straight ahead, and the z axis is vertical. This defines three standard planes, the xy or horizontal plane, the xz or frontal plane, and the yz or median plane (also called the mid-sagittal plane). Clearly, the horizontal plane defines up/down separation, the frontal plane defines front/back separation, and the median plane defines right/left separation.

However, because the head is roughly spherical, a spherical coordinate system is usually used. Here the standard coordinates are azimuth, elevation and range. Unfortunately, there is more than one way to define these coordinates, and different people define them in different ways. The vertical-polar coordinate system (shown below on the left) is the most popular. Here one first measures the azimuth θas the angle from the median plane to a vertical plane containing the source and the z azis, and then measures the elevationas the angle up from the horizontal plane. With this choice, surfaces of constant azimuth are planes through the z axis, and surfaces of constant elevation are cones concentric about the z axis.



Vertical-Polar Coordinates



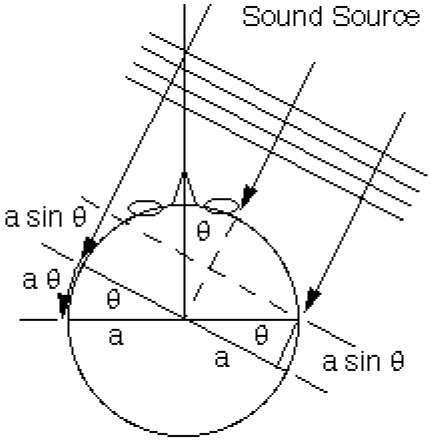
Interaural-Polar Coordinates

An important alternative is the interaural-polar coordinate system, shown above on the right. Here one first measures the elevation as the angle from the horizontal plane to a plane through the source and the x axis, which is the interaural axis; the azimuth is then measured as the angle over from the median plane. With this choice, surfaces of constant elevation are planes through the interaural axis, and surfaces of constant azimuth are cones concentric with the interaural axis.

The vertical-polar system is definitely more convenient for describing sources that are confined to the horizontal plane, since one merely has to specify the azimuth as an angle between -180Â° and +180Â°. With the interaural-polar system, the azimuth is always between -90Â° and +90Â°; surprisingly, the front/back distinction must be specified by the elevation, which is 0Â° for sources in the front horizontal plane, and 180Â° (or -180Â°) for sources in the back. While that is certainly clumsy, we shall see that the interaural-polar system makes it significantly simpler to express interaural differences at all elevations.

Azimuth Cues

One of the pioneers in spatial hearing research was John Strutt, who is better known as Lord Rayleigh. About 100 years ago, he developed his so-called Duplex Theory. According to this theory, there are two primary cues for azimuth — Interaural Time Difference (ITD) and Interaural Level Difference (ILD).



Lord Rayleigh had a simple explanation for the ITD. Sound travels at a speed c of about 343 m/s. Consider a sound wave from a distant source that strikes a spherical head of radius a from a direction specified by the azimuth angle θ. Clearly, the sound arrives at the right ear before the left, since it has to travel the extra distance $a\,\theta + a \sin\theta$ to reach the left ear. Dividing that by the speed of sound, we obtain the following simple (and surprisingly accurate) formula for the interaural time difference:

$$\mathrm{ITD} = \frac{a}{c}(\theta + \sin\theta)\ ,\ -90^\circ \le \theta \le +90^\circ$$

Thus, the ITD is zero when the source is directly ahead, and is a maximum of $(a/c)(\pi/2+1)$ when the source is off to one side. This represents a difference of arrival time of about 0.7 ms for a typical size human head, and is easily perceived.*

Lord Rayleigh also observed that the incident sound waves are diffracted by the head. He actually solved the wave equation to show how a plane wave is diffracted by a rigid sphere. His solution showed that in addition to the time difference there was also a significant difference between the signal levels at the two ears — the ILD.

As you might expect, the ILD is highly frequency dependent. At low frequencies, where the wavelength of the sound is long relative to the head diameter, there is hardly any difference in sound pressure at the two ears. However, at high frequencies, where the wavelength is short, there may well be a 20-dB or greater difference. This is called the head-shadow effect, where the far ear is in the sound shadow of the head.

