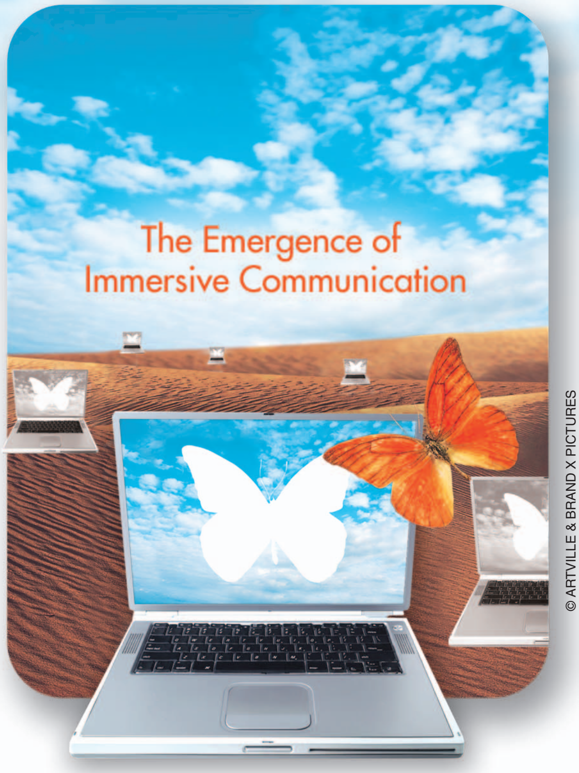
[Yiteng (Arden) Huang, Jingdong Chen, and Jacob Benesty]

Immersive Audio Schemes

[The evolution of multiparty teleconferencing]



© ARTVILLE & BRAND X PICTURES

After more than a century of accelerated advances in telecommunication technologies, people are no longer satisfied with talking to someone over a long distance and in real time. They want to collaborate through communication in a more productive way with the feeling of being together and sharing the same environment, which we refer to as an immersive experience. This need offers great opportunities for multichannel acoustic and speech signal processing, and for new ideas of voice communication services infrastructure. In this article, we present a survey of the development of various immersive audio schemes in concert with this movement according to our involvement and insights.

INTRODUCTION

Human communication has a long and rich history. But the primary goal is fairly constant: to wit, cut across spatial, temporal, and physical boundaries, and help people share information with the sense of experiential outreach. Modern communication technologies so far have made great strides in transcending the boundaries of space and time. But the issue of physical boundaries is relatively less addressed, let alone the sensory experience of "being there" when communicating. In fact, few users of today's communication systems would say that the interaction is satisfactorily natural and that they have the feeling of being in the same room and the feeling of sharing a common environment. It is this lack of immersive experience that makes people feel unfulfilled in remote information sharing and collaboration. Such an immersive experience is yet to become a reality supported by modern communication technologies and may be considered as the "last-mile" problem of telecommunications.

A person's sense of immersion is formed by his or her sensory response to the auditory and visual stimuli that exist in the ambiance of their environment. In immersive communication, thus, both audio and visual exchanges are indispensable, and probably other sensory modalities as well. The first demonstration of the so-called videoconferencing was a close-circuit test conducted by Bell Laboratories in 1927 [1]. It allowed Herbert Hoover (then the United States Commerce Secretary and later the United States President) to address an audience in New York City from Washington, D.C. The audio portion was two-way, but the video portion was one-way with only those in New York being able to see Hoover. Public interest in videoconferencing began with the display of AT&T's trademarked "Picturephone" product and service at the New York World's Fair in 1964 [2]. But Picturephone did not achieve the expected commercial success. The failure of the service is commonly attributed to its cost. It was too expensive due to the high cost of bandwidth and cameras then. Today the cost of

Digital Object Identifier 10.1109/MSP.2010.938754
Date of publication: 17 December 2010

1053-5888/11/$26.00©2011IEEE

AN IMMERSIVE AUDIO INTERFACE THAT FACILITATES BINAURAL HEARING NEEDS TO REPLICATE FOUR ATTRIBUTES OF FACE-TO-FACE COMMUNICATION.

videoconferencing has dropped to a commercially viable level, and videoconferencing has become a daily activity in businesses as well as in individual residences. Furthermore, because of increasing concerns with public safety (e.g., fear of terrorist attacks), growing travel expenses, and unanticipated delays (e.g., those caused by the volcanic ash shutdown of European air space in April 2010), there is a clear trend to reduce business travel and increase work from home. All these commercial changes and social trends add new impetus to the development of more advanced videoconferencing systems. This is evidenced by recent efforts on telepresence, which can be a high-definition videoconference and arguably facilitates eye contact, gaze awareness, and gesture recognition.

Adding video to teleconferencing gives three advantages: 1) fast interaction response and smooth conversational shifts from one talker to another can be elicited, 2) subtle emotion and opinion that we convey through body language can be faithfully shared with other participants, and 3) we can easily monitor the involvement of the others. They all help convey subconscious information in a teleconference. Subconscious cues were found more effective than conscious cues in communications of feelings and attitudes [3]. However, voice is by far the dominant media in the exchange of conference content. In fact, a teleconferencing session can still go on when the video link is broken, but it has to stop if the audio link is disrupted. So in addition to the pursuit of multimodal capabilities, we should never forget the importance of speech quality (including intelligibility and naturalness) and intermodal synergy. Moreover, there are great potentials to improve these two factors in an immersive teleconference with multiple parties being involved since binaural hearing is now allowed and can be fully exploited. This is an imperative step towards immersive communication. With both ears being kept busy, our auditory system can more easily extract a single talker's speech among multiple conversations and background noise, and can more seamlessly work together with the visual system in an adverse acoustic environment for speech perception (e.g., lip-reading).

Therefore, this article aims at audio processing and interface techniques that are necessary to support the goal of immersive communication. An immersive audio interface that facilitates binaural hearing needs to replicate four attributes of face-to-face communication [4]: 1) full-duplex exchange; 2) freedom of movement without body-worn or tethered microphones (i.e., hands free in the broad sense); 3) high-quality speech signals captured from a distance; and 4) spatial realism of sound rendering. These requirements imply that multiple microphones and loudspeakers would be used and the entire voice communication infrastructure might need to be renovated. This move incurs great challenges on multichannel acoustic and speech signal processing. While efforts from both the academic and industrial communities have been devoted to solving most of these problems and significant progress has been made over the last two decades, many fundamental challenges are still waiting for breakthroughs.

This article presents a systematic overview of the major challenges that have to be dealt with in immersive audio processing and interface. These include sound acquisition and processing, multiparty immersive audio mixing and management, and sound rendering for untethered immersive perception. The state-of-the-art technologies to solve these problems are briefly reviewed and several successful real-time systems are discussed to illustrate the advances and progress that have been made in immersive audio processing and interface.

SOUND ACQUISITION AND PROCESSING FOR IMMERSIVE ACOUSTIC RECONSTRUCTION

SIGNAL CHARACTERISTICS AND NOISE CLASSIFICATION IN IMMERSIVE VOICE COMMUNICATION

Acoustic waves are simply pressure disturbances propagating in the air. They carry information of the sound source and their energy is radiated spherically from the origin, i.e., the location of the sound source. The governing law of physics in this radiation process is the natural fall-off of the signal level inversely proportional to the distance from the origin, which is known as the inverse distance law. As a rule of thumb, the sound level decreases by 6 dB for each doubling of the distance from the source. This phenomenon makes distant acquisition of a speech signal vulnerable to interference from other concurrent sound sources and ambient noise. Moreover, in an enclosure, acoustic waves are often reflected many times by the boundaries before they reach a microphone, leading to distortion observed in the microphone signal. Therefore, acquisition of desired signals with high quality is far more difficult and challenging for immersive communications than in the classical telephony environment where the microphone is close to the user.

In immersive communications, it is more likely that multiple parties will be involved and conferencing is a more common mode of operation than point-to-point calling. In conferencing, one may hear the noise from every other participant and therefore the level of the perceived noise can grow with the number of participants. When the number is large and if noise is not well controlled, the perceived noise can reach a level such that speech is overwhelmed. So noise becomes a more quality-threatening problem for immersive voice communication.

Noise is a general term used to signify any unwanted signal that interferes with measurement, processing, and communication of the desired speech signal. This definition is, however, too broad, as it masks many important technical aspects of the real problem. In immersive communication, it is advantageous to break the general definition of noise into the following four categories [5]: additive noise, echo, reverberation, and competing speech. Due to different characteristics and (more importantly) the availability of a reference signal, the four types of noise need to be differently processed. In echo cancellation, the source (loudspeaker) signals are known. So echo control is theoretically a well-posed problem, and its practical applications have been relatively more successful than the control of the other three types of noise, in which blind or semiblind methods have to be incorporated. Therefore, in the following, the technologies for echo control are first discussed.

