Chris Kyriakakis,
Panagiotis Tsakalides,
and Tomlinson Holman

Acquisition and Rendering Methods for Immersive Audio

SURROUNDED by SOUND

Immersive audio systems are making strides in such applications as telepresence; augmented and virtual reality; entertainment; air traffic control, pilot warning, and guidance systems; displays for the visually impaired; distance learning; and professional sound and picture editing for television and film. Here, we examine signal processing issues that pertain to the acquisition and subsequent rendering of 3D sound fields over loudspeakers. On the acquisition side, recent advances in statistical methods for achieving acoustical arrays in audio applications are reviewed. Classical array signal processing addresses two major aspects of spatial filtering, namely localization of a signal of interest, and adaptation of the spatial response of an array of sensors to achieve steering in a given direction. The achieved spatial focusing in the direction of interest makes array signal processing a

1053-5888/99/$10.00©1999IEEE

necessary component in immersive sound acquisition systems. On the rendering side, 3D audio signal processing methods are described that allow rendering of virtual sources around the listener using only two loudspeakers. Finally, we discuss the commercial implications of audio DSP.

Immersive Audio Systems

Stereo sound reproduction has its origins in the U.K. where Blumlein developed a system in 1931 that could place a sound in the horizontal (azimuth) plane by using an appropriate combination of delay and level differences [1]. His work focused on the development of microphone techniques that would allow the recording of the amplitude and phase differences necessary for stereo reproduction. A few years later, Fletcher, Steinberg, and Snow, at Bell Laboratories in the U.S., discussed a "wall of sound," and concluded that an infinite number of microphones would be necessary to capture a sound scene [2]-[4]. At the reproduction end, an infinite number of loudspeakers would be required to reconstruct the sound field in a manner similar to the Huygens principle of secondary wavelets. The Bell Labs researchers proposed a practical implementation based on a three-channel system consisting of left, center, and right channels in the azimuth plane. Such a system could represent the lateralization and depth of the desired sound field with acceptable accuracy. The first such stereophonic three-channel system was demonstrated in 1934, with the Philadelphia Orchestra performing remotely for an audience in Washington, D.C., over wideband telephone lines.

The evolution of sound reproduction from its invention 60 years ago to today is in many respects similar to the evolution of orchestral music [5]. Early music, such as that of Mozart, was confined to a relatively small stage. Beethoven's symphonies widened the stage, as well as the dynamic range of the sound. Later, composers such as Berlioz and Mahler placed instruments or soloists off to the side or even behind the audience. Today, we experience music, sound, and image in films like "Terminator II" that utilize practically the entire audible dynamic range (peaks can reach 115 dB SPL) and present sounds that fully surround the listeners. However, even with today's cinema or home theater multichannel sound systems, sound is mainly confined in the horizontal (azimuth) plane. Immersive audio seeks to create a seamless, *fully 3D* aural environment that preserves the correct spatial sound localization for a listener that may be moving.

Sound Acquisition with Microphone Arrays

The acquisition of audio signals from spatially distributed sound sources can be achieved using several methods. For example, in order to implement high-quality teleconferencing or telepresence systems, it is necessary to address problems that are specific to these environments, such as background noise, reverberation, and echo cancellation. Microphone-array methods have been proposed for fast blind identification of multipath acoustic channels, as well as for audio signal copy i.e., for estimation of the transmitted acoustic waveform. Following are several key issues that arise in the implementation of such systems.

The wideband nature of the signal found at the source in audio applications adds complexity not commonly found in other signal processing applications.

Subspace Tracking of Speech and Sound

In applications such as hands-free audio communication and videoconferencing, microphone array processing has been used to suppress noise and speech reverberation that results from reflections induced by undesirable multipath propagation. The class of fixed or static arrays addresses this problem by using microphones with a very narrow spatial directivity. It has been shown that in an enclosed space, the direct-to-reverberant-power ratio at a given distance between the source and the receiver is related to the directivity factor of the sensor [6]. Combining the output of several such sensors can enhance the signal from a desired location and improve the quality of recorded speech. Fixed microphone arrays are robust because they do not make strict assumptions about the statistics of the acoustical environment or the signals. The downside is that their performance is limited by the number of sensors in the array. In addition, fixed arrays cannot handle complex, time-changing noise and reverberation environments.

Adjusting the adaptive array coefficients to match the sound field constitutes a complex inverse filtering problem. Traditional techniques combine the measurements at the array output by inserting time delays so that a direct acoustical source signal from a particular direction of interest adds coherently. This processing step, commonly known as *steering*, maximizes the array system response along a certain propagation path, and is justified by the fact that sources are usually modeled to propagate along planar or spherical waves. In real applications of speech acquisition in acoustic environments, spatially colored ambient noise, reflections, and reverberation cannot be neglected by the adaptive processor. In these cases, simple delay-and-sum beamforming forming may result in signal cancellation [7]. Hence, in random reverberant acoustic environments, it has been suggested that steering should be treated as a matched filtering problem that involves the inversion of the acoustic channel impulse response rather than a problem of time delay compensation of the direct path [8], [9].

Dereverberation of Acoustic Channels

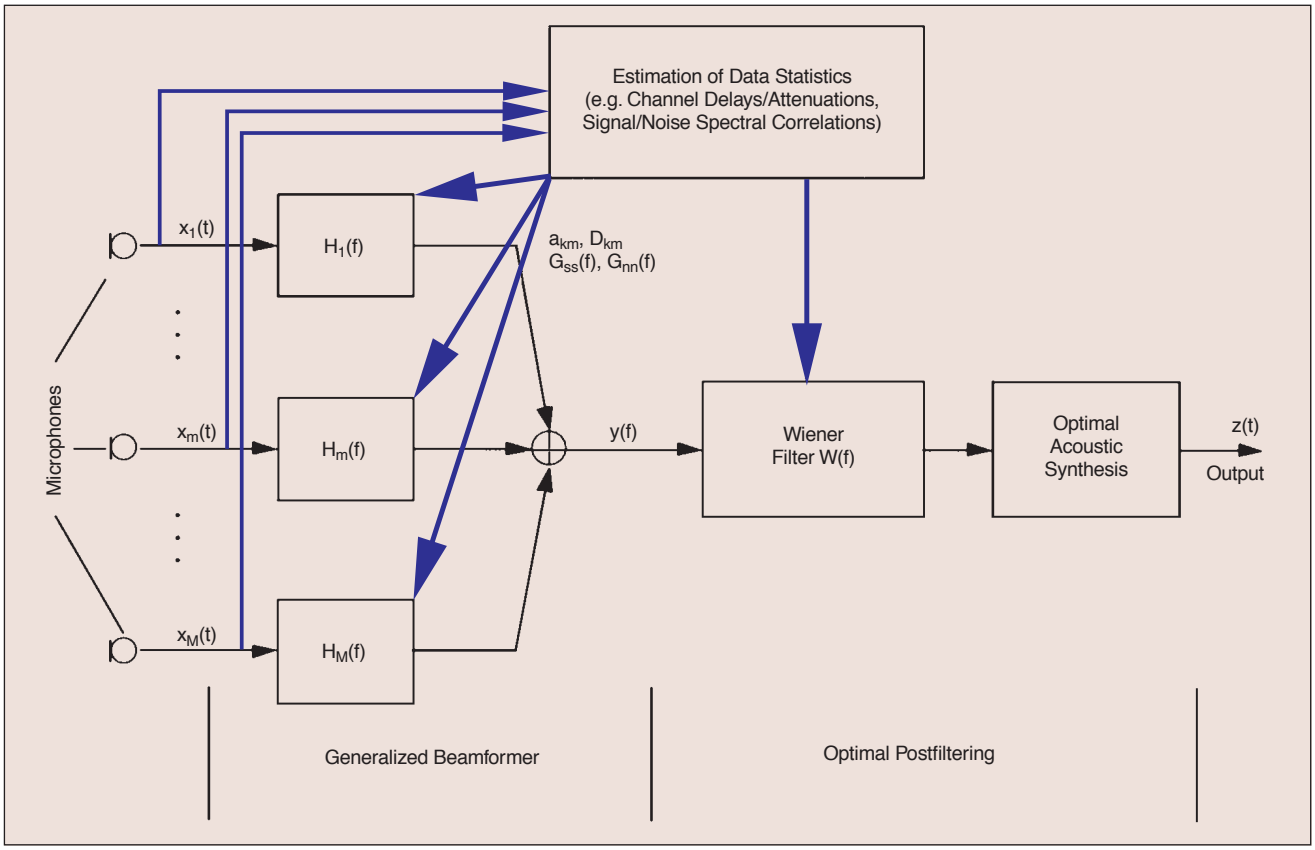
The inversion of the reverberant acoustic channel between the speaker and the sensor is complicated by the non-minimum phase characteristics of the channel [10]. Miyoshi and Kaneda solved the problem by using an array of sensors, and assuming that the acoustic paths between the sources and sensors do not have common zeros [11]. This method requires exact knowledge of the transfer functions of the paths and, thus, it is difficult to implement in real, time-varying acoustic environments.