The Duplex Theory asserts that the ILD and the ITD are complementary. At low frequencies (below about 1.5 kHz), there is little ILD information, but the ITD shifts the waveform a fraction of a cycle, which is easily detected. At high frequencies (above about 1.5 kHz), there is ambiguity in the ITD, since there are several cycles of shift, but the ILD resolves this directional ambiguity. Rayleigh's Duplex Theory says that the ILD and ITD taken together provide localization information throughout the audible frequency range.

Footnotes

Clearly, one can invert this equation and obtain the azimuth from the ITD. The auditory system must perform a more or less equivalent function in recovering the azimuth from ITD information.

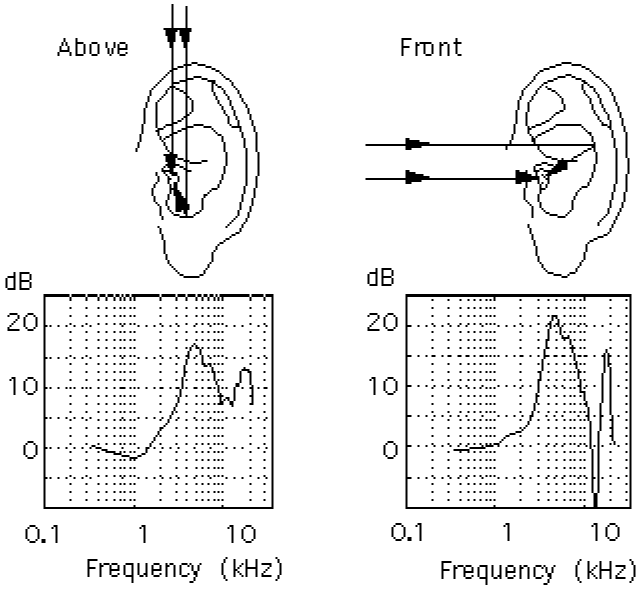
The accuracy with which this can be done depends on the circumstances. For speech in normally reverberant rooms, typical human accuracies are on the order of 10Â° to 20Â°. However, under optimum conditions, much greater accuracy (on the order of 1Â°) is possible if the problem is to decide merely whether or not a sound source moves. This is rather remarkable, since it means that a change in arrival time of as little as 10 microseconds is perceptible. (For comparison, the sampling rate for audio CD's is 44.1 kHz, which corresponds to a sampling interval of 22.7 microseconds. Thus, in some circumstances, less than a one-sample delay is perceptible.)

It is also worth noting that if we use an interaural-polar coordinate system and hold the azimuth constant, then we obtain a constant value for the ITD. Thus, there is a simple one-to-one correspondence between the ITD and the cone of constant azimuth, which is sometimes called the "cone of confusion". This is not the case for the vertical-polar system.

Finally, note that the ITD alone only constrains the source to be somewhere on the cone of confusion. In particular, it is not sufficient for determining whether the source is in front or in back.

Elevation Cues

While the primary cues for azimuth are binaural, the primary cues for elevation are often said to be monaural. They stem from the fact that our outer ear or pinna acts like an acoustic antenna. Its resonant cavities amplify some frequencies, and its geometry leads to interference effects that attenuate other frequencies. Moreover, its frequency response is directionally dependent.



The figure above shows measured frequency responses for two different directions of arrival. In each case we see that there are two paths from the source to the ear canal — a direct path and a longer path following a reflection from the pinna. At moderately low frequencies, the pinna essentially collects additional sound energy, and the signals from the two paths arrive in phase. However, at high frequencies, the delayed signal is out of phase with the direct signal, and destructive interference occurs. The greatest interference occurs when the difference in path length d is a half wavelength, i.e., when f = c / 2d. In the example shown, this produces a “pinna notch” around 10 kHz. With typical values for d, the notch frequency is usually in the 6-kHz to 16-kHz range.

Since the pinna is a more effective reflector for sounds coming from the front than for sounds from above, the resulting notch is much more pronounced for sources in front than for sources above. In addition, the path length difference changes with elevation angle, so the frequency of the notch moves with elevation. Although there are still disputes about what features are perceptually most important (for example, see Han), it is well established that the pinna provides the primary cues for elevation.