IN IMMERSIVE COMMUNICATION, IT IS ADVANTAGEOUS TO BREAK THE GENERAL DEFINITION OF NOISE INTO THE FOLLOWING FOUR CATEGORIES: ADDITIVE NOISE, ECHO, REVERBERATION, AND COMPETING SPEECH.

ACOUSTIC ECHO CONTROL

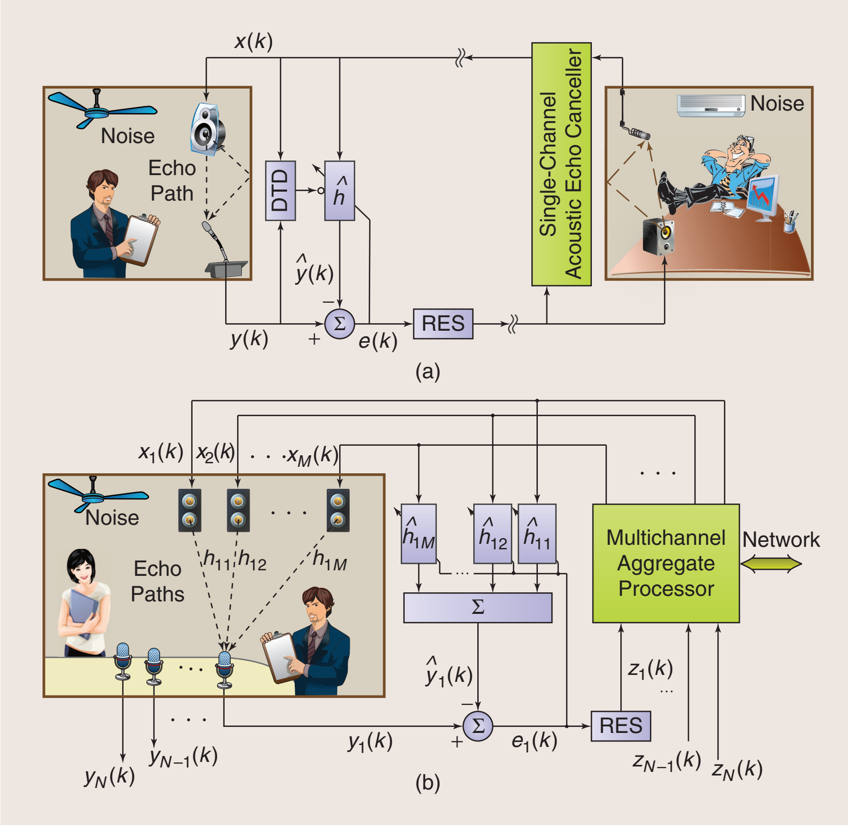
Acoustic echo in a hands-free voice communication system is produced by the acoustic coupling between the loudspeaker(s) and the microphone(s). The perception of an echo depends on not only its level but also its delay. Through long-distance transmission, the echo features a long delay time and would significantly reduce the quality of voice communication. When the delay approaches a quarter of a second, most people find it difficult to carry on a normal conversation. Full-duplex voice telecommunication was implausible, if not impossible, before the echo cancellation theory was developed by Bell Labs researchers in the 1960s [6].

For an immersive audio system with several microphones and loudspeakers, multiple echo paths need to be identified. Regardless of how many microphones there are, acoustic echo cancellation is always carried out individually with respect to each of them. But the number of loudspeakers present in the system draws a theoretical difference between monophonic (one loudspeaker) and multichannel (multiple loudspeakers) echo cancellations in the difficulty of tracking the echo paths.

MONOPHONIC ECHO CANCELLATION

As illustrated by Figure 1(a), an adaptive filter plays a central role in a monophonic echo cancellation system. It attempts to dynamically identify the acoustic echo path. As long as the channel impulse response of the echo path can be quickly and accurately determined, it is then straightforward to generate a good estimate of the echo and subtract it from the microphone signal. Since the loudspeaker signal as the reference is available, numerous nonblind adaptive filtering methods for system identification are applicable for solving this problem. The most widely known algorithms include the least mean square (LMS), normalized LMS (NLMS), affine projection (AP), recursive least square (RLS), proportionate NLMS (PNLMS), and frequency-domain and subband adaptive filters [7], [8]. Historically, the study of acoustic echo cancellation substantially enriched the adaptive filtering and system identification literature.

In the presence of doubletalk, i.e., when the far-end and near-end talkers are active at the same time, the near-end signal acts as a strong noise signal. This is likely to cause the adaptive filter to diverge, resulting in insufficient echo cancellation. To prevent this from happening, a doubletalk detector (DTD) is typically used and whenever doubletalk is detected, the adaptation is frozen until the end of the doubletalk [8].



[FIG1] **Illustration of (a) single-channel and (b) multichannel acoustic echo cancellation systems that reduce echoes arising from coupling between loudspeakers and microphones where DTD stands for doubletalk detector and RES for residual echo suppression.**

Acoustic impulse responses are usually long, and filter lengths of thousands of taps are not uncommon. Human ears have an extremely wide dynamic range and are very sensitive to weak tails of the channel impulse responses. But a practical real-time echo cancellation system cannot afford an equally long adaptive filter under complexity constraints. As a result, residual echoes usually still is audible. In addition, residual echoes can be produced by the nonlinear part of an echo path, which is unable to be characterized by an impulse response. Therefore an echo suppressor is usually applied to the residual echo after acoustic echo cancellation. The idea of echo suppression is similar to that of single-channel noise reduction, which will be explored in greater detail in a section below.

MULTICHANNEL ECHO CANCELLATION

When there are multiple loudspeakers, the echo problem becomes quite distinct from the monophonic case. As shown in Figure 1(b), the echo picked up by the nth ($n = 1, 2, \ldots, N$) microphone is due to M loudspeaker signals $x_m(k)$ ($m = 1, 2, \ldots, M$), where k is the discrete time index. Consequently, M channel impulse responses h_{nm} need to be jointly estimated for each microphone. While the M loudspeaker signals are different, they are obtained presumably from common sound sources and contain linearly related components. This leads to a singular signal covariance matrix in the normal equations, causing a nonuniqueness problem [9] that does not exist in the monophonic echo cancellation. So the loudspeaker signals have to be decorrelated first. Early straightforward ideas included signal dithering (i.e., adding Schroeder noise) and time-varying all-pass filtering, but were found unsatisfactory. An effective approach is to pass the loudspeaker signals through a memoryless nonlinearity [10]. The beauty of this is that the nonlinearity allows the adaptive channel identification algorithm to converge to a unique, true solution, but it is hardly perceptible for speech and complex musical signals. Furthermore, for a reasonable amount of nonlinearity, it does not distort the spatial information embedded in the multichannel signals. For a more detailed discussion of these techniques, please refer to [8] and the references therein.

MICROPHONE ARRAYS

To control noise, reverberation, and competing speech, multiple-microphone systems are generally more powerful than a single microphone [11]. Based on how the microphones are arranged, these systems have two basic forms: organized and distributed arrays. In an organized array, the sensors are arranged to form a particular geometry (such as a line, a circle, or a sphere) in which each sensor's position with reference to a common point is known. These sensors spatially sample the sound field and are required to have the same sensitivity. By applying filters to the outputs of the sensors and combining the results together, the desired source signals and their locations can be estimated, while the noise and interference can be reduced or even eliminated. This filtering process is called beamforming, which comprises a wide variety of array processing algorithms. A beamformer forms a response with different sensitivities to sounds arriving from different directions [12].

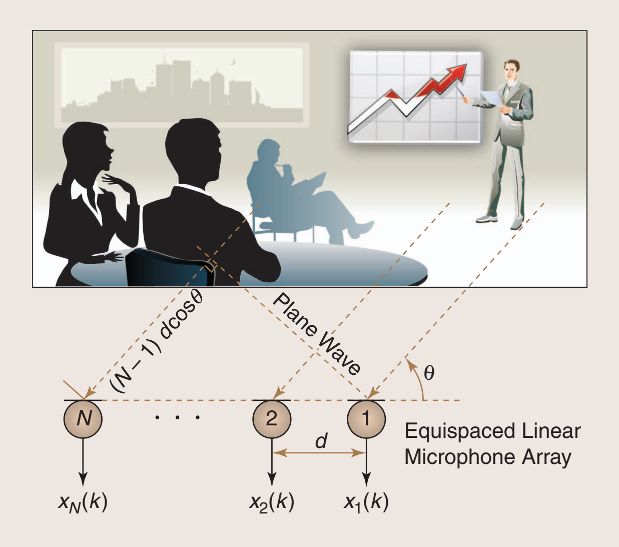
In comparison, a distributed array consists of randomly placed microphones. It offers the advantage of logistic convenience during installation and later operations. Typically, distributed arrays have a large number of elements forming a large sensor network. The microphone positions and the pattern of the array are usually not known, and a uniform response among the microphones cannot be presumed beforehand. So new ideas other than beamforming are needed to tackle the problem of speech acquisition with such a distributed microphone system.