Another family of dereverberation methods is based on microphone arrays followed by optimal postfiltering, which allows the removal of noncoherent parts of the recorded signal (Fig. 1). According to this class of systems, first described by Allen [12], the sum of properly delayed microphone signals is filtered by a time-varying postfilter, whose transfer function is derived from the cross-spectral densities (coherence functions) of the sensor signals. The underlying assumption is that the noise and reverberation form a diffuse acoustic field, and hence can be considered to be uncorrelated with the direct path of the desired signal. Simmer *et al.* were the first to derive a formal expression of the optimal postfilter based on the Wiener approach [13]. He also included a generalized sidelobe canceler to remove coherent noise components. Recently, Marro *et al.* performed the theoretical and experimental analysis of the methods to draw conclusions about their behavior in real environments [14]. They showed that when a strong correlation exists between the desired speech signal and the reverberant components, the improvement yielded by the postfilter is limited. However, in a videoconferencing context, the postfilter reduces other perturbations such as acoustic echo and localized noise sources.

Adaptive Beamforming

Classical, adaptive, statistically-optimum beamformers maximize the signal-to-noise ratio of the array output for optimal signal detection. The multiple side lobe canceler (MSC) is perhaps the earliest statistically-optimum beamformer introduced by Applebaum *et al.* [15]. Widrow *et al.* studied the adaptive antenna problem by minimizing the mean square error between the beamformer former output and a reference signal [16]. The array adapts its pattern to form a main lobe with its direction and bandwidth determined by the reference signal, and to reject signals and noise outside the main lobe. Frost generalized Widrow's approach into a constrained optimization problem with a goal of minimizing the array output power, while maintaining a certain frequency response in the desired direction [17].

Conventional adaptive beamforming methods cancel the desired signal when they operate in environments



▲ *1. Block diagram of a microphone array with optimal postfiltering. The system consists of three processing modules: the generalized beamformer attempts to compensate for the effects of the multipath acoustic environment, the Wiener filter tries to reduce the incoherent noise components that have been partially removed by the beamformer, and the data statistics estimator functions in tandem with the two other modules.*

The design challenge in desktop audio systems is to successfully map these standards onto the desktop environment through appropriate acoustical and psychoacoustical scaling and system design.

where the interference is correlated with the signal of interest. This is particularly true in hands-free speech acquisition using microphone arrays, because reflections from the enclosed room surfaces produce interference that is highly correlated with the useful signal. In the context of coherent signal classification, a spatial smoothing scheme, first suggested by Evans *et al.* [18], and subsequently studied by Shan *et al.* [19], was exploited to overcome this problem. The method involves a uniform linear array using sub-aperture sampling that essentially decorrelates the coherent signals. To deal with a reduction of the effective array aperture induced by smoothing techniques, Williams *et al.* [20] introduced the modified spatial smoothing method, specifying the conditions under which such a goal is reached. Other advanced techniques used forward and complex conjugate backward sub-arrays of the original array [21]. For the case of broadband signals, spectral averaging [22] as well as frequency-invariant beamforming has been employed to destroy the correlation [23].

Acoustic Arrays for Binaural Hearing Aids

Microphone arrays with binaural hearing aids combine the spatial filtering functionality of the array and the natural benefits of sound localization and speech intelligibility due to binaural listening. Currently, arrays of acoustic sensors with appropriate signal processing achieve directional selectivity to enhance sound sources from a specific direction, while canceling jamming and noise sources. Adaptive processing of the microphone signals is employed to cancel interference, with good results when the direct component of the interference is stronger than its reverberation [24], [25]. Recently, Desloge *et al.* pointed out that a user of array systems with one-channel output cannot localize sounds or perceive speech that may result naturally by ear. Hence, they developed and evaluated single- and dual-array, fixed-processing systems that achieve a compromise between the goals of a maximally directional response and a faithful preservation of binaural cues [26]. In related work, Welker *et al.* devoted the low-pass frequency part to binaural processing for left-right source localization, signal detection, and speech reception, and the high-pass frequency part to adaptive array processing for noise cancellation [27].

Optimization Methods for Sensor Positioning

One of the design aspects of a microphone array is concerned with the optimal positioning of the acoustic sensors. Using the Nyquist sampling theorem for a uniformly spaced linear array, sensor spacing should be at most $\lambda_{min}/2$, where λ_{min} is the smallest waveform wavelength to avoid spatial aliasing effects [28]. On the other hand, the antenna length should be at least $\lambda_{max}/\Delta\theta$, where λ_{max} is the greatest waveform wavelength, to achieve a desired main lobe beamwidth equal to $\Delta\theta$ in radians. Hence, the number of sensors should be greater than $2\lambda_{max}/\Delta\theta\lambda_{min}$, which may be too large for audio applications. Efforts to reduce the number of sensors are based on non-uniformly spaced arrays. In addition, circular and other planar multiple sensor configurations have been employed to expand the azimuth localization region of an array. Minimum redundancy linear arrays allow for increased array aperture by reducing the number of redundant spacings in the array [29]. Several optimization criteria are proposed by Gazor and Grenier in [30] and references therein. The described criteria are based on the quality of the array for beamforming, the array performance for linear filtering, minimum variance beamforming, and mutual information criteria.

New Advances in Robust Array Processing

Recently, the importance of extending the statistical array signal processing methodology to the so-called alpha-stable framework has been demonstrated. Array processing methods have been developed for a larger class of random processes, which include the Gaussian processes as special elements. The proposed methods, based on alpha-stable statistical theory and fractional lower-order statistics (FLOS), can be applied in environments which, while sharing many characteristics, also differ from Gaussian environments in significant ways. In particular, techniques for source detection and localization, and time delay estimation were developed, which perform optimally over a wide range of heavy-tailed noise environments [31]-[33].

Currently, research interests lie in wireless distributed multimedia applications, where the spatial selectivity of an array of sensors can be used to operate in close frequency bands and suppress undesirable noise and interference in favor of the signal of interest. Work has focussed on the development of methods for beam steering and waveform recovery. The objective has been to allow successful acoustic signal direction-of-arrival tracking in the presence of a large number of side-lobe interferers in a possibly hostile environment.