Range Cues

When it comes to localizing a source, we are best at estimating azimuth, next best at estimating elevation, and worst at estimating range. In a similar fashion, the cues for azimuth are quite well understood, the cues for elevation are less well understood, and the cues for range are least well understood. The following cues for range are frequently mentioned:

- Loudness
- Motion parallax
- Excess interaural level difference (ILD)
- Ratio of direct to reverberant sound

The physical basis for the loudness cue obviously stems from the fact that the captured sound energy coming directly from the source falls off inversely with the square of range. Thus, as a constant-energy source approaches a listener, the loudness will increase. It is equally obvious that the received energy is proportional to the energy emitted by the source, and that there cannot be a one-to-one relation between loudness and range. Just playing a sound at a low volume level will not, in itself, make it seem to be far away. To use loudness as a cue to range, we must also know something about the characteristics of the source. In the case of human speech, each of us knows from experience the different quality of sound associated with whispering, normal talking, and shouting, no matter what the sound level. The combination of loudness and knowledge of the source provides useful information for range judgments.

Motion parallax refers to the fact that if a listener translates his or her head, the change in azimuth will be range dependent. For sources that are very close, a small shift causes a large change in azimuth, while for sources that are distant there is esentially no azimuth change.

In addition, as a sound source gets very close to the head, the ILD will increase. This increase becomes noticeable for ranges under about one meter. An extreme case is when there is an insect buzzing in one ear, or when someone is whispering in one ear. In general, sounds that are heard in only one ear are threatening and are uncomfortable to listen to. It is particularly important to keep this in mind when designing HCI systems for headphone listening. As we will see, to get the listener to think that the sound is on one side, it is not at all necessary to have all of the sound in that ear and nothing in the other ear.

The final cue listed is the ratio of direct to reverberant sound. As we mentioned above, the energy received directly from a sound source drops of inversely with the square of the range. However, in ordinary rooms, the sound is reflected and scattered many times from environmental surfaces, and the reverberant energy reaching the ears does not change much with the distance from the source to the listener. Thus, the ratio of direct to reverberant energy is a major cue for range. At close ranges, the ratio is very large, while at long ranges it is quite small. Fortunately, this is a relatively easy and effective cue to manipulate for HCI applications.

Reverberation and Echoes

Most of the time we are unaware of how much of the sound that we hear comes from reflections from environmental surfaces. Even out of doors, a significant amount of energy is reflected by the ground and by surrounding structures and vegetation. However, we only notice these reflections when the time delay gets longer than about the 30- to 50-ms echo threshold, in which case we become consciously aware of them and call them echoes. Special rooms called anechoic chambers are built to absorb sound energy, so that only the directly radiated energy reaches the ears. Upon entering an anechoic chamber for the first time, most people are astonished by how much softer and duller everything sounds.

If reflected sound is so common in ordinary acoustic environments, one might wonder why these reflections do not interfere with our ability to localize sources. The answer seems to be that we quickly adapt to a new environment, and our auditory system uses only partially understood mechanisms to suppress the effects of reflections and reverberation. The fact that we localize on the basis of the signals that reach our ears first is known as the precedence effect or the Law of the First Wavefront (see Blauert). This is not to say that we are unaware of the reflections that follow. Indeed, we subconsciously use this information to estimate range. However, unless reverberation is severe, the reflections have relatively little effect on our ability to localize sounds.

However, the precedence effect does force us to modify Rayleigh's Duplex Theory. In a typical room, reflections begin to arrive a few milliseconds after the initial sound. For a low-frequency sound whose period is longer than the time for reflections to arrive (for example, below 250 Hz), the reflections begin arriving before even one cycle is completed. By the time several cycles have arrived and the auditory system can begin to estimate pitch, the sound pattern in the room is a jumble of standing waves, and it is now impossible for the auditory system to estimate interaural time differences. Thus, in a reverberant room, low-frequency information is essentially useless for localization.

However, that does not mean that interaural timing differences are unimportant. The important timing information comes from the Interaural Envelope Difference (IED), e.g., from the transients at the onset of a new sound. This is vividly demonstrated by the Franssen Effect. If a sine wave is suddenly turned on and a high-pass-filtered version is sent to Loudspeaker A while a low-pass filtered version is sent to Loudspeaker B, most listeners will localize the sound at Loudspeaker A. This is true even if the frequency of the sine wave is sufficiently low that in steady state most of the energy is coming from Loudspeaker B. Basically, the starting transient provides unambiguous localization information, while the steady-state signal is very difficult to localize, and in this circumstance the auditory system simply ignores the ambiguous information. With some risk of oversimplification, we can generalize and say that in reverberant environments it is the high-frequency energy, not the low-frequency energy, that is important for localization.