IMMERSIVE AUDIO FOR THE DESKTOP

C. Kyriakakis and T. Holman

Integrated Media Systems Center
University of Southern California
3740 McClintock Ave., EEB 432
Los Angeles, California 90089-2564, USA

ABSTRACT

Integrated media workstations are increasingly being used for creating, editing, and monitoring sound that is associated with video or computer-generated images. While the requirements for high quality reproduction in large-scale systems are well understood, these have not yet been adequately translated to the workstation environment. In this paper we discuss several factors that pertain to high quality sound reproduction at the desktop including acoustical considerations, signal processing requirements, and listener location issues. We also present a novel desktop system design with integrated listener-tracking capability that circumvents several of the problems faced by current digital audio and video workstations.

1. INTRODUCTION

Several applications are envisioned for integrated media workstations. The principal function of such systems is to manipulate, edit, and display still images, video, and computer animation and graphics. The necessity to edit and accurately monitor the sound associated with such visual images in the desktop environment has only recently been recognized. Accurate reproduction of audio program material combined with accurate spatial perception of sound are necessary to create a seamless aural environment and to achieve sound localization relative to visual images. In fact, a mismatch between the aurally-perceived and visually-observed positions of a particular sound causes a cognitive dissonance that can seriously limit the desired suspension of disbelief.

In this paper we examine several key issues in the implementation of high quality desktop-based audio systems. Such issues include the optimization of the frequency response over a given frequency range, the dynamic range, and stereo imaging subject to constraints imposed by room acoustics and human listening characteristics. Several problems that are particular to the desktop environment are discussed including the frequency response anomalies that arise due to the local acoustical environment, the proximity of the listener to the loudspeakers, the acoustics associated with small rooms, and the need to accurately track the listener's position relative to the loudspeakers.

2. REPRODUCTION REQUIREMENTS

2.1 Acoustical Considerations

In a typical desktop sound monitoring environment delivery of stereophonic sound is achieved through two loudspeakers that are typically placed on either side of a video or computer monitor. This environment, combined with the acoustical problems of small rooms, causes severe acoustical problems that contribute to audible distortion of the reproduced sound.

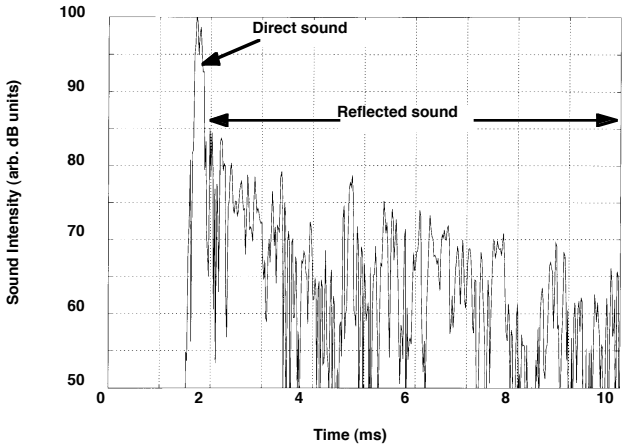

Figure 1. Desktop sound system time response that meets the requirements for low-level early reflections. All the peaks of the reflected sound are at least 15 dB below the direct sound.

Among these problems the one most often neglected is the effect of discrete early reflections. The effects of such reflections on sound quality has been studied extensively [1-3] and it has been shown that they are the dominant source of monitoring non-uniformities when all the other standards discussed above have been met. These non-uniformities appear in the form of colorations (frequency response anomalies) in rooms with an early reflection level that exceeds –15 dB spectrum level relative to the direct sound for the first 15 ms (Fig. 1). Such a high level of reflected sound gives rise to comb filtering in the frequency domain that in turn causes severe changes in timbre. The perceived effects of such distortions were not quantified until psychoacoustic experiments [1, 4] demonstrated their importance.