BEAMFORMING FOR ORGANIZED ARRAYS

To show how an array can effectively be used in immersive communications, we first discuss beamforming for organized arrays and then address noise reduction, dereverberation, and source separation techniques that can be used for both organized and distributed arrays. Broadly, beamforming has three basic forms depending on the array geometry and how the algorithm is designed: additive, differential, and eigenbeamforming.

ADDITIVE BEAMFORMING

Additive beamforming, the oldest form of array signal processing, remains a powerful approach today. To illustrate the basic idea, we use a simple example with a uniformly spaced linear microphone array as shown in Figure 2. Assume that there is a single source in the far field such that its spherical wavefront appears planar at the array. Then the signal recorded by each microphone is simply a phase-shifted replica of the signal at the reference sensor, i.e., $x_n(k) = x_1(k - \tau_n)$, $n = 2, 3, \ldots, N$, where the time delay τ_n can be written as



[FIG2] Illustration of using an equispaced linear microphone array to capture a sound in the far field.

$$\tau_n = (n-1)d\cos\theta/c, \tag{1}$$

d is the spacing between two neighboring microphones, c is the speed of sound in air, and θ is the signal incident angle. Suppose that the array "looks" at the direction ϕ. Then we need to compensate each $x_n(k)$ by a delay equal to $\tau_0-(n-1)d\cos\phi/c$, where τ_0 is a constant processing delay. The time-shifted signals are averaged to produce the classical delay-and-sum (DS) output. The transfer function between the output and the signal can then be computed as

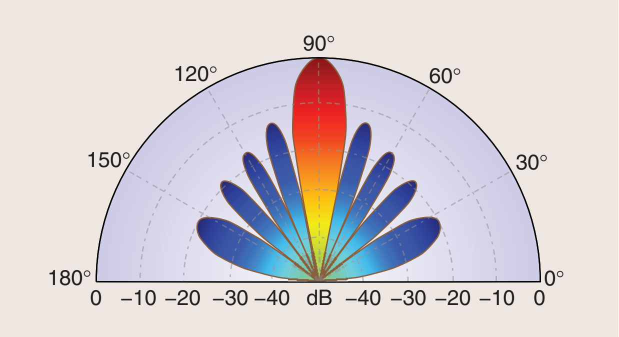
$$\mathcal{S}_{\mathrm{DS}}(\theta;f,\phi)=\frac{e^{-j\tau_0}}{N}\sum_{n=1}^{N}e^{-j2\pi(n-1)fd(\cos\theta-\cos\phi)/c}, \tag{2}$$

BEAMFORMING HAS THREE BASIC FORMS DEPENDING ON THE ARRAY GEOMETRY AND HOW THE ALGORITHM IS DESIGNED: ADDITIVE, DIFFERENTIAL, AND EIGENBEAMFORMING.

where $j=\sqrt{-1}$. The array gain pattern, which is often called the beampattern or directivity pattern, is defined as the magnitude of the transfer function and has the following form:

$$\mathcal{B}_{\mathrm{DS}}(\theta;f,\phi)=\left|\frac{\sin[N\pi fd(\cos\theta-\cos\phi)/c]}{N\sin[\pi fd(\cos\theta-\cos\phi)/c]}\right|. \tag{3}$$

Figure 3 plots the beampattern for the case with ten sensors, $d=8$ cm, $\theta=90°$, and $f=2$ kHz. It consists of a total of nine beams (in general, the number of beams in the range between 0° and 180° is equal to $N-1$). The one with the highest amplitude is called the mainlobe and all the others are called sidelobes. One important parameter regarding the mainlobe is the beamwidth (or mainlobe width), which is defined as the region between the first zero-crossings on either side of the mainlobe. For a DS beamformer using a linear array, the beamwidth is $2\cos^{-1}[c/(N\cdot f\cdot d)]$. This beampattern indicates that the DS beamformer allows the desired signal from the look direction (i.e., $\phi=\theta$) pass through without attenuation, while suppressing noise and other interfering signals coming from directions other than the look direction. The degree of suppression depends on the number of sensors, the microphone spacing, the angular separation between the desired signal and the signals to be suppressed, and the signal frequency. Though powerful, this simple beamformer suffers from a prominent drawback: it is a narrowband technique and would not yield the same beampattern at different frequencies for broadband signals like speech. If the speech source moves away from the look direction, it will be low-pass filtered. In addition, noise is not uniformly attenuated over its entire spectrum, resulting in some disturbing artifacts in the array output. This is why broadband beamforming techniques have been developed for voice communication. A common way to design a broadband beamformer is to perform subband decomposition and then design narrowband beamformers independently for each subband. This is equivalent to applying a spatiotemporal filter to the array outputs, which is widely known as the filter-and-sum (FS) structure [13]. The core problem of broadband beamforming then becomes one of determining the coefficients of the spatiotemporal filter. These coefficients can be determined using many different criteria. So-called fixed beamformers are designed independently of the acoustic environment and array data. Alternatively, adaptive beamformers are estimated according to the received data. Examples include the minimum variance distortionless filter (MVDR) and linearly constrained minimum variance (LCMV) algorithms. These can be more efficient than fixed beamformers in suppressing reverberation and competing sources. However, adaptive algorithms may suffer from the signal cancellation problem, which deserves careful attention [12].



[FIG3] Beampattern of the traditional delay-and-sum beamformer with respect to far-field sound sources for a ten-element equispaced linear microphone array with $\phi=90°$, $d=8$ cm, and $f=2$ kHz.

DIFFERENTIAL BEAMFORMING (DIFFERENTIAL ARRAYS)

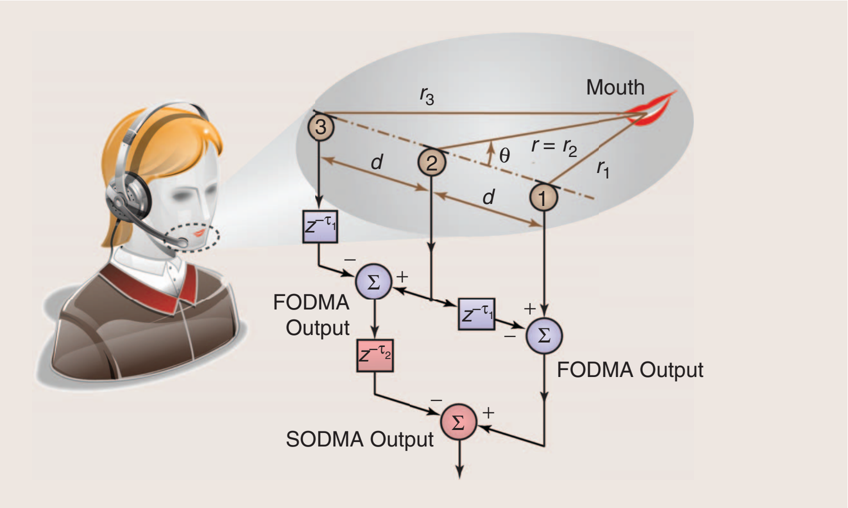
Differential beamforming is an inherent part of a differential microphone array, in which the microphones are placed much closer than in an additive array so the array is responsive to the spatial derivatives of the acoustic pressure field. An underlying assumption in the construction of differential arrays is that the true sound pressure differentials can be approximated by finite differences. Figure 4 illustrates how first-order and second-order differential microphone arrays are constructed. A general nth-order array has a response proportional to a linear combination of signals derived from spatial derivatives up to, and including order n. In a scalar pressure field, the (zero-order) responses of the three omnidirectional microphones in Figure 4 due to a sound source at a distance of r as a function of frequency f are

