Localization of Sound to Image

A conceptual approach to a closer-to-reality moviegoing experience

by Mark Mayfield,
Director of Cinema Products
Eastern Acoustic Works, Inc.

There is no doubt that presentation quality in the movie theatre has taken dramatic steps in the past two decades. Beginning with Dolby processing in the late 1970s, film exhibition was finally offered a cost-effective way to play back films with a noise-free, dynamically large soundtrack. Improvements in picture quality have also been made possible with brighter light sources, better lenses, high quality filmstock and screens. The introduction of flat power response loudspeaker systems in the early eighties was a major improvement in cinema loudspeaker systems, and the Lucasfilm/THX program helped to promote the concept of paying attention to all aspects of motion picture exhibition. And the most recent milestone, digital sound for film playback, has taken sound to yet another level of quality.

So is that as far as we need to go? Do we now have the technology to say it's "as good as it gets"? The answer, for any creatively-minded industry, is, of course, no. What digital sound technology has really brought to the motion picture industry is more degrees of freedom to experiment and enhance the moviegoing experience. And this is what will keep our industry a strong and viable one, even in the face of continuously improving home theatre technology and entertainment alternatives.

Now that a digital source is available for nearly perfect sound storage and playback, we can look beyond the soundtrack itself and focus on different ways to enhance to moviegoing experience -- in ways that can't easily be duplicated in the home.

First, let's make a distinction: there is the *soundtrack*, where the sound is actually stored in synch with the image. The soundtrack, and all of the electronics used to read it from the film, is typically part of what is known as the "A-chain". Then there is *playback system*, which includes loudspeakers, amplifiers, and the room itself. This is also known as the "B-chain". With high quality, modern equipment throughout the A and B-chains, today's digital soundtracks offer any well-designed theatre the potential for truly outstanding sound quality in the auditorium. But this article is not about sound quality; this time we are talking about sound *location*. In the early days of sound on film, a single channel of soundtrack was all that was available. A single loudspeaker system, placed behind a perforated screen, was used to link the sound with the picture. As multiple channel soundtrack formats began to emerge in the 1950s, more speakers were placed behind the screen in the house, to take advantage of discrete sound elements on the soundtrack. The conventional approach became a three-channel, three-speaker arrangement corresponding to Left, Center, and Right stage/screen systems, and an array of surround loudspeaker in the house. The center speaker "anchors" the dialog, left and right speakers provide a canvas for music, effects, sometimes panned dialog, and occasional "motion pans" -- cars zooming left to right across the screen, for example. The rule of thumb has been to elevate these three speakers from 1/2 to 2/3 the screen's height.

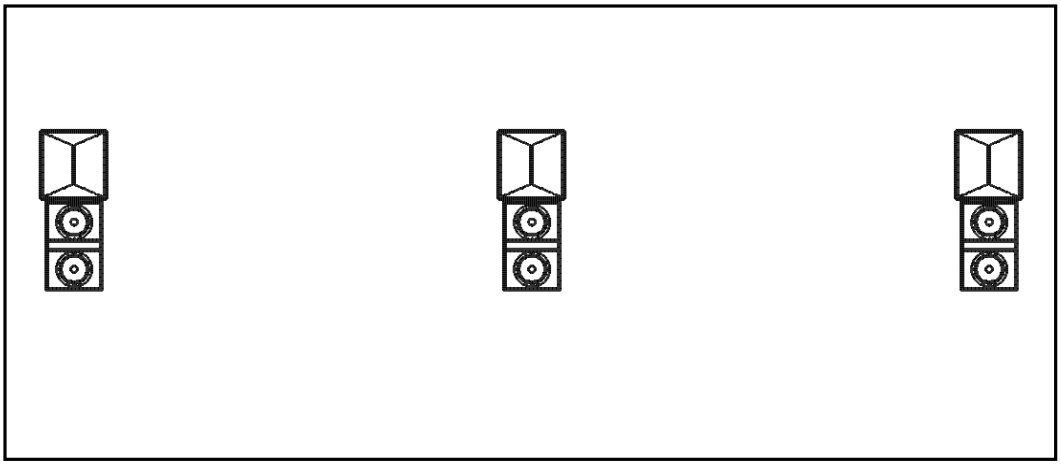

Figure 1. Three speakers behind screen, corresponding to Left, Center, and Right.