A potential solution that alleviates the problems of early reflections in small rooms is near-field monitoring. In theory, the direct sound is dominant when the listener is very close to the loudspeakers thus reducing the room effects to below audibility. In practice, however, there are several issues that must be addressed in order to provide high quality sound. One such issue relates to the large reflecting surfaces that are typically present near the loudspeakers. Strong reflections from a console or a video/computer monitor act as baffle extensions for the loudspeaker resulting in a boost of mid-bass frequencies. Furthermore, even if it were possible to place the loudspeakers far away from large reflecting surfaces, this would only solve the problem for middle and high

frequencies. Low frequency room modes do not depend on surfaces in the local acoustical environment, but rather on the physical size of the room. These modes produce standing waves that give rise to large variations in frequency response. Finally, the physical size of the loudspeakers has a negative effect on the quality of reproduced sound relates to. Typical two-way designs in which the tweeter is physically separated from the woofer exhibit strong radiation pattern changes in the crossover frequency range. Amplitude and phase matching in this frequency range becomes critical and as a result such speakers are extremely sensitive to placement and typically produce a flat frequency response for direct sound *in one exact position*. This limitation makes typical two-way speakers unsuitable for near-field monitoring.

The current state-of-the-art in desktop reproduction systems is rather poor. Large deviations from flat response arise from a combination of loudspeaker design and acoustical environment drawbacks (Fig. 2). The sound reproduced by such systems does not meet the standards required for professional applications and does a very poor job at producing a high fidelity experience for the desktop.

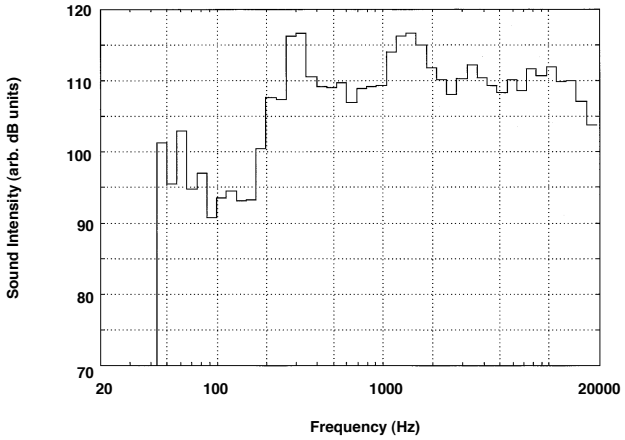

Figure 3. Frequency response of a typical desktop loudspeaker system. The two large peaks centered at 300 Hz and 1500 Hz represent significant distortion in the reproduced sound. Also note the effect of standing waves in the 40 Hz to 100 Hz region.

3. DESIGNREQUIREMENTS

3.1 Frequency Response

In order to address the problems described above, a set of solutions has been developed for single listener desktop reproduction that delivers sound quality equivalent a calibrated dubbing stage. These solutions include:

Direct-path dominant design. By combining elements of psychoacoustics in the system design, it is possible to place the listener in a direct sound field that is dominant over the reflected and reverberant sound. The colorations that arise due to such effects are eliminated and this results in a listening experience that is dramatically different than what is achievable through traditional near-field monitoring methods. The design considerations for this direct-path dominant design include the effect of the video/computer monitor that extends the loudspeaker baffle, as well as the large reflecting surface on which the computer keyboard typically rests.

Correct low-frequency response. There are severe problems in the uniformity of low-frequency response that arise from the standing waves associated with the acoustics of small rooms. Such anomalies can give rise to variations as large as ±15 dB for different listening locations in a typical room. The advantage of desktop audio systems lies in the fact that the position of the loudspeakers and, to a large extent, the listener are known *a priori*. It is, therefore, possible to use equalization to produce very smooth low-frequency response. One fundamental limitation imposed by small room acoustics is that this can only be achieved for a relatively-small volume of space centered around the listener. One possible solution to this problem can be found by tracking the listener's position and adjusting the equalization dynamically.

3.2 3-D Audio Rendering

A critical issue in the implementation of immersive audio is the reproduction of 3-D spatially-correct sound fields. There are two methods for 3-D audio rendering. The first is based on headphones that are capable of reproducing binaural signals to each ear with no crosstalk. The second method of sound delivery is based on loudspeaker reproduction. Current multichannel (5.1 channel) systems can convey precisely-localized sound images that are primarily confined to the horizontal plane and diffuse (ambient) sound to the sides and behind the listener.