Several major design considerations are involved in the aforementioned scenario, in which moving antenna arrays play an important role. First, signal processing design must accommodate arrays with a time-varying manifold. The calibration of such arrays will have to be done blindly because in many situations it is either impossible or impractical to deploy calibrating sources. Second,

the environment can be highly non-stationary and unpredictable, characterized by unfavorable ever-changing propagation conditions due to the presence of interferers and jammers. Therefore, the design of robust microphone arrays that perform well in a wide range of interference and noise environments is of great importance. Third, the proposed solutions must be impervious to problems in hardware. Because it is costly and sometimes impossible to replace faulty hardware, the proposed methods should be able to work by suitable processing of the remaining available measurements. Finally, audio applications require beamforming techniques that are general enough to work for wideband signals.

Time-Delay Estimation with Two-Sensor Arrays

The problem of microphone-array steering and sound source localization is of great interest in immersive telepresence and teleconferencing applications. For example, it is often required to automatically redirect a video camera so that the person speaking is in the field-of-view. In multi-participant environments it is desirable to provide spatially-selective speech acquisition as well as noise and echo cancellation.

Inter-sensor Time-Delay Estimation (TDE) is a method commonly used (see [34] and references therein) to estimate the position of a source using bearing information. The majority of TDE methods proposed so far in audio applications use second- or higher-order statistics of the measurements to locate the signal of interest. A drawback of these methods is that they are not robust in heavy-tailed noise or severe interference environments.

The wideband nature of the signal found at the source in audio applications adds complexity not commonly found in other signal processing applications. Additionally, the *a priori* noise and interference statistics are not known, and they may vary with time. Traditional Gaussian modeling of noise signals fails when the signals exhibit impulsive behavior. Therefore, a new model is used, namely the Symmetric α-Stable (SaS), which can better account for the outliers that exist in real-world signals.

In the following, we formulate the TDE problem, we describe the Phase Transform (PHAT) method for TDE, and we present a recently proposed variation of PHAT that is based on the Fractional Lower-Order Statistics (FLOS-PHAT) of the received signals. The FLOS-PHAT stands up to the presence of a severe noise background.

Robust TDE in Heavy Noise

Consider the two-element microphone array system, shown in Fig. 2, which measures the signals

$$r_1(t) = x(t) + n_1(t) \text{ and } r_2(t) = x(t-\tau) + n_2(t) \qquad (1)$$

in which the noise components $n_1(t)$ and $n_2(t)$ are assumed to be zero mean and uncorrelated with the desired speech signal $x(t)$. The goal is to estimate the delay τ from the two microphone measurements in order to localize the sound source $x(t)$. Transforming the measurements into the frequency domain yields

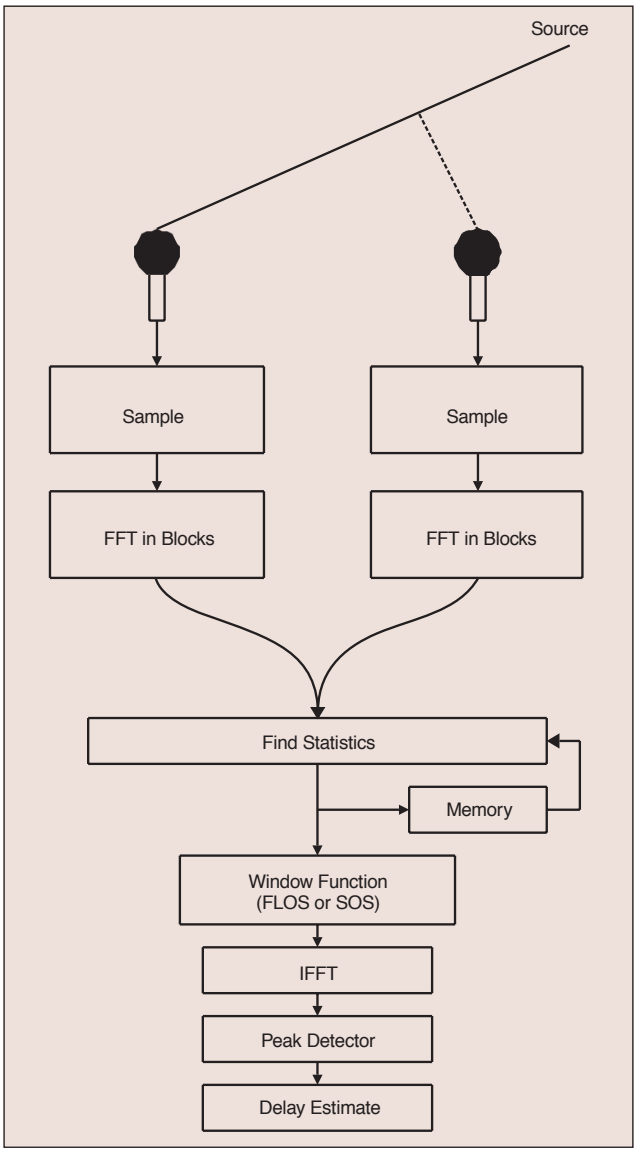
$$\begin{aligned} R_1(k) &= \left[X(k) + N_1(k)\right] \\ R_2(k) &= \left[X(k)\cdot e^{-j\omega_k \tau} + N_2(k)\right]. \end{aligned} \qquad (2)$$

Then, the cross-spectrum of the two signals can be found from (2) to be

$$C_{R_1R_2}(k) = E\left\{R_1(k)\cdot R_2^*(k)\right\} = \sigma_x^2 e^{j\omega_k \tau} \qquad (3)$$

where $\sigma_X^2 = E\{|X(k)|^2\}$ is the signal power.

PHAT method [35] smoothes the cross-spectrum $C_{R_1R_2}(k)$ by a window inversely proportional to the mag-



▲ *2. Block diagram of an array system with two microphone sensors used for time-delay estimation.*

nitude of the cross-spectrum. This results in a weighted cross-correlation function

$$C^{w}_{R_1R_2}(k) = \frac{C_{R_1R_2}(k)}{\left|C_{R_1R_2}(k)\right|} = e^{j\omega_k \tau} \tag{4}$$

whose inverse Fourier transform to the time domain generates a peak corresponding to the value of the delay τ. The PHAT method performs well when the noise components are Gaussian. However, when the noise deviates from the ideal Gaussian assumption, and is better characterized by the alpha-stable class of distributions, the performance of PHAT degrades significantly.