So what we have lived with for nearly fifty years (and the standard of today) is a refined loudspeaker system that can realistically reproduce sound along the horizontal axis. But the visual action uses the whole screen -- left to right, top to bottom. So why do we only move sound along the horizontal axis? One reason is that the human hearing mechanism is more sensitive to sound along the horizontal axis than the vertical axis. This is because we have two ears located on the sides of our heads. A sound originating from our left side will reach our left ear first, and our right ear slightly later. This "interaural time difference"

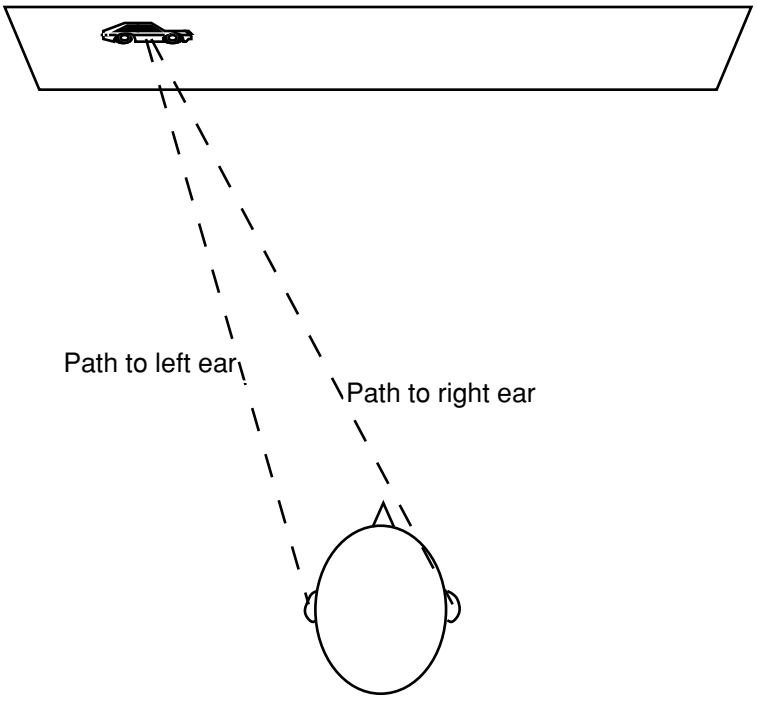

Figure 2. The interaural time difference is created by different path lengths from the sound source to each ear. (Perspective is from overhead, looking down at a seated listener in a theatre).

(ITD) is what allows our brain to localize sound. Generally, the sound will be perceived to be coming from the direction of the first-arriving sound, and its exact lateral position is

determined by the ITD information from both ears. But since we don't have ears at the top and bottom of ours heads, no such ITD is available as a location cue in the vertical plane. However, according to Dr. Richard Duda, a leading researcher in the field from San Jose State University, although people are more sensitive to horizontal changes, they can definitely detect changes in elevation of a sound source.

So if we know that we can "localize" (find a sound's location) so well in the horizontal plane, why doesn't the sound move from side to side with the picture? In cinemagraphic terms, the close-up presents no problem, since the talking head is usually centered on the screen exactly where the center channel speaker is located. The "two-shot" begins to dislocate sound to image, since two talkers are in the scene simultaneously while both actors' dialog typically is mixed to the Center channel. There have been experiments with mixing dialog Left and Right in exact accordance with the actors on screen. For example, while mixing the epic "Spartacus" in 1960, director Stanley Kubrick insisted on panning (moving the sound position) almost everything. According to Don Rogers, sound director at Warner Hollywood studios (from the book "Sound-on-Film" by Vincent LoBrutto), "They had a pan pot to place the sound of every actor, every sound effects, every footstep, every bang and every crash exactly where is was on the screen." The mix, in this case, took nine months to complete. Beside being very costly, it has been commented that constant movement of sound images becomes distracting and fairly tedious to listen to. Besides, as ventriloquists have known for centuries, it is very easy to "fool" the brain into assuming a sound source's location, by simply providing a plausible visual cue with which to pair the sound. It is well-known in perception research that while the auditory system is highly refined and accurate, it always takes cues provided by the visual system. When we see a bird or an airplane, for example, we will assume the sound source to be above the horizontal plane. Or when we see footsteps, our we assume the sound to be coming from the bottom of the screen. For these reasons, it has generally been accepted that the ability to move sound left and right along the horizontal plane provided a close enough match to image movement on a standard movie screen. Whatever mis-matches there were between location of the image and actual location of the sound would be compensated for by the brain.