$$H_0(r_i;f)=\frac{e^{-j2\pi fr_i/c}}{r_i},\quad i=1,2,3. \tag{4}$$

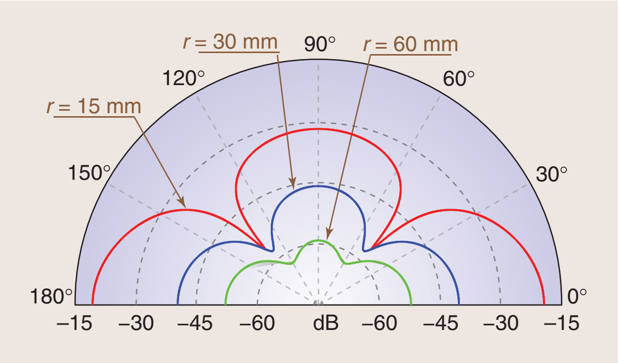
Then the response of the second-order differential microphone array (SODMA) can be written as

$$H_{\mathrm{SODMA}}(r,\theta;f) = [H_0(r_1;f)-H_0(r_2;f)] - [H_0(r_2;f)-H_0(r_3;f)], \quad (5)$$

where θ is the incident angle of the sound source with respect to the sensor axis. Figure 5 plots the SODMA beampattern to an on-axis sound source ($\theta = 0°$) evaluated at $r = 15$ mm, 30 mm, and 60 mm for $d = 10$ mm. Similar to additive beamforming, the differential array forms a response with different sensitivities at different directions (the shape of the directivity pattern depends on the delays τ_i and the order of array). As a matter of fact, it has been shown that for a given number of microphone sensors in an array, differential arrays have the potential to attain maximum directional gain [15]. It is also seen from Figure 5 that the array gain decreases with the distance r. When the sound source moves from 15 mm to 60 mm away from the array, the array gain decrease more than 30 dB. This indicates that a differential array inherently suppresses far-field noise and interference.



[FIG4] **Schematic diagram of a second-order differential microphone array.**



[FIG5] **Illustration of the second-order differential microphone array's sensitivity to the incident angle of the near-field sound source of interest with *d* = 1 cm and *f* = 2 kHz.**

Differential microphones are very compact in size and are found very useful in situations where the background noise level is very high. However, they also have many prominent drawbacks. First, the response of an nth order array has a high-pass filter nature with a slope of $6n$ dB/octave, so its frequency response has to be compensated to process wideband signals like speech. Second, the frequency response and level of differential arrays are extremely sensitive to the position and orientation of the arrays relative to the sound source, which makes it necessary to perform frequency and level equalization to its response according to the range and incident angle of the sound source [16]. Though feasible, this equalization process is in general very difficult for arrays with order higher than two. Finally, the mainlobe cannot be electrically steered to the desired source.

EIGENBEAMFORMING

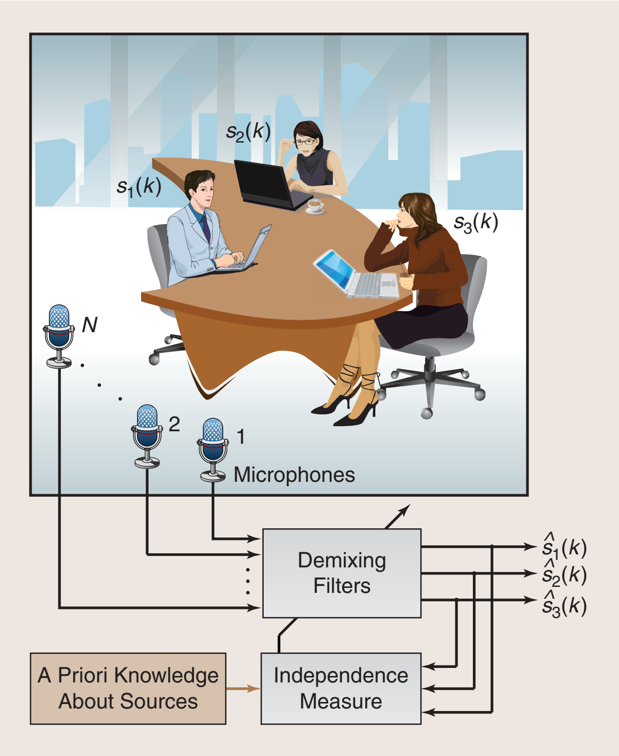
Eigenbeamforming consists of two steps: decomposition and synthesis [14]. The decomposition step transforms the multichannel sensor signals into an orthonormal space. Each base vector is called an eigenbeam. It is so named based on the analogy of an eigenvector to a matrix. The synthesis step forms the desired array response by weighting the eigenbeams and summing the results together. Theoretically, any directivity pattern that an eigenbeamformer can attain can also be realized using the additive beamforming method; but eigenbeamforming offers some unique advantages from an engineering viewpoint [15]: 1) it requires fewer signals to be stored since the number of eigenbeams can be much less than the number of sensors; and 2) it can form a desired response in a computationally very efficient way. However, to use eigenbeamforming, the array has to be carefully designed so that the sensor positions can meet the orthonomality condition. This is a nontrivial job. Moreover, the number of sensors with this beamformer is usually large to guarantee the required spatial resolution. That is why this technique is currently used mainly for spherical arrays even though theoretically plausible also for other shapes such as cylindrical, oblate, and prolate. The magnitude response of the eigenbeams is dependent on the eigenbeam order and has a high-pass filtering nature. So frequency equalization is required, as in differential beamforming.

NOISE REDUCTION

Noise reduction techniques intend to mitigate the effect of additive noise. This noise can come from various sources and can profoundly affect the processing and perception of speech signals in voice communication. Noise reduction is typically formulated as an estimation problem where the optimal estimate of the clean speech is achieved by optimizing some criteria, such as the mean-squared error (MSE) between the clean speech and its estimate, the signal-to-noise ratio (SNR), the a posteriori probability of the clean speech given its noisy observations, etc.

The complexity of this problem depends on the number of accessible microphones. Most of today's communication terminals are equipped with only one microphone. Existing single-channel noise reduction techniques fall into one of the following three classes [17]: filtering, spectral restoration, and model-based methods. The basic principle underlying the filtering technique is to pass the noisy speech through a filter/transformation. Since speech and noise normally have very different characteristics, the filter/transformation can be designed to significantly attenuate the noise level while leaving the clean speech relatively unchanged. The Wiener filter [18] and subspace method [19] are the two most representative algorithms in this category. Comparatively, the spectral restoration technique treats the problem as one of spectral estimation, i.e., estimating the spectrum of the clean speech from that of the noise-corrupted speech. Many algorithms have been developed for this purpose, such as spectral subtraction [20], the minimum-MSE (MMSE) estimator [21], the maximum likelihood (ML) estimator, and the maximum a posteriori (MAP) estimator, to name a few. Similar to the spectral restoration technique, the model-based approaches also formulate noise reduction as a parameter estimation problem. The difference between the two is that, in the model-based methods, a mathematical model is used to represent human speech production and parameter estimation is carried out in the model space. The model space normally has a much lower dimensionality than the original signal space. Typical algorithms in this class are the linear-prediction-model-based Kalman filtering approaches [22] and hidden Markov model-based statistical methods [23].

ONE COMMON PROBLEM WITH THE SINGLE-CHANNEL TECHNIQUES IS THAT SPEECH DISTORTION IS INEVITABLE AND GENERALLY THE MORE THE NOISE REDUCTION ACHIEVED, THE MORE THE SPEECH IS DISTORTED.



[FIG6] An acoustic source separation system assumes independence among multiple sound sources and attempts to maximize the independence measure of the output signals by adjusting the demixing filters. The procedure can be either blind or semiblind depending on whether a priori knowledge about the sound sources are used. Examples of a priori knowledge include the source positions, speech models, and sparseness of speech in the time-frequency domain.

One common problem with the single-channel techniques is that speech distortion is inevitable and generally the more the noise reduction achieved, the more the speech is distorted [18]. One way to circumvent this dilemma is to use multiple microphones. For instance, if two microphones are allowed and it is possible to use one microphone to pick up the noisy signal and the other to measure the noise field, the second microphone signal can be used as a noise reference to reduce the noise in the first microphone by means of adaptive cancellation [24]. In this way, the desired speech is not modified since no filter is applied to the primary signal. More generally, microphone arrays with beamforming techniques described previously can be used to reduce noise. However, beamforming is formulated to estimate the source signal, which attempts to perform both noise reduction and speech dereverberation at the same time. It is therefore not optimal from the noise reduction perspective. To more efficiently use the array for reducing noise, multichannel noise reduction techniques were developed. They attempt to estimate the speech signal observed at the reference microphone (instead of the source signal) exploiting the redundancy among different microphones [5]. The great advantage of using multiple microphones over a single microphone is that noise reduction can be achieved without adding speech distortion.