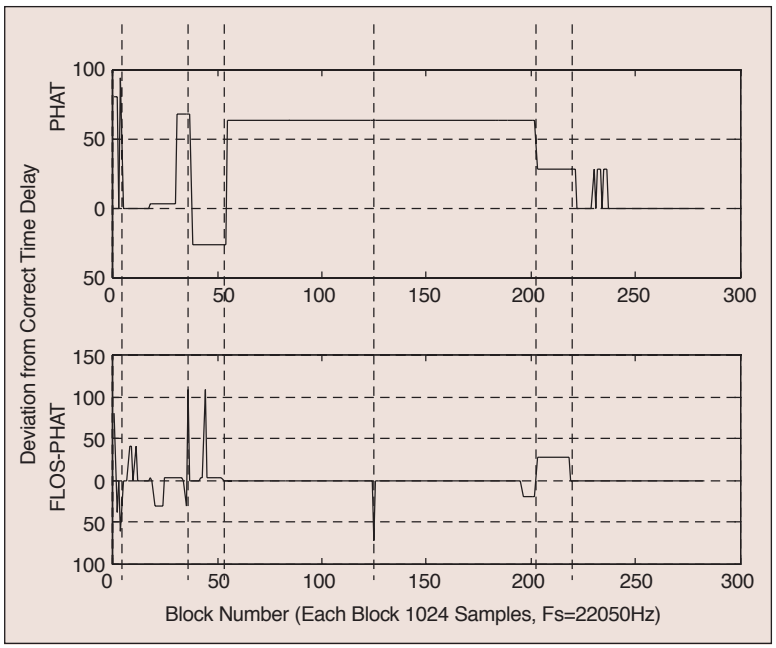
A dual tool to the second-order cross-correlation function can be defined based on the FLOS of the measurements [36]. The fractional lower-order correlation function is expressed as

$$A_{R_1R_2}(k) = E\left\{R_1^*(k)^{\langle a \rangle} \cdot R_2(k)^{\langle b \rangle}\right\} \tag{5}$$

where $y^{\langle p \rangle} = |y|^{p-1} y^* (0 < p < 1)$ is called the "signed-power non-linearity," and affects only the magnitude of the measurement. Clearly, when $a = b = 1$, the above definition reduces to the usual cross-correlation function between the measurements. Fractional lower-order statistics of the form (5) have been used in the design of signal-processing algorithms that can handle the existence of heavy-tailed noise in the data [36]. The FLOS-PHAT method uses a smoothed version of the fractional lower-order correlation function defined in (5) to estimate the time delay in a robust way:

$$A^{w}_{R_1R_2}(k) = \frac{A_{R_1R_2}(k)}{\left|A_{R_1R_2}(k)\right|}. \tag{6}$$

The TDE accuracy of the PHAT method, vis-a-vis the FLOS-PHAT technique is shown in Fig. 3, using actual sound measurements in heavy noise. The figure depicts the reaction of the two algorithms to the occurrence of noise outliers in the measurements. It is apparent that the FLOS-PHAT time-delay estimates are not influenced as much by the presence of heavy noise as the PHAT estimates.



▲ *3. Transient performance of the PHAT (top) and the FLOS-PHAT (bottom) methods for time-delay estimation of an audio signal in heavy noise. The figure shows the offset of the estimated TDE from the actual value. The use of fractional lower-order statistics makes the FLOS-PHAT method more resistent to the presence of noise outliers in the sound measurements.*

Immersive Audio Rendering

The methods for sound acquisition described above represent one important element of immersive audio. The other element pertains to the seamless reproduction of real and virtual sound sources in 3D space around a listener. In this section, we discuss several issues that relate to immersive audio reproduction over loudspeakers.

Human sound-source localization is based on the estimation of frequency-dependent differences in intensity and time of arrival at both ears for source localization in the horizontal (azimuth) plane. In the median plane time, differences are constant. Localization is based on spectral filtering by the outer ear. The reflection and diffraction of sound waves from the head, torso, shoulders, and pinnae, combined with resonances caused by the ear canal, form the physical basis for the head-related transfer function (HRTF). This system can be modeled as linear and time-invariant, and is fully characterized by the HRTF in the frequency domain [37].

Immersive audio-rendering systems are based on digital implementations of such head-related transfer functions. The spectral information provided by the HRTF can be used to implement a set of filters that can process non-directional (monaural) sound to simulate a real HRTF. Early attempts in this area calculated the attenuation and delay of the sound field by the head by assuming a simplified spherical head model [38]. More recent methods are based on the measurement of HRTF's for each desired virtual sound source direction using a microphone placed inside the ear canals of a mannequin (see [39] and references therein). The main advantage of measured HRTF's is that they account for the effects of the pinnae, diffraction from the head, and reflections from the upper body.

In principle, it is possible to achieve excellent reproduction of 3D sound fields using such methods, however, this requires precise measurement of each listener's individual HRTF's. In fact, the magnitude and phase of these head-related transfer functions vary significantly not only for each sound direction, but

Just map the conventional processes already available into the digital domain, execute with DSP processing, and voilá, better quality over analog techniques.

also from person to person. Current research in this area is focused on achieving good localization performance while using synthetic (non-individualized) HRTF's derived through averaging, modeling, or by using HRTF's of subjects that have been determined to be "good localizers" [40]-[42]. Begault [43] found that there are currently three major barriers in 3D audio implementations: psychoacoustic errors, such as front-to-back confusion; the requirement of long FIR filter lengths to accurately represent measured HRTF's; and frequency- and phase-response errors that arise from mismatches between non-individualized and measured HRTF's.

Crosstalk Cancellation for Loudspeaker Rendering

Delivery of immersive audio can be achieved through headphones or loudspeakers. Here, we focus our discussion on loudspeaker methods for two reasons: There is a large installed base of desktop computers with two loudspeakers on either side of the monitor, and headphone delivery requires the user to be tethered to the system. While with headphones it is possible to deliver the appropriate sound field to each ear, with loudspeakers it is necessary to eliminate the crosstalk. This crosstalk arises because each loudspeaker sends sound to the same-side (ipsilateral) ear, as well as undesired sound to the opposite-side (contralateral) ear.

Several methods have been proposed to address crosstalk cancellation. The first such scheme was proposed by Atal and Schroeder [44], and later again by Damaske and Mellert [45]. The main limitation of these early systems was the fact that any listener movement that exceeded 75 to 100 mm completely destroyed the spatial effect. A method proposed by Cooper and Bauck modeled the head as a sphere, and then calculated the ipsilateral and contralateral terms [38]. A similar method by Gardner approximates the effect of the head with a low-pass filter, a delay, and a gain (less than 1) [46]. Cooper and Bauck [47], [48] 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.

The two-loudspeaker, two-ear system can be fully characterized by a 2×2 matrix of transfer functions in the frequency domain (Fig. 4). In order to render a virtual sound source in a particular direction, it is necessary to deliver to the left and right eardrums of the listener the signals E_L and E_R, respectively,

$$\begin{bmatrix} E_L \\ E_R \end{bmatrix} = \begin{bmatrix} H_L & 0 \\ 0 & H_R \end{bmatrix} \begin{bmatrix} S \\ S \end{bmatrix} \tag{7}$$

in which H_L is the virtual sound source HRTF for the left ear, H_R is the virtual sound source HRTF for the right ear, and S is the monaural input sound.

Rendering over loudspeakers, however, introduces an additional transfer matrix that includes ipsilateral (H_i) and contralateral (H_c) terms, and as a result, the actual signals arriving at the ears are given by

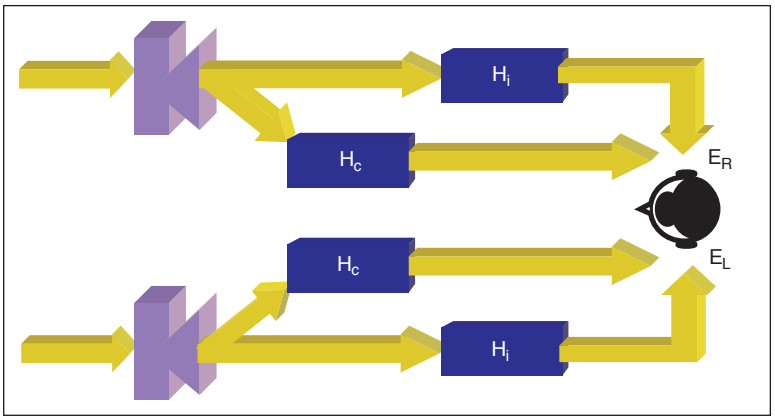
$$\begin{bmatrix} E'_L \\ E'_R \end{bmatrix} = \begin{bmatrix} H_i & H_c \\ H_c & H_i \end{bmatrix} \begin{bmatrix} H_L & 0 \\ 0 & H_R \end{bmatrix} \begin{bmatrix} S \\ S \end{bmatrix}. \tag{8}$$

Note that if the position of the listener changes over time, then the ipsilateral and contralateral transfer functions will not be symmetrical as shown above, but will vary to reflect the HRTF's for the listener's new position. Adjustments for these variations can be achieved through tracking of the listener's head in 3D space [49]-[52]. In the analysis below, we assume that the listener is seated symmetrically, with respect to the loudspeakers.