So why bother to more closely-link sound with image? Could it really be worth all this fuss? For one thing, can we really detect elevation changes in a sound source's location. According to research, the sensitivity to change is greatest near the horizontal plane when the sound is moved more than three degrees. When the sound is nearly overhead, we are less sensitive to movement unless the sound moves greater than twenty degrees. In other words, if you were seated, looking straight ahead, you could only detect a sound changing elevation if it moved in an arc greater than three degrees.

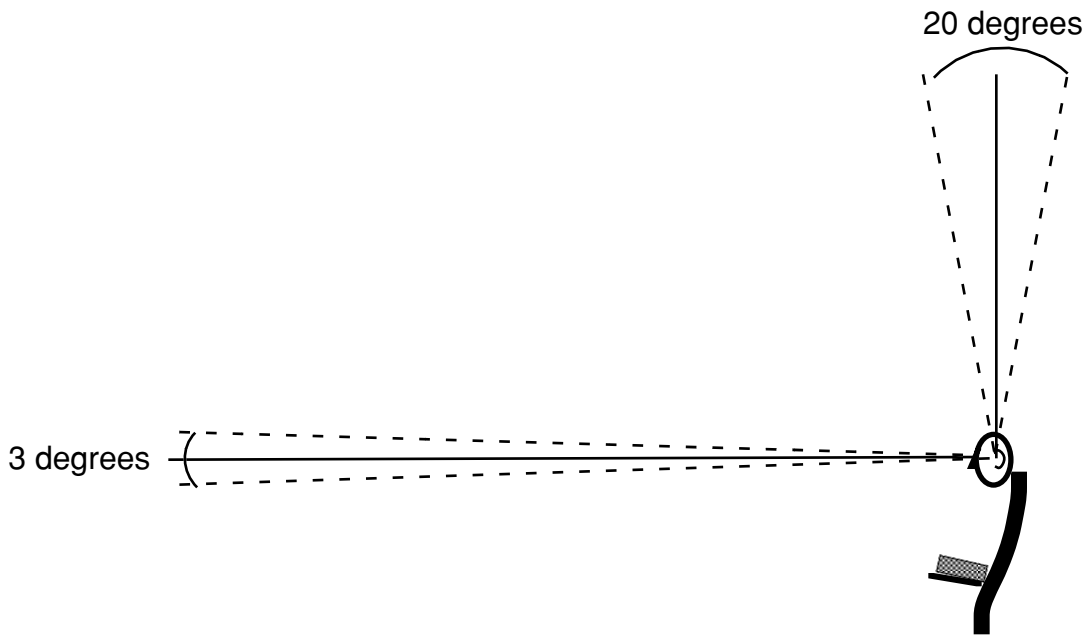

Figure 3. Sensitivity to sound elevation in the vertical place, assuming the head is stationery.

At eight feet from a sound source, three degrees translates to about five inches in elevation difference. In a movie theatre, however, if the nearest seats were 16 feet from the screen, listeners should able to detect elevation differences of ten inches. At a distance of 68 feet (the screen to last row distance of an average multiplex theatre), three degrees is the equivalent of three-and-a-half feet on the screen. When you consider that action takes place with a much broader vertical spread than 3.5 feet, you begin to wonder whether a more realistic sense of "action" would be possible of we didn't have to "fool" our brain with psychoacoustics. Instead, a more involving sense of the action should be possible.