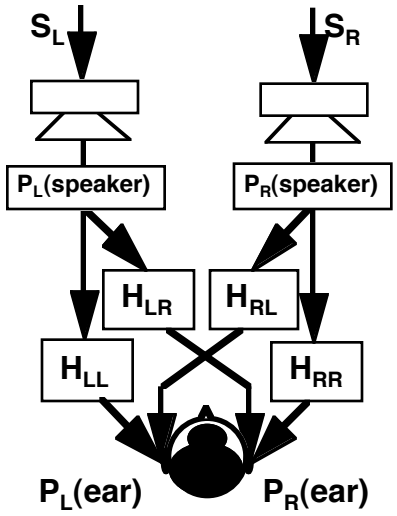

Figure 3. Crosstalk cancellation is required for loudspeaker reproduction of binaural signals. The signal to each loudspeaker is prefiltered so that the cross terms are canceled at the listener's ears.

While such systems are very successful at reproducing sound associated with film, they are not capable of reproducing fully three-dimensional soundfields. Binaural methods are more suitable for capturing and rendering such soundfields. These methods seek to accurately reproduce at each eardrum the sound pressure generated by a set of sources and their

interactions with the acoustic environment [5]. Recordings are made with a dummy-head microphone system that is based on average human characteristics. Sound recorded using binaural methods is then reproduced through headphones that attempt to deliver the desired sound to each ear.

There are, however, serious limitations associated with headphone listening that can be summarized as follows: (1) individualized head-related transfer functions (HRTF) are difficult to measure for each listener and the averaged HRTFs that are used make it impossible to match each individual's perception of sound; (2) there are large errors in sound position perception associated with headphones, especially for the most important visual direction, out in front; (3) headphones are uncomfortable for extended periods of time; and (4) it is very difficult to externalize sounds and avoid the "inside-the-head" sensation.

The use of loudspeakers for reproduction can circumvent the limitations associated with headphone reproduction of binaural recordings. In order, however, to deliver the appropriate binaural sound field to each ear it is necessary to eliminate the crosstalk that is inherent in all loudspeaker-based systems. This is a technological limitation of *all* loudspeaker systems and it arises from the fact that while each ear receives the desired sound from the same-side loudspeaker (ipsilateral), it also receives undesired sound from the opposite-side loudspeaker (contralateral).

Several schemes have been proposed to address crosstalk cancellation. The basic principle of such schemes relies on preconditioning the signal into each loudspeaker such that the output sound generates the desired binaural sound pressure at each ear. If we denote the sound pressures that must be delivered to each ear as P_L(ear) and P_R(ear) and the transfer functions from each loudspeaker to each ear as H_{LL}, H_{LR}, H_{RL}, and H_{RR} then we can write (Fig. 3):

$$\begin{aligned} P_L(\text{speaker}) &= H_{LL}S_L + H_{RL}S_R \\ P_R(\text{speaker}) &= H_{LR}S_L + H_{RR}S_R \end{aligned} \tag{1}$$

in which we denote by S_L and S_R the input signals to each loudspeaker and $P_{L,R}$(speaker) the sound pressure delivered by each loudspeaker. In order to accurately reproduce the desired binaural signal at each ear the input signals S_L and S_R must be chosen such that:

$$\begin{aligned} P_L(\text{ear}) &= P_L(\text{speaker}) \\ P_R(\text{ear}) &= P_R(\text{speaker}) \end{aligned} \tag{2}$$

The desired loudspeaker input signals are then found from:

$$\begin{aligned} S_L &= \frac{H_{RR}P_L(\text{ear}) - H_{RL}P_R(\text{ear})}{H_{LL}H_{RR} - H_{LR}H_{RL}} \\ S_R &= \frac{H_{LL}P_R(\text{ear}) - H_{LR}P_L(\text{ear})}{H_{LL}H_{RR} - H_{LR}H_{RL}} \end{aligned} \tag{3}$$

The only requirement is that S_L and S_R must be realizable filter responses (causal) [6]. The first implementation of a crosstalk cancellation scheme based on the theory described above was shown by Atal and Schroeder [6]. Later work by Cooper and Bauck [7, 8] showed that under the assumption of left-right symmetry a much simpler shuffler filter can be used to implement crosstalk cancellation as well as synthesize virtual loudspeakers in arbitrary positions. Cooper and Bauck went on to use results from Mehrgard and Mellert [9], who showed that the head-related transfer function is minimum phase to within a frequency-independent delay that is a function of the angle of incidence.