There is an extremely rich literature on the subject of noise reduction. But due to space limitation, only a small number of references are cited here. For more comprehensive coverage of the related references, the interested reader can refer to [5] and [17].

SOURCE SEPARATION

Sound source separation was motivated by observing the remarkable human ability of focusing on one particular voice or sound amid a cacophony of distracting conversations and background noise, an interesting psychoacoustic phenomenon referred to as the cocktail party effect [25]. This is a common experience that our hearing system and brain can handle well. However, it is a very difficult problem for speech processing. Beamforming has long been studied for this problem, but over the last two decades more effort has been devoted to blind source separation (BSS) after the tool of independent component analysis (ICA) was introduced. As illustrated by Figure 6, the multiple sound sources are presumably independent and BSS by ICA processes multiple microphone observations (linear mixtures of the sound sources) with a de-mixing system. The demixing system is determined in a learning procedure after which the outputs become as independent as possible with respect to each other. Existing BSS methods differ in the way the dependence of the separated signals is defined. These include second order statistics (SOSs), higher (than second)-order cumulant-based statistics, and information-theory-based measures [4].

In voice communications, acoustic channels are not instantaneous but convolutive due to room reverberation, which makes the problem much more complicated to solve. A common way is to transform the time-domain convolutive mixtures into the frequency domain via the fast Fourier transform (FFT) such that the mixtures in each frequency band are instantaneous. ICA is then preformed with respect to instantaneous mixtures independently in each frequency bin. With ICA, independent source signals in instantaneous mixtures can at best be blindly separated up to a scale and a permutation. This limitation leads to the possibility that the recovered signal is not a consistent estimate of one of the source signals over all frequencies. This is known as the permutation inconsistency problem. In addition, ICA algorithms require that the number of sources is less than or equal to the number of microphones, which is not always guaranteed. Therefore, the incorporation of a priori knowledge about speech sources has been proposed, making them only semiblind solutions. Examples of the a priori knowledge include the speech articulation model and the sparseness of speech in the time-frequency domain [26].

SPEECH DEREVERBERATION

Reverberation adds warmth to sound, which is essential for music, and helps people better orient themselves in the listening environment. However, it leads to temporal and spectral smearing, which would distort both the envelope and fine structure of a speech signal. As a result, speech becomes difficult to understand in the presence of room reverberation, especially for hearing-impaired and elderly people, and for automatic speech recognition systems. This gives rise to a strong need for effective speech dereverberation algorithms. Since neither the source speech signal nor the acoustic channel impulse responses are known a priori, the procedure of speech dereverberation is blind and is hence very challenging. A great amount of research on this topic has been carried out in the last four decades. Existing algorithms fall roughly into three classes [27]: 1) speech model-based dereverberation, 2) separation of speech and reverberation via homomorphic transformation, and 3) speech dereverberation by channel inversion and equalization [28]. While significant progress has been made, the problem is far from solved. It is worth mentioning that a spectral subtraction-based approach may be able to remove some reverberation [29]. But currently, the most effective techniques of reverberation control for teleconferencing are arguably still those that use acoustic wave absorption materials to cover room surfaces.

SPEECH SOURCE LOCALIZATION AND SPATIAL SOUND REPRODUCTION

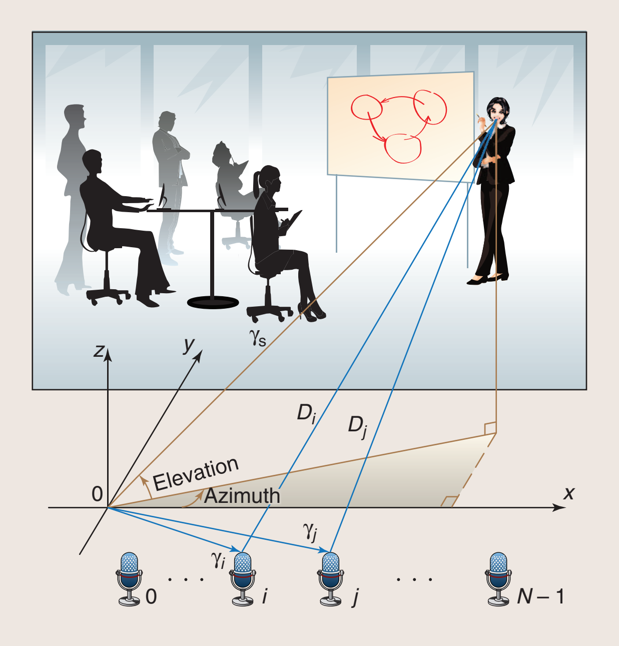
Creating a spatially rich acoustic environment is essential for immersive experiences. So while the capability of instantaneously localizing and continuously tracking speech sources is also a fundamental requirement of a number of distant speech acquisition techniques (e.g., beamforming) as discussed in the previous section, we devote a separate section here on speech source localization and spatial sound reproduction.

SPEECH SOURCE LOCALIZATION

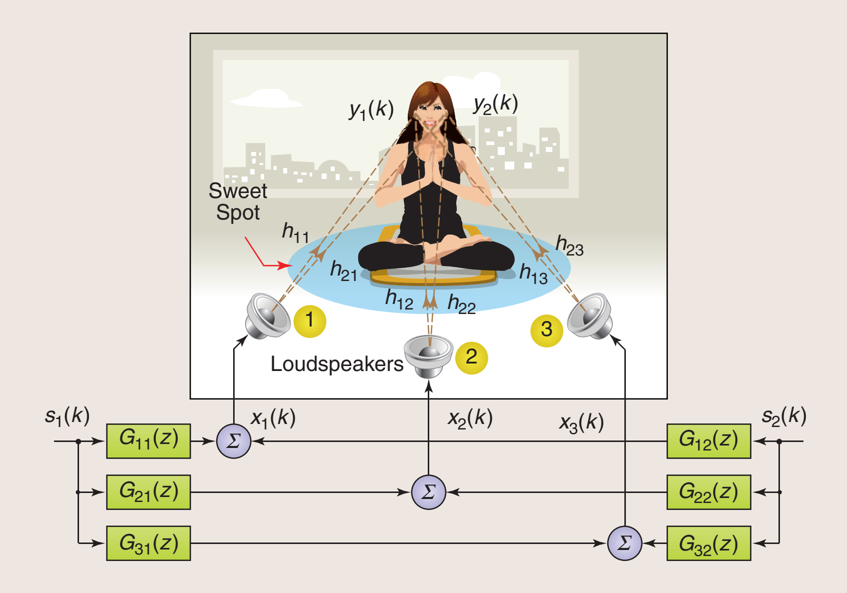
Locating a radiative point source using sensor arrays is more commonly referred to as direction of arrival (DOA) estimation in radar, underwater sonar, and seismology. Electrical beam scanning and high-resolution spatial-spectral analysis are two classes of the most celebrated approaches. The following four assumptions are reasonably realistic in these applications:

1) The source signal is narrowband.
2) It is stationary.
3) The source is in the far field.
4) The multipath effect (i.e., reverberation) is weak.

Unfortunately these assumptions do not hold generally for speech in common room acoustical environments. It was recently found [30] that time delay estimation (TDE)-based methods work better than the aforementioned two classes of DOA approaches with microphone arrays for speech sources as illustrated by Figure 7.



[FIG7] An acoustic sound source localization and tracking system is used to determine the position of active talkers in a teleconference.



[FIG8] Schematic diagram of a cross-talk cancellation system for stereophonic spatial sound reproduction that targets to create proper spatial cues at the two ears of a listener. Here three loudspeakers are used while more can be incorporated to increase the size of the sweet spot.

The generalized cross-correlation (GCC) algorithm [31] is the most widely known and used method for TDE. It is simple and can be extended to the case of multiple (say $M > 1$) sound sources by searching for the M largest peaks of the cross correlation function. But it does not cope well with room reverberation since an open-space signal model is used in its problem formulation. A more recent approach is based on the blind single-input multiple-output (SIMO) system identification techniques. The channel impulse responses from the single sound source to the multiple microphones are first estimated (which implies that a more realistic reverberant model is used) and then the relative time delays of arrival (TDOAs) are determined [32], [33]. This approach works better in conferencing environments where reverberations are not always well controlled. But it has some difficulties dealing with multiple simultaneous sound sources since blind identification of a multiple-input multiple-output (MIMO) system is a much more challenging and complicated problem. Alternatively TDE can only be carried out intermittently by taking advantage of the periods when only one speech source is active.