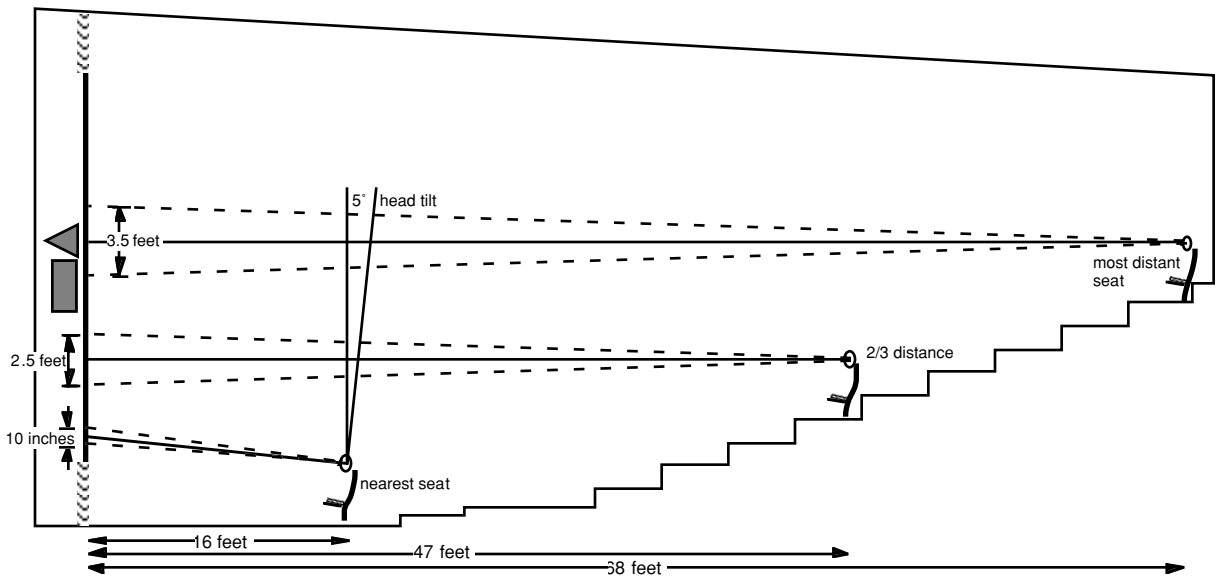

Figure 4. Three-degrees as it translates to screen image in a medium sized theatre.

Trends in current motion picture genres and exhibition suggest that achieving a better image to sound match might be desirable. First of all, by far the biggest boxoffice successes of the day are the large, special effects-laden action films. The success of these films have much to do with their ability to create a sense of realism, even though the plot or scene may be totally implausible. Also, screens are getting larger and larger, and the standard Left, Center, and Right speaker locations (with conventional loudspeakers) may be too widely spaced to create a seamless horizontal sound image. The "7.1" soundtrack format available from sound formats such as Sony's SDDS addresses this to some degree by placing five channels of speakers behind the screen. Panning audio on very wide screen image across five channels has distinct advantages over only three channels, since fewer "phantom sources" are created. Phantom sources have the disadvantage of not always providing accurate audio localization for every seat in a large theatre. Recognizing the need to expand the sound field on very large screens, many of the so-called "large format" film systems (such as IMAX) use multiple center channel speakers in an over-under arrangement, in order to create a better sense of vertical audio imaging.

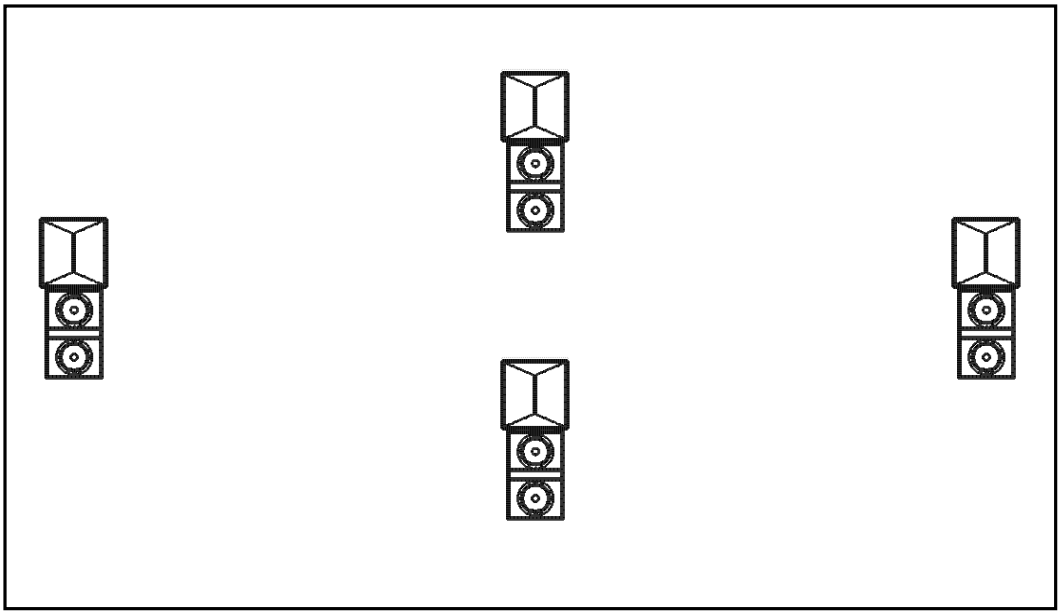

Figure 5. Behind screen speaker arrangement in use with some large format sound systems. Center channel is fed to two speakers to create sense of vertical audio imaging.