4. LISTENERTRACKING

4.1 Desktop Sound Rendering

\For desktop applications, in which a single user is located in front of a CRT display, the use of a center loudspeaker is not possible because that position is occupied by the display. In such cases sound is reproduced through two loudspeakers placed symmetrically on either side of the CRT, two surround loudspeakers placed to the side and above the listening position. The two front loudspeakers can create a virtual (phantom) image that appears to originate from the exact center of the display provided that the listener is seated symmetrically with respect to the loudspeakers. With proper head and loudspeaker placement, it is possible to recreate a spatially-accurate soundfield with the correct frequency response in *one* exact position, the sweet spot. However, even in this static case, the sound originating from each loudspeaker arrives at each ear at different times (about 200 μs apart), thereby giving rise to acoustic crosstalk. These time differences combined with reflection and diffraction effects caused by the head lead to frequency response anomalies that are perceived as a lack of clarity.

This problem can be solved by adding a crosstalk cancellation filter to the signal of each loudspeaker. While this solution may be satisfactory for the static case, as soon as the listener moves even slightly, the conditions for cancellation are no longer met and the phantom image moves towards the closest loudspeaker because of the precedence effect. In order, therefore, to achieve the highest possible quality of sound for a non-stationary listener and preserve the spatial information in the original material it is necessary to know the precise location of the listener relative to the loudspeakers.

4.2 Video-Based Tracking

Computer vision has historically been considered problematic particularly for tasks that require object recognition. Up to now the complexity of vision-based approaches has prevented them from being incorporated into desktop-based integrated media systems. Recently, however, the Laboratory of Computational and Biological Vision at USC, under the direction of Prof. Christoph von der Malsburg, has developed a vision architecture that is capable of recognizing the identity, spatial position (pose), facial expression, gesture identification, and movement of a human subject, in real time.

This highly versatile architecture integrates a broad variety of visual cues in order to identify the location of a person's head within the image. Object recognition is achieved through pattern-based analysis that identifies convex regions with skin color that are usually associated with the human face, and a stereo algorithm that determines the disparities among pixels that have been moving [10]. This pattern-recognition approach is based on the Elastic Graph Matching method that places graph nodes at appropriate fiducial points of the

pattern [11]. A set of features is extracted at each graph node corresponding to the amplitudes of complex Gabor wavelets. The key advantage of this method is that a new pattern (face or ear) can be recognized on the basis of a small number of example images (10-100). Finally, there is a module that keeps track of object position over time. This is critical for audio applications in which the system must "remember" the last position of the listener that may have stopped moving. The tracking module uses a hysteresis mechanism to exploit time continuity and estimation of the current position and velocity of the head is achieved through a linear predictive filter.

While there are several alternative methods for tracking humans (*e.g.,* magnetic, ultrasound, infrared, laser), they typically based on tethered operations or require artificial fiducials to be worn by the user. Furthermore, these methods do not offer any additional functionality to match what can be achieved with vision-based methods (*e.g.,* face and expression recognition, ear classification).

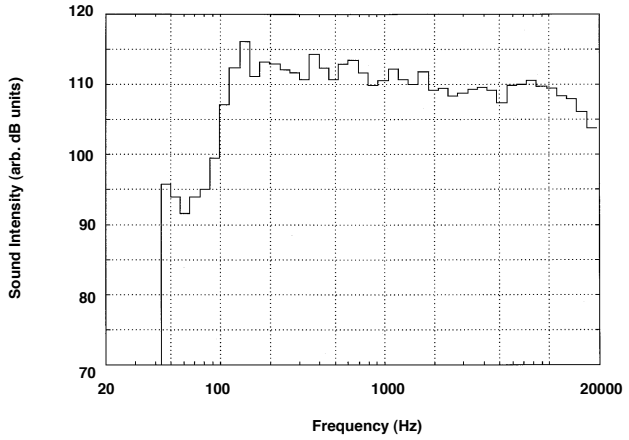

Figure 4. Frequency response of desktop loudspeaker system designed using the direct-path dominant and correct low-frequency response guidelines.