In order to deliver the signals in (7), given that the physical system results in (8), it is necessary to preprocess the input signal S by inverting the matrix introduced by the loudspeakers

$$\begin{bmatrix} E_L \\ E_R \end{bmatrix} = \begin{bmatrix} H_i & H_c \\ H_c & H_i \end{bmatrix} \begin{bmatrix} H_i & H_c \\ H_c & H_i \end{bmatrix}^{-1} \begin{bmatrix} H_L & 0 \\ 0 & H_R \end{bmatrix} \begin{bmatrix} S \\ S \end{bmatrix}. \tag{9}$$

This can be written as



▲ *4. With a loudspeaker-based rendering system, the listener's ears receive the desired ipsilateral sound (represented by the transfer functions H_i) , as well as undesired contralateral sound (represented by the transfer functions H_c) . In order to render the correct signal in each ear, it is necessary to remove the crosstalk terms.*

$$\begin{bmatrix} E_L \\ E_R \end{bmatrix} = \begin{bmatrix} H_i & H_c \\ H_c & H_i \end{bmatrix} \begin{bmatrix} 1 & H_c / H_i \\ -H_c / H_i & 1 \end{bmatrix} \begin{bmatrix} H_L / H_i & 0 \\ 0 & H_R / H_i \end{bmatrix} \begin{bmatrix} S \\ S \end{bmatrix} \tag{10}$$

under the assumption that the second-order ipsilateral term is much larger than the second-order contralateral term (i.e., $1 - H_c^2 / H_i^2 \cong 1$). This assumption is justified due to the natural "crosstalk cancellation" that arises from shadowing of the opposite ear by the human head.

The terms H_L / H_i and H_R / H_i in (10) correspond to inversion of the HRTF's introduced by the loudspeakers, while the off-diagonal terms $-H_c / H_i$ represent the crosstalk cancellation. The required loudspeaker signals are then found to be

$$X_L = \left(\frac{H_L}{H_i} - \frac{H_c}{H_i} \frac{H_R}{H_i} \right) S = F_L S$$
$$X_R = \left(\frac{H_R}{H_i} - \frac{H_c}{H_i} \frac{H_L}{H_i} \right) S = F_R S \tag{11}$$

in which F_L and F_R are the filters that must be implemented for preprocessing the input signal so that it is perceived from the desired direction.

Inversion of Non-Minimum Phase Filters

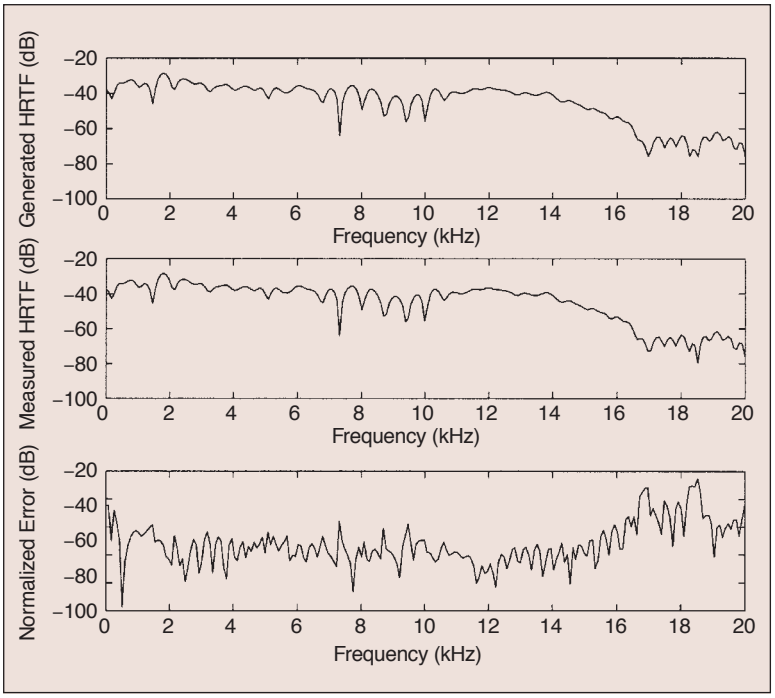
The implementation of the filters F_L and F_R shown in (11) is complicated by the fact that the ipsilateral transfer function (H_I) is a non-minimum phase filter. A typical solution to this problem is to use cepstrum analysis to design a new filter with the same magnitude as H_i but keep it minimum phase at the same time[53]. The drawback is that information contained in the excess phase is lost. Jot *et al.* [54] presented a method in which every HRTF is separated into a minimum phase, and an all-pass component is approximated by a linear phase response.

Another method has been proposed [55] that maintains the HRTF phase information. The non-causal, but stable impulse response that corresponds to H_x/H_i in a different Region of Convergence is found. A delay is then introduced in order to make the filter causal. The trade-off and the corresponding challenge is to make the delay small enough to be imperceptible to the listener, particularly when there is associated video material required to be synchronized to the sound.

One possible implementation of this method is by using an adaptive algorithm. The problem of finding the filter H_{inv} such that the mean squared error $E[e(n)]$ is minimized, can be classified as a combination of a system identification and inverse modeling problem. Its solution can be based on standard adaptive methods such as the LMS algorithm [56]. An example of an HRTF generated using this method is shown in Fig. 5. It can be seen that the resulting synthetic HRTF is in very good agreement with the originally measured HRTF for that azimuth and elevation. The FIR filter length, as well as the delay d, are selected so that the mean squared error is minimized. Furthermore, iterative adaptation of the step size μ leads to faster convergence as well as fewer incorrect adjustments. This method is particularly well-suited to real-time applications, in which it is important to account for movement of the virtual sound source positions and movement of the listener's head. A current implementation is under development that incorporates this algorithm with a video-based head-tracking algorithm that can operate in real-time on a Pentium/Win NT platform [57]. In parallel with FIR filter methods, IIR methods are also being examined for HRTF implementations in order to reduce the computational complexity inherent in long FIR filters [54], [58]-[60].