A large action scene projected onto a large, wall-to-wall, ceiling-to-floor screen will undoubtedly result in important visual elements appearing in the same frame at significant distances from each other. If the sound elements do not closely match the visual elements location, then some sense of "dis" location is likely to be noticed. Even though psychoacoustics and "the ventriloquist effect" will compensate for much of this disparity, it seems logical that a heightened sense of realism would be experienced with a more accurate match of sound to image.

So just how could this be accomplished? New processes and practices both in soundtrack production and sound playback would be necessary to make it work. Of course this means virtually abandoning many of the "accepted" practices of film sound as we know them today. First, would need some way to record the actual location of a certain sound on the sound track relevant to and in synchronization with the visual image. We would also need a corresponding process for "moving" the image in playback.

Recording the sound location relative to the image

Fortunately, we have already mastered the process of storing complex and prodigious amounts of information on a piece of film, in synch with picture. The development of digital sound for film has forced our industry to face the issues of data compression, storage, and playback. So adding another parameter of information, such as the X and Y coordinates of a sound's physical location, would seem to be a incremental undertaking in software programming. Perhaps there would be another function in the sound post-production process called "sound location". Imagine a stylus device, much like a pen, and an image pad, which, when touched would position a location mark ("dot") superimposed over the scene. We have already seen this process when television sports commentators draw over a football play to diagram what has happened on the field. As the stylus pinpoints the action, the coordinates are recorded in the data bitstream along with the audio information.

Playing back the localized sound

The most obvious way to approach this would be to physically move a loudspeaker behind the screen so that it tracks to movement of the image on the screen. But almost any system of physical movement would suffer from mechanical complexity and perhaps noise of moving cables and Too many moving parts -- reliability issues.

Another way to do this involves a total shift in thinking about cinema sound playback as we know it today (and for the past 70 years). Let's begin with the concept of the baffle wall. Since the beginning of cinema sound, baffle wings, or extension on the side of the loudspeaker enclosure have been recommended to enhance low frequency output. Today, with the problem of low cost high output low frequency speakers and amplifiers, we use baffle walls more for their ability to minimize rear of screen reflections and thus enhance dialog intelligibility.

For this theoretical exercise, assume that we keep the baffle wall, but instead of putting only three speakers, corresponding to Left, Center and Right, we use many more. The speakers can be much smaller in size, since we know their main purpose is to localize the sound, not necessarily provide many locations of full range sound. The low frequencies (below 150 Hz) can be handled by standard floor mounted low frequency cabinets. Since we have determined that the closest listener could detect vertical image changes less than one foot, our minimal spacing could be one foot from center to center of each loudspeaker. This would suggest over a thousand speakers for a 45 by 25 foot screen! A more practical approach might be to maximize the effect for the "2/3" distance seating area. In the average medium-sized multiplex style theatre (based on a statistical analysis of over 200 theatres worldwide), this "2/3" distance is about 47 feet from screen. At this position, the minimum vertical elevation distance which humans could detect would be 2.5 feet. Assume the spacing between any two speaker centers would be 2.5 feet, and for a 45 by 25 foot screen, that calculates to about 180 speakers (ten rows of eighteen speakers).