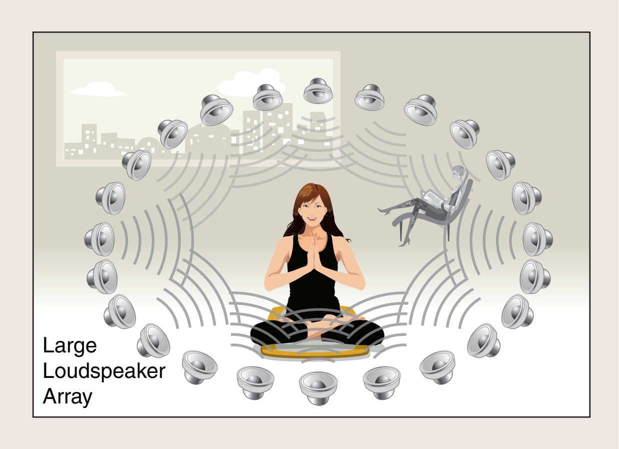
There are two ways to draw the connections between the estimated TDOAs and the source positions in the three-dimensional (3-D) space. The first is basically a search method where a grid is used to cover the space. For each grid point, its corresponding TDOAs are calculated, and they are compared to the estimated set of TDOAs. The computational complexity of this method grows with the density of the grid that is set by the required localization resolution. The methods of the second class may differ remarkably in the level of involved math skills, but the main idea is similar to that of triangulation. The difficulty of the problem lies in its nonlinear nature since a TDOA defines a hyperboloid [or hyperbola for two-dimensional (2-D)] instead of a flat plane (or a straight line for 2-D) between two microphones. For the methods of the first class, the two-step procedure can be simplified into one step by using a TDOA-based cost function rather than TDOAs explicitly. But their merits and limitations would not dramatically change.

For a human auditory system, the process of learning a new acoustic environment may be short thanks to its high efficiency, but it is nevertheless progressive. So intuitively, tracking is equally important in speech source localization (particularly when there are multiple speech sources). Particle filtering is a useful tool that has recently attracted a lot of research interest and deserves further exploration but is beyond the scope of this article.

SPATIAL SOUND REPRODUCTION

Spatial sound can be reproduced and presented to a listener or listeners using headphones or multiple loudspeakers. Headphones are obtrusive and cause noticeable ear discomfort after long use. Therefore they are not what we expect to use in immersive communications. Alternatively multiple loudspeakers ought to be employed; both stereophony [34] and wave field synthesis (WFS) [35] approaches may be considered. As shown in Figure 8, stereophony systems try to reproduce the pressure waves at the eardrums of the listener's left and right ears only. The underlying belief is that what matter to the brain in the perception of the spatial characteristics of incoming sound are only these two pressure waves. If they are reproduced, the listener would hear the sound exactly as in the original sound field. For such an open acoustic system, cross talk between the two ears must be canceled [36]. Unfortunately the best results can only be obtained in a fairly small "sweet spot." The size of the spot can be increased by using more loudspeakers. Two additional pitfalls should be noted: 1) typically head-related transfer functions (HRTFs) measured with a mannequin are used in determining the cross-talk cancellers, but there is always a mismatch in the size, shape, and acoustic properties between the mannequin and a particular listener, and 2) the dynamic cues that arise from the motion of the listener's head are missing, this leads to a feeling of unrealness, similar to a headphone system for which the perceived sound field does not move as the listener's head turns.

WFS uses a large number of loudspeakers (tens to hundreds) to reproduce a sound field not only at the two ears of a listener but in a larger space enclosing possibly multiple listeners [37], as illustrated by Figure 9. The principle relies on the so-called Kirchhoff-Helmholtz integral, which states



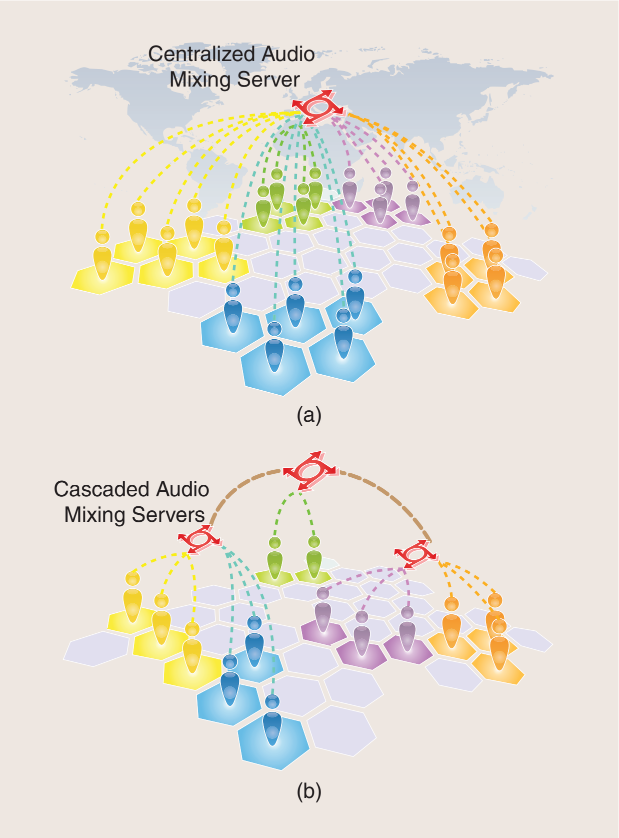
[FIG9] Illustration of wave field synthesis in which a large number of loudspeakers are used to reproduce the sound field around listeners.

that the distribution of acoustic pressure and particle velocity on a surface uniquely defines the sound field within the surface if no sources or obstacles are enclosed within it. While the spatial effect produced by WFS is more realistic and is appealing for immersive communications, the system can be very costly.