Acoustical and Psychoacoustical Considerations for Desktop Rendering

A significant amount of work in the area of high-quality sound production and reproduction has originated from the film industry. A well-defined set of standards has been developed for sound monitoring conditions in dubbing stages to ensure the transparent reproduction of program material in theaters. Such standards include loudspeaker



▲ *5. Top: HRTF generated using the LMS method described in the text. Middle: Original measured HRTF for 0° azimuth and 0° elevation. Bottom: Signal-to-error ratio of the generated and original HRTF's.*

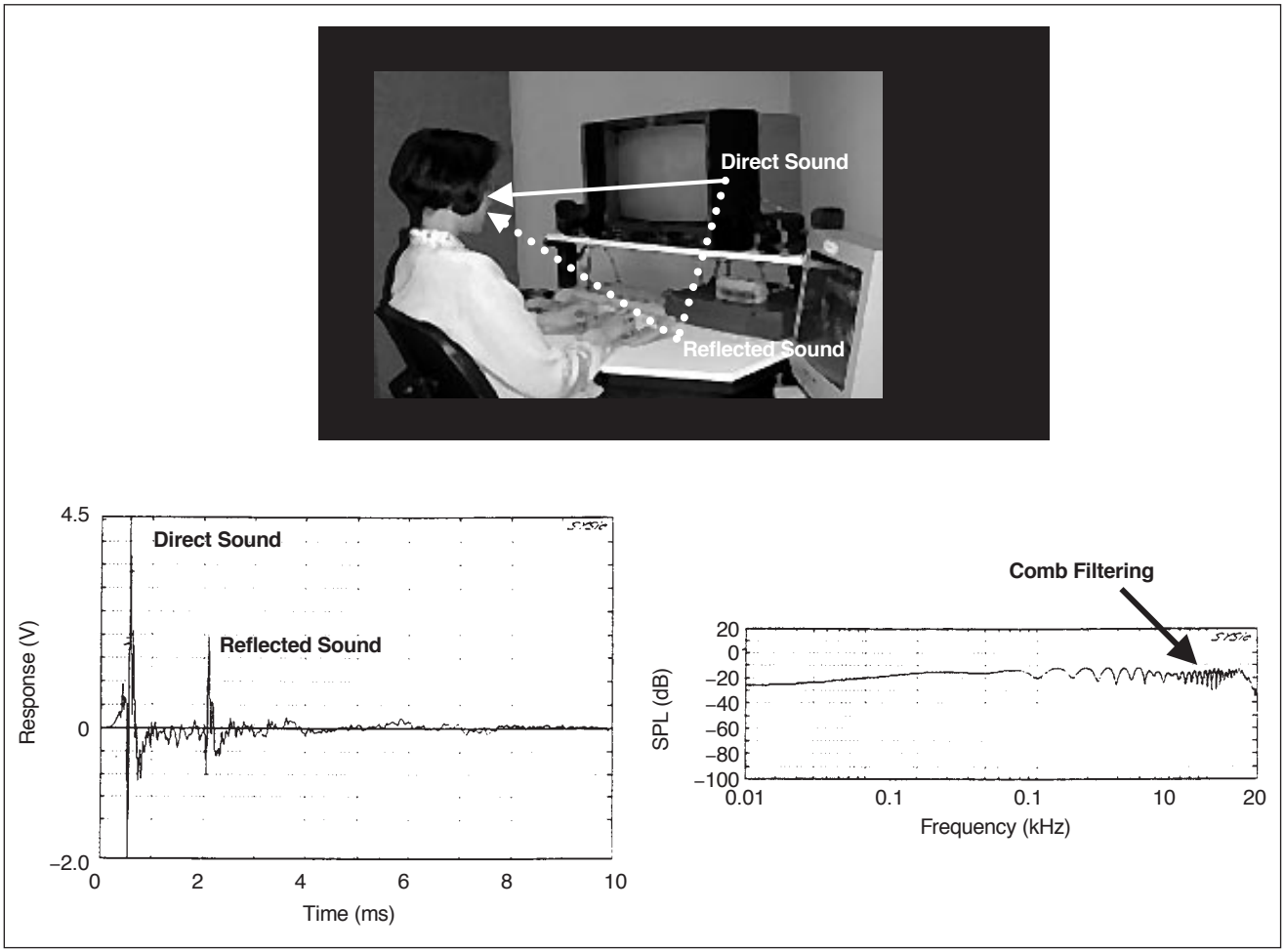
positioning for multichannel monitoring, loudspeaker frequency response and directivity requirements, precise sound-pressure level calibration, control of room acoustics parameters (such as reverberation time and discrete reflections), and background noise levels. Meeting these standards ensures that material produced in one professional dubbing stage can be monitored under identical conditions in another dubbing stage or in a movie theater. The design challenge in desktop audio systems is to successfully map these standards onto the desktop environment through appropriate acoustical and psychoacoustical scaling and system design.

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 problems that contribute to audible distortion of the reproduced sound [61]. Among these problems, the one most often neglected is the effect of discrete early reflections. The effects of such reflections on sound quality have been studied extensively [61]-[65], and it has been

The noise of the living space tends to be low-frequency in origin, and the noise of the DAC tends to be white with an emphasis on high frequencies.

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 as frequency-response colorations in rooms with an early reflection level that exceeds –15 dB in spectrum level, relative to the direct sound for the first 15 ms. Such a high level of reflected sound gives rise to comb filtering in the frequency domain that, in turn, causes noticeable changes in timbre (Fig. 6).

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



▲ *6. The local acoustical environment can influence the quality of the reproduced sound. Sound reflected from nearby surfaces interferes with the direct sound and gives rise to audible comb filtering. The bottom left plot shows a time-domain impulse measurement, with the direct sound followed by a second peak due to the reflection from the flat table surface. The frequency response of the combined effect is shown on the bottom right.*

effects to inaudible levels. 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, which give rise to large variations in frequency response.

To address the problems described above, a set of solutions has been developed for single-listener desktop reproduction that delivers sound quality equivalent to a calibrated dubbing stage [49], [61]. These solutions include:

Direct-path dominant design. By combining elements of psychoacoustics in the system design, it is possible to place the listener in the direct sound field that is dominant over the reflected and reverberant sound. The colorations that arise from these reflections are eliminated, resulting 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 rests.

Correct low-frequency response. There are problems in the uniformity of low-frequency response that arise from the standing waves associated with the acoustics of small rooms. Such anomalies can produce 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 positions 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. Due to the limitations of small-room acoustics, 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 [57].

Digital Signal Processing for Consumer Audio Applications

When digital signal processing was first introduced in the early 1980s, it looked as if most audio-related signal processing would be performed by DSP chips within a few short years. Manufacturers' application notes for components such as graphic equalizers seemed to show the way: Just map the conventional processes already available into the digital domain, execute with DSP processing, and voilá, better quality over analog techniques.

In fact, one reason that audio signal processing was used to demonstrate the capabilities of the first DSP chips is that it was audio hobbyists at semiconductor companies who led the way by showing off their technology to the rest of the world, through audio applications. However, they did so without regard to the sensitive cost-performance trade-offs of the consumer market. (After all, they had a back room full of chips!) Also, such early designs may not, in fact, have considered all of the problems associated with mapping from the analog to the digital domain. These problems include frequency warping of equalization and filter characteristics when converting to a sampled data system, as well as complications from gain changes between a properly dithered input linear PCM signal and an output signal. If the signal is attenuated, it becomes under-dithered on the output, revealing quantization distortion of output DACs. Another reason that DSP has not taken over the consumer audio market is the difficulty in programming DSP processors, which are largely hand-coded for optimum performance.

Audio designers are becoming knowledgeable about DSP, or at least looking for DSP solutions as prices drop and performance improves. Additional application areas once thought too difficult or expensive to consider in analog technology are now becoming interesting topics. In the last several years, one principle application area has emerged where DSP is essential, because other techniques are just too slow and expensive, or require large capital investment and quick time to market. This is the area of low-bit-rate coding systems. To date, about 785,000 DSP decoders have been sold, incorporating AC-3 or DTS multichannel codec algorithms, as well as several million MPEG and AC-2 two-channel codecs. With design wins in the ATSC digital television system and DVDs now being fielded in the U.S., geometric growth rates for low-bit-rate coding applications are expected in the next few years. In turn, this will bring costs down, as volume builds and expected performance gains are realized.