5. SUMMARY

We have developed a novel multichannel desktop audio system that meets all the design requirements and acoustical considerations described above (Fig. 4). This experimental desktop audio system uses two loudspeakers that are positioned on the sides of a video monitor at a distance of 45 cm from each other and 52 cm from the listener's ears. The seating position height is adjusted so that the listener's ears are at the tweeter level of the loudspeakers (117 cm from the floor) thus eliminating any colorations in the sound due to off-axis lobing. We have also incorporated the vision-based tracking algorithm described above using a standard video camera connected to an SGI Indy workstation. This tracking system provides us with the coordinates of the center of the listener's head relative to the loudspeakers and is currently capable of operating at 10 frames/sec with a 3% accuracy.

In this single-camera system we can track listener movement that is confined in a plane parallel to loudspeakers and at a fixed distance from them. When the listener is located at the exact center position (the sweet spot), sound from each loudspeaker arrives at the corresponding ear at the exact same time (*i.e.,* with zero ipsilateral time delay). At any other position of the listener in this plane, there is a relative time difference of arrival between the sound signals from each loudspeaker. In order to maintain proper stereophonic perspective, the ipsilateral time delay must be adjusted as the listener moves relative to the loudspeakers.

The head coordinates provided from the tracking algorithm are used to determine the necessary time delay adjustment. This information is processed by a 32-bit DSP processor board (ADSP-2106x SHARC) resident in a separate PC. In this early version, the DSP board is used to delay the sound from the loudspeaker that is closest to the listener so that sound arrives with the same time difference as if the listener were positioned in the exact center between the loudspeakers. Using this set-up we have demonstrated stereophonic reproduction with an adaptively-optimized sweet spot.

There are still several limitations that must be addressed. We are currently in the process of identifying the bottlenecks of both the tracking and the audio signal processing algorithms and integrating both into a single, PC-based platform for real-time operation. Furthermore, we are expanding the capability of the current single-camera system to include a second camera in a stereoscopic configuration that will provide distance information.

6. REFERENCES

[1] F. E. Toole, "Loudspeaker measurements and their relationship to listener preferences," *Journal of the Audio Engineering Society*, vol. 34, pp. 227-235, 1986.

[2] S. Bech, "Perception of timbre of reproduced sound in small rooms: influence of room and loudspeaker position," *Journal of the Audio Engineering Society*, vol. 42, pp. 999-1007, 1994.

[3] T. Holman, "Monitoring Sound in the One-Person Environment," *SMPTE Journal*, vol. October, 1997.

[4] F. E. Toole, "Subjective measurements of loudspeaker sound quality and listener performance," *Journal of the Audio Engineering Society*, vol. 33, pp. 2-32, 1985.

[5] H. Moller, "Fundamentals of Binaural Technology," *Applied Acoustics*, vol. 36, pp. 171-218, 1992.

[6] M. R. Schroeder and B. S. Atal, "Computer Simulation of Sound Transmission in Rooms," *IEEE International Convention Record*, vol. 7, 1963.

[7] D. H. Cooper and J. L. Bauck, "Prospects for Transaural Recording," *Journal of the Audio Engineering Society*, vol. 37, pp. 3-19, 1989.

[8] J. Bauck and D. H. Cooper, "Generalized Transaural Stereo and Applications," *Journal of the Audio Engineering Society*, vol. 44, pp. 683-705, 1996.

[9] S. Mehrgard and V. Mellert, "Transformation Characteristics of the External Human Ear," *Journal of the Acoustical Society of America*, vol. 51, pp. 1567-1576, 1977.

[10] O. Groetenherdt, "Video-Based Detection of Heads Using Motion and Stereo Vision (in German)," in *Institute for Neuroinformatics*. Bochum: University of Bochum, 1997.

[11] L. Wiskott, J. M. Fellous, N. Krueger, and C. von der Malsburg, "Face Recognition by Elastic Bunch Graph Matching," Institute for Neuroinformatics, Bochum 8, 1996.