MULTIPARTY IMMERSIVE AUDIO MIXING AND MANAGEMENT

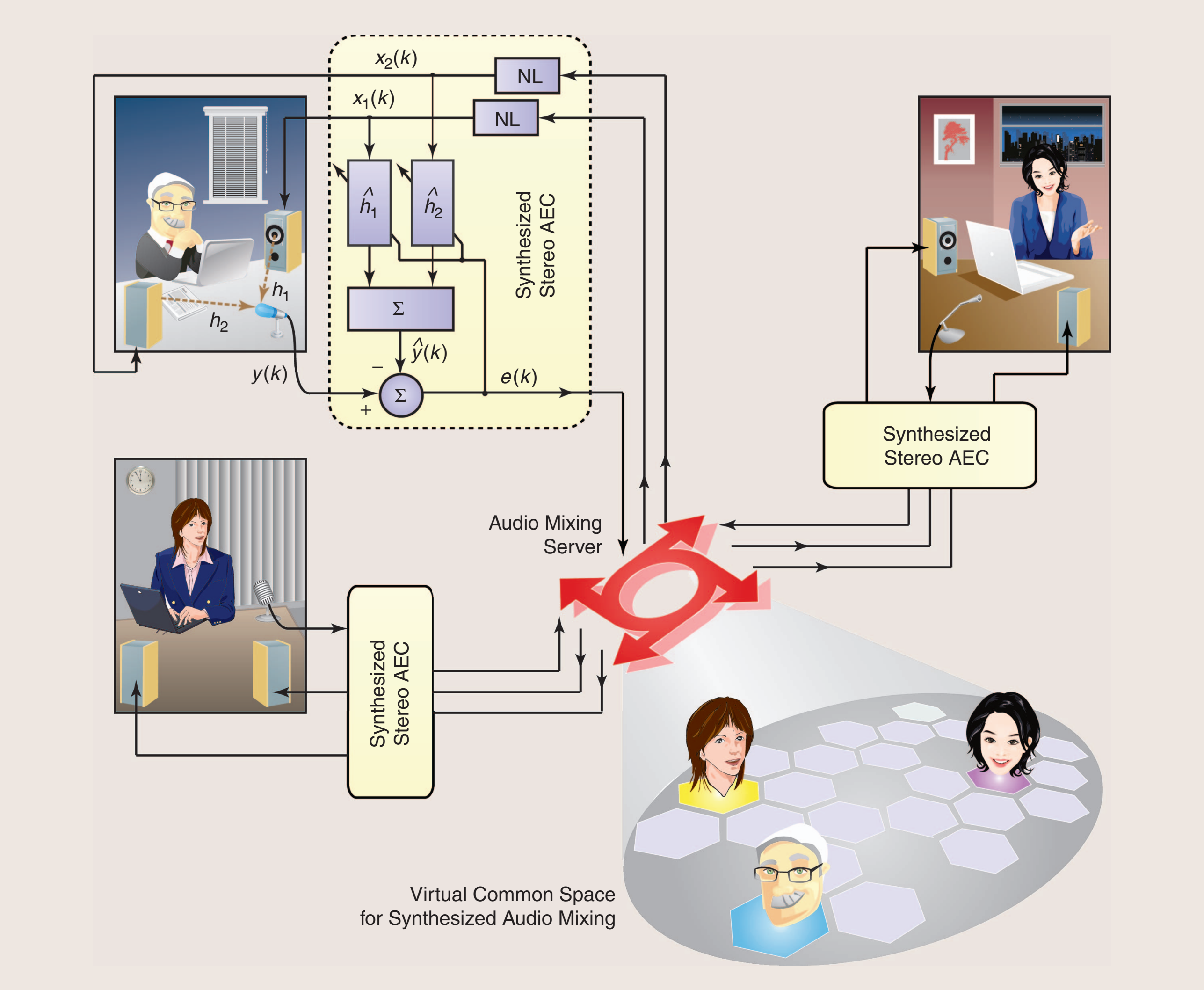
MIXING PREPROCESSING AND INTELLIGENT INPUT STREAM SELECTION CRITERIA

In immersive communication, the signals recorded in each conference site are sent to a mixer. The mixer processes all the inbound signals and presents a mixed signal to each conference room so that all the participants, regardless of their physical locations, can hear and interact with each other as if they are in the same acoustic environment. The mixing algorithms can vary greatly, depending on the sound system in each conference room. If an array of loudspeakers is used in each site, it is possible to render the voice from a different site into a spatially different location. In this way, the human binaural system can be fully exploited to discriminate concurrent speakers. But this would require transmission of multichannel signals from the mixer to the conference room as well as efficient sound rendering and management techniques. In situations where each conference room has only one loudspeaker, all the inbound signals need to be mixed to form a single output stream for each conference site. The most straightforward approach to mixing the signals for a particular conference room is to sum the inbound voice streams from all the other conference sites together and then normalize the aggregated signal to an appropriate range to prevent clipping. However, the voice quality of the mixed signal with such a simple method is often not guaranteed owing to many sophisticated reasons such as uneven voice levels, unbalanced voice qualities, and unequal SNR among different channels. In addition, when too many channels are mixed together and too many users speak simultaneously, the listener can hardly distinguish one particular speaker among the others. To limit the number of channels in the mixed signal at a time, the functionality of loudest N selection is added to the straightforward mixing algorithm. In this modified approach, the voice level of each inbound channel is estimated and is used as a selection criterion. Those channels with energy above a certain threshold are selected for mixing and all the others are discarded. This improved method can improve the perceptual quality of the mixed speech signal by limiting the number of mixed channels and has been widely adopted in the currently available conferencing systems. However, using signals' volume as the selection criterion may not be optimal since a higher volume does not necessarily mean that this channel is more important. Consequently, this method may block off some important speakers with low voice volume. In addition, due to the inherent fluctuation of the energy estimation, the presence of a certain channel in the mixed signal may not be continuous and consistent.



[FIG10] Two organizations of audio mixing network: (a) horizontal with one centralized mixing server and (b) hierarchical with cascaded mixing servers.

To control voice quality, we can divide the mixing algorithms into two steps: preprocessing, and mixing and management. In the first step, automatic gain control techniques can be applied to each channel so that all the channels can have similar volume and estimation techniques can also be



[FIG11] A synthesized stereo audio mixing system takes advantage of the normal audio accessories (i.e., one microphone and two loudspeakers) of current regular personal computers to deliver 2-D immersive experience for multiparty conferencing services. In the synthesized stereo acoustic echo canceller (AEC), NL stands for nonlinearity.

used to estimate the SNR, voice activity, and the signal quality of each channel. If noise is an issue, noise reduction techniques can be used to enhance the signal in each channel. Also, error concealment can be used if necessary to deal with the packet loss issue in voice over Internet Protocol (IP) channels. The mixing algorithm can then use the estimated parameters to determine which channels should be mixed. For example, if some channel have very low SNR, they should not be mixed. Driven by the estimated parameters and using the preprocessed signals, the mixer can perform an intelligent job so that the voice quality of the mixed signal can be optimized.

It should be noted that it is also desirable that the mixer could provide feedback to the preprocessor. For instance, if the preprocessor knows that the channels are not going to be mixed, noise reduction, error concealment, and perhaps other techniques are not needed for these channels. This leads to a computationally more efficient preprocessor.

NETWORK ARCHITECTURE

Similar to other network applications, voice communication networks fall within one of two network types: server based and peer-to-peer. Traditional and IP telephony services (e.g., Skype) usually adopt the peer-to-peer framework. For conferencing, only three-way calling is supported and audio mixing is carried out at one of the end points. As the number of conference participants increases, the peer-to-peer configuration has difficulty achieving the following desirable features: 1) unified and coherent auditory perception among all participants, 2) low bandwidth and hence minimum delay, 3) high scalability, 4) strong security, and 5) easy administration. Therefore the server-based configuration appears to be a better choice for immersive teleconferences. There are two possible organizations for the server-based framework: horizontal and hierarchical as illustrated by Figure 10. The former has a single centralized server while the latter encompasses a cascade of mixing servers. A hierarchical structure helps evenly distribute the intense processing loads among the mixing servers. When more participants join, more mixing servers can be added seamlessly to sustain a consistent quality-of-service. All these topics are open for research. Efforts from audio engineers, computer scientists, and network protocol developers will be required.

PRACTICAL STEPS TO FULLY IMMERSIVE COMMUNICATIONS

The nearly unprecedented success of the 3-D science fiction movie *Avatar* leads many to believe that immersive technologies will revolutionize filmmaking and the way they experience the cinema. We believe that the effect will not stop at the box offices but will impact other areas where we interact with the world. Three-dimensional TV and 3-D video games are possibly ready for prime time and what follows next could arguably be immersive communication. But before building fully immersive communication systems and services, we should take practical steps to develop and validate the core techniques. Presented in the following are two examples that effectively demonstrate the technologies of the synthesized stereo audio bridge and stereophonic echo cancellation.

SYNTHESIZED STEREO AUDIO BRIDGE FOR MULTIPARTY CONFERENCING

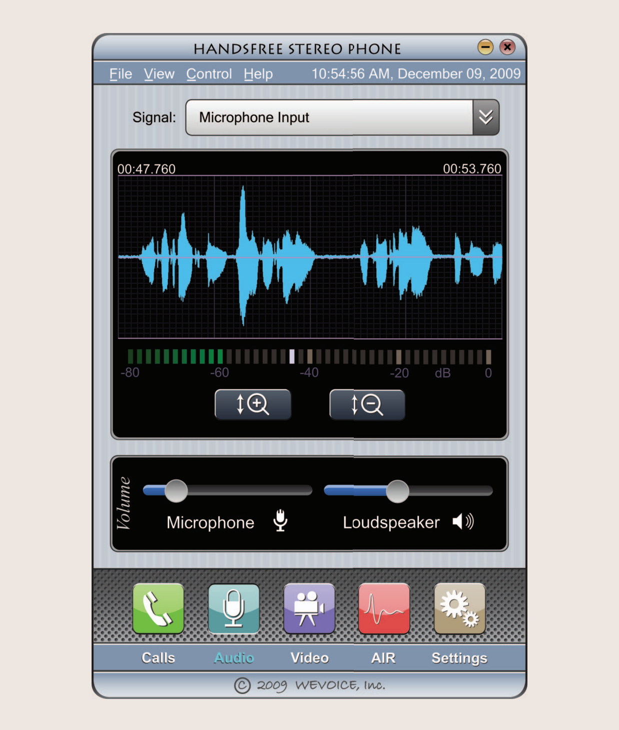
While an audio communication system that is equipped with multiple microphones and loudspeakers can certainly deliver good sound realism, it is prohibitively expensive. Today, all personal computers (desktops or laptops) have at least one microphone and a pair of loudspeakers. They can be readily employed for lifelike multiparty conferencing with the support of a synthesized stereo audio bridge [38], as visualized in Figure 11. This economical system has a common virtual 2-D space in which every conference participant takes a unique position. Accordingly, it presents speech signals from different sites to the listener with different spatial cues and impressions.

STEREOPHONIC ECHO CANCELLATION

It has been well known that going from a single-channel to a two-channel hands-free audio communication system produces a basic change in the requirements of an acoustic echo canceler (due to the aforementioned nonuniqueness problem). But extension further to more than two channels does not introduce any new problems other than higher complexity. Therefore, a real-time stereophonic echo canceller is deemed as a good demonstration system for multichannel acoustic echo cancellation. The world's first such system was developed in the late 1990s by the authors and their colleagues at Bell Labs. A recently redeveloped similar system shows better performance. It runs on Windows XP and has a more friendly user interface. A screen copy of the system is shown in Figure 12.