What low-bit-rate processing has shown is that incorporating psychoacoustic principles can be most productive, with greater than ten-fold compression possible with only fairly minor degradation of the original quality. Low-bit-rate coding uses two elements from psychoacoustics—frequency and temporal masking—to do the job. There are many more principles of psychoacoustics that remain to be exploited. These include:

The loudness effect. This is the reason for the loudness control on audio equipment. It is based on the psychoacoustic effect that human hearing is not as sensitive to low frequencies as it is to middle and high frequencies. This sensitivity varies with reproduction level, more severe as the playback level is reduced from the original level. Previous solutions to this problem have been plagued with a lack of calibration standards, and no understanding that it is the differential equal-loudness contours of hearing that

need to be derived to obtain correct results. Once such psychoacoustic effects are understood, and calibration made available through a tie in between electrical and acoustical levels, a system can be designed to overcome what is a very real defect in most people's hearing [66].

Room equalization. Research conducted at the National Research Council of Canada [62, 63, 67, 68] has shown that the effects of room acoustics on reproduced sound are far greater than differences between loudspeaker brands. This problem has long been recognized, but is difficult to solve in small rooms primarily due to the fact that it is theoretically impossible to invert the room response [10]. However, by applying psychoacoustics, in addition to acoustic principles, it may be possible to develop a transparent inversion method, rather than one that solves the general response problem in rooms.

Most high-end equipment today employs DACs with more than 16-bits. Practically speaking, this is in order to ensure monotonicity to the 16-bit level, and good differential non-linearity performance, since most source material has a 16-bit dynamic range. The problem with 16 bits is that the dynamic range is about 95 dB from instantaneous peak to broadband noise levels. If the replay level is set to motion picture theater standards, 0 dBFS will be set to a 105-dB sound pressure level. This means that –95 dBFS will correspond to 10 dB SPL. While this seems like a particularly low level, in fact it was shown to be above the audible threshold in a survey of 50 living rooms [69]. The noise of the living space tends to be low-frequency in origin, and the noise of the DAC tends to be white with an emphasis on high frequencies. Thus, for good performance, an external level control following the DAC is necessary. This needs to be interlocked with the DAC, with gain ranging performed by a combination of post-DAC analog gain control within the digital domain gain control before the DAC. Even so, 16-bit performance, in terms of noise level, is considered limiting, and 20-bit performance may solve this DAC problem, offering a peak replay level of 120 dB, with a noise floor of 0 dB SPL.

These are a few examples in which psychoacoustic knowledge can be brought to bear on audio problems that are perhaps best solved by DSP techniques. Impediments due to cost-performance trade-offs and difficulty in programming are decreasing, and with new immersive audio technology on the horizon, application areas are increasing.

Chris Kyriakakis is an Assistant Professor in the Electrical Engineering Department at the University of Southern California (USC) and an investigator in the Integrated Media Systems Center, a National Science Foundation Engineering Research Center at USC. *Dr. Panagiotis Tsakalides* is a Research Assistant Professor with the Signal and Image Processing Institute at the Department of Electrical Engineering—Systems, at USC. Tomlinson Holman is an Associate Professor of film sound at the USC School of Cinema-Television.

References

[1] A.D. Blumlein, "Improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems," U.K. Patent no. 394325, 1931.

[2] H.Fletcher, "Auditory perspective—Basic requirements," *Elect. Eng.*, vol. 53, pp. 9-11, 1934.

[3] J.C. Steinberg and W. B. Snow, "Physical Factors," *Bell Sys. Tech. J.*, vol. 13, pp. 245-258, 1934.

[4] W.B. Snow, "Basic principles of stereophonic sound," *SMPTE Journal*, vol. 61, pp. 567-589, 1953.

[5] T. Holman, "The Number of Audio Channels," presented at AES 100th Convention, Copenhagen, Denmark, 1996.

[6] H. Kuttruff, *Room Acoustics*. Amsterdam, The Netherlands: Elsevier, 1991.

[7] B. Widrow, K.M. Duvall, R.P. Gooch, and W.C. Newman, "Signal cancellation phenomena in adaptive antennae: Causes and cures," *IEEE Trans. Antennas Prop.*, vol. 30, pp. 469-478, 1982.

[8] J.L. Flanagan, A.C. Surendran, and E.E. Jan, "Spatially selective sound capture for speech and audio processing," *Speech Commun.*, vol. 13, pp. 207-222, 1993.

[9] S. Affes and Y. Grenier, "A signal subspace tracking algorithm for microphone array processing of speech," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 5, pp. 425-437, 1997.

[10] S.T. Neely and J.B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Am.*, vol. 66, pp. 165-169, 1979.

[11] M. Miyoshi and Y. Kaneda, "Inverse filtering of room impulse response," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 36, pp. 145-152, 1988.

[12] J.B. Allen, D.A. Berkley, and J. Blauert, "Multimicrophone signal processing technique to remove reverberation from speech signals," *J. Acoust. Soc. Am.*, vol. 62, pp. 912-915, 1977.

[13] K.U. Simmer, S. Fischer, and A. Wasiljeff, "Suppression of coherent and incoherent noise using a microphone array," *Ann. Télécommun.*, vol. 49, pp. 439-446, 1994.

[14] C. Marro, Y. Mahieux, and K.U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 6, pp. 240-259, 1998.

[15] S.P. Applebaum and D.J. Chapman, "Adaptive arrays with main beam constraints," *IEEE Trans. Antennas Prop.*, vol. 24, pp. 650-662, 1976.

[16] B. Widrow, P.E. Mantey, L.J. Griffiths, and B.B. Goode, "Adaptive antenna systems," *Proc. IEEE*, vol. 55, pp. 2143-2159, 1967.

[17] O.L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, pp. 926-935, 1972.

[18] J. Evans, J. Johnson, and D. Sun, "High-resolution angular spectrum estimation techniques for terrain scattering analysis and angle of arrival estimation," presented at 1st ASSP Workshop on Spectral Estimation, 1981.

[19] T.J. Shan, M. Wax, and T. Kailath, "On spatial smoothing for direction-of-arrival estimation of coherent signals," *IEEE Trans. Acoust., Speech, and Signal Process.*, 1985.

[20] R. Williams, S. Prasad, A.K. Mahalanabis, and L. Sibul, "An improved spatial smoothing technique for bearing estimation in a multipath environment," *IEEE Trans. Acoust., Speech, and Signal Process.*, 1988.

[21] S. Pillai and B. Kwon, "Forward/backward spatial smoothing techniques for coherent signal identification," *IEEE Trans. Acoust., Speech, and Signal Process.*, 1989.

[22] J.F. Yang and M. Kaveh, "Coherent signal-subspace transformation beamformer," *IEE Proceedings-F*, vol. 137, pp. 267-275, 1990.

[23] D.B. Ward, "Technique for broadband correlated interference rejection in microphone arrays," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 6, pp. 414-417, 1998.

[24] J.E. Greenberg, P.M. Peterson, and P.M. Zurek, "Intelligibility-weighted measures of speech-to-interference ratios and speech system performance," *J. Acoust. Soc. Am.*, vol. 94, pp. 3309-3010, 1993.

[25] M.W. Hoffman, T.D. Trine, K.M. Buckley, and D.J.V. Tasell, "Robust adaptive microphone array processing for speech enhancement," *J. Acoust. Soc. Am.*, vol. 96, pp. 759-770, 1994.

[26] J.G. Desloge, W.M. Rabinowitz, and P.M. Zurek, "Microphone array hearing aids with binaural output, Part I: Fixed-processing systems," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 5, pp. 529-542, 1997.

[27] D.P. Welker, J. Greenberg, J.G. Desloge, and P.M. Zurek, "Microphone array hearing aids with binaural output, Part II: Two-microphone adaptive systems," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 5, pp. 543-551, 1997.