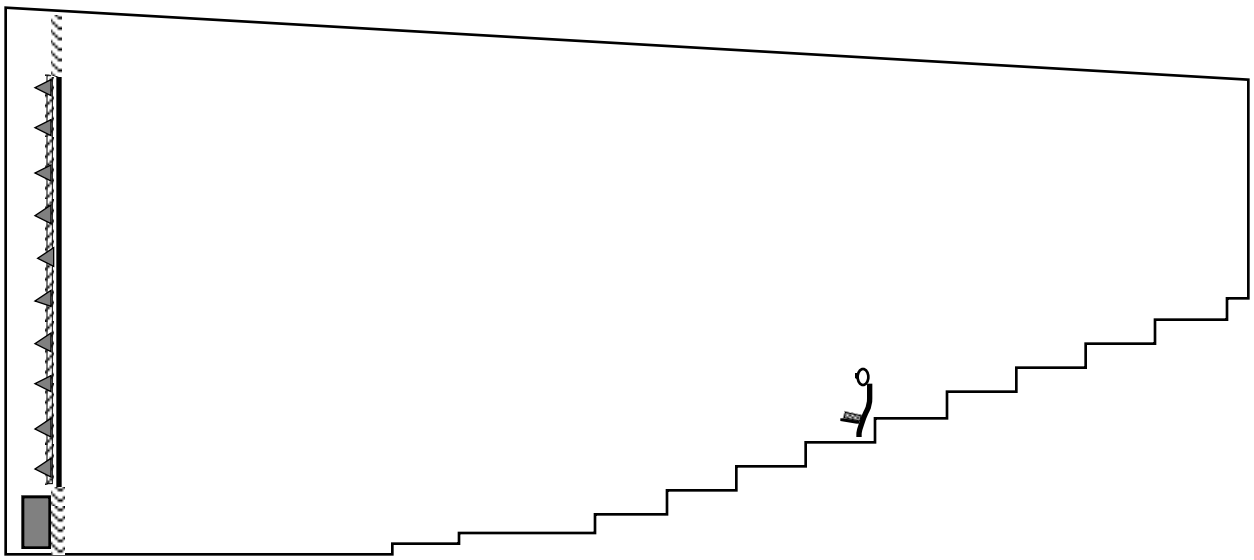

Figure 6. Multiple speaker array built into baffle for Cartesian-coordinate cinema playback system, side view.

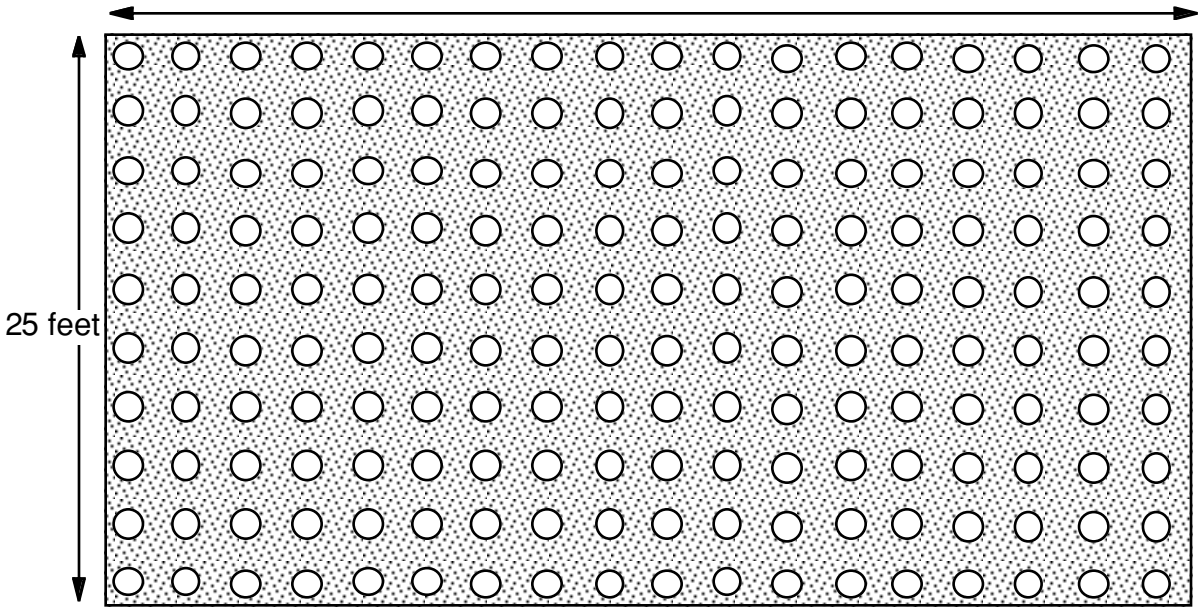

Figure 7. Multiple speaker array for Cartesian-coordinate cinema playback system, front view.

Now the coordinates that were recorded during the sound location phase of the post production can be directed the nearest X/Y loudspeaker coordinate position -- and you have nearly perfect sound to image localization.

Obviously, there is a large chasm between this theoretical solution and its practical implementation, but the tools are here today. There would also be an "unlearning" period before we could convince our brain that dialog doesn't always have to come from only the screen center position. Since sound and image are generally matched with precision in real life, creating a similar effect in the cinema should dramatically enhance a sense of reality on the screen.

Information Sources:

Blauert, J. (1997). *Spatial Hearing* (Revised Edition) (MIT Press, Cambridge, MA).

Center for Image Processing and Interactive Computing, University of California at Davis.

Duda, Dr. Richard. (correspondences) San Jose State University.

Holman, Tomlinson (1997). *Sound for Film and Television*. (Focal Press, Newton, Massachusetts).

LoBrutto, Vincent (1994). *Sound-on-Film: Interviews with Creators of Film Sound*. Praeger Publishers, Westport, Connecticut.

Middlebrooks, J.C., and Green, D.M. (1991). "Sound localization by human listeners", Annual Review of Psychology, Vol. 42, pp.135-139.