CONCLUSIONS

Today there stands before us a great opportunity to revolutionize communications with immersive technologies. The advance is partially attributed to the progress in multichannel acoustic and speech signal processing. In this article, we presented, according to our involvement and insights, the audio schemes that are behind this important evolution. The topics that have been explored include speech acquisition and processing, spatial information extraction and spatial sound reproduction, and mixing preprocessing and management. Practical steps towards fully immersive conferencing services were also discussed, and two real-time demonstration systems were introduced as success stories to explain the technologies and key system components.


HANDSFREE STEREO PHONE

Signal: Microphone Input

[FIG12] Screen copy of the recently developed stereo echo cancellation (including synthesized stereo echo cancellation) and audio conferencing system. Figure created by Yiteng Huang.

ACKNOWLEDGMENTS

We would like to thank Prof. Fred Juang for his inspiring discussions with us, and the three anonymous reviewers for their helpful criticisms and comments on an earlier draft, which have greatly improved the quality of this article.

AUTHORS

Yiteng (Arden) Huang (arden_huang@ieee.org) is the founder and CTO of WeVoice, Inc., in Bridgewater, New Jersey. He received his Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology. His research interests are in acoustic, speech, and audio signal processing, and multimedia communications. He is a coauthor and coeditor of six books and was an associate editor for *IEEE Signal Processing Letters* and *EURASIP Journal on Applied Signal Processing*. He is currently a member of the Audio and Acoustic Signal Processing Technical Committee and the Signal Processing Theory and Method Technical Committee of

the IEEE Signal Processing Society. He received the 2002 IEEE Signal Processing Society Young Author Best Paper Award and the 2008 Best Paper Award.

Jingdong Chen (jingdongchen@ieee.org) received his Ph.D. degree in pattern recognition and intelligence control from the Chinese Academy of Sciences. He has broad experiences and interests in acoustic signal processing, speech enhancement, and automatic speech recognition. He is currently a member of the Audio and Acoustic Signal Processing Technical Committee of the IEEE Signal Processing Society, an associate editor of *IEEE Transactions on Audio, Speech, and Language Processing,* and a member of the editorial board of *Open Signal Processing Journal*. He coauthored and coedited four books. He received the 2008 IEEE Signal Processing Society Best Paper Award, the 1998–1999 Research Grant Award from the Japan Key Technology Center, and the 1996–1998 President's Award from the Chinese Academy of Sciences. He is a Senior Member of the IEEE.

Jacob Benesty (benesty@emt.inrs.ca) is a faculty member at Université du Québec, INRS-EMT, Montreal, Canada. He holds a Ph.D. degree in control and signal processing from Orsay University, Paris, France. His research interests are in signal processing, acoustic signal processing, and multimedia communications. He coauthored and coedited 11 books, published 80 journal and 110 conference papers, and has held 14 U.S. patents. He was an associate editor of the *EURASIP Journal* on *Applied Signal Processing* and a member of the IEEE Signal Processing Society Audio and Electroacoustics Technical Committee. He cochaired two international and IEEE workshops. He received the 2001 and 2008 Best Paper Awards from the IEEE Signal Processing Society.

REFERENCES

[1] "Washington hails the test: Operator there puts through the calls as scientists watch," *The NY Times*, p. 20, Apr. 8, 1927. [Online]. Available: http://select.nytimes.com/gst/abstract.html?res=F20A17F73F5F147A93CAA9178FD85F438285F9&scp=4&sq=telephone+%22Herbert+Hoover%22&st=p

[2] "The evolution of PICTUREPHONE service," *Bell Lab. Rec.*, vol. 47, pp. 160–161, May/June 1969.

[3] A. Mehrabian, *Silent Messages*, 1st ed. Belmont, CA: Wadsworth, 1971.

[4] Y. Huang, J. Benesty, and J. Chen, *Acoustic MIMO Signal Processing*. Berlin: Springer-Verlag, 2006.

[5] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing*. Berlin: Springer-Verlag, 2009.

[6] M. M. Sondhi and A. J. Presti, "A self-adaptive echo canceler," *Bell Syst. Tech. J.*, vol. 45, pp. 1851–1854, 1966.

[7] M. M. Sondhi, "Adaptive echo cancellation for voice signals," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Berlin: Springer-Verlag, 2008, ch. 45, pt. H, pp. 903–927.

[8] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, and S. L. Gay, *Advances in Network and Acoustic Echo Cancellation*. Berlin: Springer-Verlag, 2001.

[9] M. M. Sondhi, D. R. Morgan, and J. L. Hall, "Stereophonic acoustic echo cancellation—An overview of the fundamental problem," *IEEE Signal Processing Lett.*, vol. 2, pp. 148–151, Aug. 1995.

[10] J. Benesty, D. R. Morgan, and M. M. Sondhi, "A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 156–165, Mar. 1998.

[11] J. L. Flanagan, J. D. Johnson, R. Zahn, and G. W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Amer.*, vol. 75, pp. 1508–1518, Nov. 1985.

[12] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.

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

[14] J. Meyer and G. W. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'02)*, 2002, pp. 1781–1784.

[15] G. W. Elko and J. Meyer, "Microphone arrays," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Berlin, Germany: Springer-Verlag, 2008, ch. 50, pt. I, pp. 1021–1041.

[16] H. Teutsch and G. W. Elko, "An adaptive close-talking microphone array," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001, pp. 163–166.

[17] J. Chen, J. Benesty, Y. Huang, and E. J. Diethorn, "Fundamentals of noise reduction," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Berlin: Springer-Verlag, 2008, ch. 43, pt. H, pp. 843–871.

[18] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Speech Audio Processing*, vol. 14, pp. 1218–1234, July 2006.

[19] Y. Ephraim and H. L. van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 251–266, July 1995.

[20] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 113–120, Apr. 1979.

[21] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, pp. 1109–1121, Dec. 1984.

[22] K. K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'87)*, 1987, pp. 177–180.

[23] Y. Ephraim, D. Malah, and B.-H. Juang, "On the application of hidden Markov models for enhancing noisy speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1846–1856, Dec. 1989.

[24] B. Widrow, J. R. Glover, J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Zeidler, E. Dong, and R. C. Goodwin, "Adaptive noise canceling: principles and applications," *Proc. IEEE*, vol. 63, pp. 1692–1716, Dec. 1975.

[25] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Amer.*, vol. 25, pp. 975–979, Sept. 1953.

[26] S. Makino, T.-W. Lee, and H. Sawada, Eds., *Blind Speech Separation*. Berlin: Springer-Verlag, 2007.

[27] Y. Huang, J. Benesty, and J. Chen, "Speech dereverberation," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Berlin: Springer-Verlag, 2008, ch. 46, pt. H, pp. 929–943.

[28] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech Signal Processing*, vol. ASSP-36, pp. 145–152, Feb. 1988.

[29] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acust.*, vol. 87, no. 3, pp. 359–366, 2001.

[30] M. S. Brandstein and H. F. Silverman, "A practical methodology for speech source localization with microphone arrays," *Comput., Speech, Language*, vol. 2, pp. 91–126, Nov. 1997.

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

[32] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Amer.*, vol. 107, pp. 384–391, Jan. 2000.

[33] Y. Huang and J. Benesty, "Adaptive multichannel time delay estimation based on blind system identification for acoustic source localization," in *Adaptive Signal Processing: Application to Real-World Problems*, J. Benesty and Y. Huang, Eds. Berlin: Springer-Verlag, 2003.

[34] B. B. Bauer, "Stereophonic earphones and binaural loudspeakers," *J. Audio Eng. Soc.*, vol. 9, pp. 148–151, 1961.

[35] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *J. Acoust. Soc. Amer.*, vol. 93, pp. 2764–2778, May 1993.

[36] B. S. Atal and M. R. Schroeder, "Apparent sound source translator," U.S. Patent 3 236 949, Feb. 1966.

[37] R. Rabenstein, S. Spors, and P. Steffen, "Wave field synthesis techniques for spatial sound reproduction," in *Topics in Acoustic Echo and Noise Control*, E. Hänsler and G. Schmidt, Eds. Berlin: Springer-Verlag, 2006, pp. 517–545.

[38] J. Benesty, D. R. Morgan, J. L. Hall, and M. M. Sondhi, "Synthesized stereo combined with acoustic echo cancellation for desktop conferencing," *Bell Labs. Tech. J.*, vol. 3, pp. 148–158, July–Sept. 1998.

[SP]