[28] D. Johnson and D. Dudgeon, *Array Signal Processing: Concepts and Techniques*. Englewood Cliffs: Prentice Hall, 1993.

[29] A.T. Moffet, "Minimum-redundancy linear arrays," *IEEE Trans. Antennas Prop.*, vol. 16, pp. 172-175, 1968.

[30] S. Gazor and Y. Grenier, "Criteria for positioning sensors for a microphone array," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 3, pp. 294-303, 1995.

[31] P. Tsakalides and C.L. Nikias, "Maximum likelihood localization of sources in noise modeled as a stable process," *IEEE Trans. Signal Process.*, vol. 43, pp. 2700-2713, 1995.

[32] P. Tsakalides and C.L. Nikias, "The robust covariation-based MUSIC (ROC-MUSIC) algorithm for bearing estimation in impulsive noise environments," *IEEE Trans. Signal Process., vol. 44, pp. 1623-1633, 1996.*

[33] P.G. Georgiou, C. Kyriakakis, and P. Tsakalides, "Robust time-delay estimation for sound source localization in noisy environments," presented at 1997 Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk, New Paltz, NY, 1997.

[34] M.S. Brandstein, "A Framework for speech source localization using sensor arrays," Brown University, Ph.D. Dissertation 1995.

[35] C.H. Knapp and G.C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. ASSP-24, pp. 320-327, 1976.

[36] C.L. Nikias, *Signal Processing with Alpha-Stable Distributions and Applications*. New York: John Wiley and Sons, 1995.

[37] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization, Revised Edition*. Cambridge, Massachusetts: MIT Press, 1997.

[38] D.H. Cooper, "Calculator program for head-related transfer functions," *J. Aud. Eng. Soc.*, vol. 30, pp. 34-38, 1982.

[39] W.G. Gardner and K.D. Martin, "HRTF measurements of a KEMAR," *J. Acoust. Soc.,* vol. 97, pp. 3907-8, 1995.

[40] F.L. Wightman and D.J. Kistler, "Headphone simulation of free-field listening: psychophysical validation," *J. Acoust. Soc. Am.*, vol. 85, pp. 868-78, 1989.

[41] F.L. Wightman, D.J. Kistler, and M. Arruda, "Perceptual consequences of engineering compromises in synthesis of virtual auditory objects," *J. Acoust. Soc. Am.*, vol. 101, pp. 1050-1063, 1992.

[42] E.M. Wenzel, M. Arruda, and D.J. Kistler, "Localization using nonindividualized head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 94, pp. 111-23, 1993.

[43] D.R. Begault, "Challenges to the successful implementation of 3D sound," *J. Aud. Eng. Soc.*, vol. 39, pp. 864-70, 1991.

[44] M.R. Schroeder and B.S. Atal, "Computer simulation of sound transmission in rooms," *IEEE International Convention Record*, vol. 7, 1963.

[45] P. Damaske and V. Mellert, "A procedure for generating directionally accurate sound images in the upper-half space using two loudspeakers," *Acustica*, vol. 22, pp. 154-162, 1969.

[46] W.G. Gardner, "Transaural 3D audio," MIT Media Laboratory Tech. Report No. 342, January/February 1995.

[47] D.H. Cooper and J.L. Bauck, "Prospects for transaural recording *J. Aud. Eng. Soc.*, vol. 37, pp. 3-19, 1989.

[48] J. Bauck and D.H. Cooper, "Generalized transaural stereo and applications," *J. Aud. Eng. Soc.*, vol. 44, pp. 683-705, 1996.

[49] C. Kyriakakis, T. Holman, J.S. Lim, H. Hong, and H. Neven, "Signal processing, acoustics, and psychoacoustics for high-quality desktop audio," *J. Vis. Comm. and Im. Rep.*, vol. 9, pp. 51-61, 1997.

[50] C. Kyriakakis, "Fundamental and technological limitations of immersive audio systems," *IEEE Proceedings*, vol. 86, pp. 941-951, 1998.

[51] C. Kyriakakis, T. Holman, H. Neven, and C.V.D. Malsburg, "Immersive audio for desktop systems," presented at ICA/ASA '98, Seattle, Washington, 1998.

[52] W.G. Gardner, "Head-tracked 3D audio using loudspeakers," presented at WASPAA '97, New Palz, New York, 1997.

[53] A.V. Oppenheim and R.W. Shafer, *Discrete Time Signal Processing*: Prentice Hall, 1989.

[54] J.M. Jot, V. Larcher, and O. Warufsel, "Digital signal processing issues in the context of Binaural and transaural stereophony," presented at 98th Convention of the Audio Engineering Society, Paris, 1995.

[55] A. Mouchtaris, J.S. Lim, T. Holman, and C. Kyriakakis, "Signal processing considerations for immersive audio rendering," presented at 10th Tyrrhenian International Workshop on Digital Communications: Multimedia Communications, Ischia, Italy, 1998.

[56] S. Haykin, *Adaptive Filter Theory, 3rd Edition*: Prentice Hall, 1996.

[57] C. Kyriakakis and T. Holman, "Video-based head tracking for improvements in multichannel loudspeaker audio," presented at 105th Convention of the Audio Engineering Society, San Francisco, California, 1998.

[58] J. Sandvad and D. Hammershoi, "Binaural auralization: Comparison of FIR and IIR filter representation of HIR's," presented at 96th Audio Engineering Society Convention, Amsterdam, The Netherlands, 1994.

[59] B. Beliczynski, I. Kale, and G.D. Cain, "Approximation of FIR by IIR digital filters: An algorithm based on balanced model reduction," *IEEE Trans. Sig. Proc.*, vol. 40, pp. 532-542, 1992.

[60] J. Mackenzie, J. Huopaniemi, V. Välimäki, and I. Kale, "Low-order modeling of head-related transfer functions using balanced model truncation," *IEEE Sig. Proc. Letters*, vol. 4, pp. 39-41, 1997.

[61] T. Holman, "Monitoring sound in the one-person environment," *SMPTE Journal*, vol. 106, pp. 673-678, 1997.

[62] F.E. Toole, "Loudspeaker measurements and their relationship to listener preferences," *J. Aud. Eng. Soc.*, vol. 34, pp. 323- 348, 1986.

[63] S.E. Olive and F.E. Toole, "The detection of reflections in typical rooms," *J. Aud. Eng. Soc.*, vol. 37, pp. 539-53, 1989.

[64] S. Bech, "Perception of timbre of reproduced sound in small rooms: influence of room and loudspeaker position," *J. Aud. Eng. Soc.*, vol. 42, pp. 999-1007, 1994.

[65] R. Walker, "Early reflections in studio control rooms: The results from the first controlled image design installations," presented at 96th Meeting of the Audio Engineering Society, Amsterdam, 1994.

[66] T. Holman and F. Kampmann, "Loudness compensation: Use and abuse," *J. Aud. Eng. Soc.*, vol. 26, pp. 526-536, 1978.

[67] F.E. Toole, "Subjective measurements of loudspeaker sound quality and listener performance," *J. Aud. Eng. Soc.*, vol. 33, pp. 2-32, 1985.

[68] F.E. Toole and S.E. Olive, "The modification of timbre by resonances: perception and measurement," *J. Aud. Eng. Soc.*, vol. 36, pp. 122-42, 1988.

[69] L.D. Fielder and E.A. Cohen, "Determining noise criteria for recording environments," *J. Aud. Eng. Soc.*, vol. 40, pp. 384, 1992.