Introduction

At present, most teleconferencing systems use a single full-duplex audio channel for voice communication. These systems usually employ an acoustic echo canceler (AEC) to remove undesired echoes that result from the coupling between a loudspeaker and a microphone. As these systems evolve to an ever more lifelike and transparent audio/video medium, the need for enhanced sound realism becomes more important. This situation leads to the consideration of multichannel audio, which consists of at least two channels—that is, stereophonic sound. However, before full-duplex stereophonic teleconferencing can be deployed, the AEC problem must first be solved.

In this paper, we focus on one particular application: multiparticipant stereo desktop conferencing. With single-channel sound, simultaneous talkers are overlaid and it is difficult to concentrate on one particular voice. On the other hand, by using our binaural auditory system together with multichannel presentation, we can concentrate on one source to the exclusion of others (the so-called cocktail party effect). Moreover, localization helps us identify which person is actually talking. This is a very difficult task in a monophonic presentation. Communication using stereo (or multichannel) sound likely will grow rapidly in the near future, especially over the Internet.

The general scenario is as follows. Several persons in different locations would like to communicate with each other, and each one of them has a workstation. Each participant would like to see on the screen pictures of the other participants arranged in a reasonable fashion and to hear them in perceptual space in a way that facilitates identification and understanding. For example, the voice of a participant whose picture is located on the left of the screen should appear to come from the left.

We assume that we are located in a hands-free environment, where the composite acoustic signal is presented over loudspeakers. This study will be limited to two channels, so we assume that each workstation is equipped with two loudspeakers (one on each side of the screen) and one microphone (somewhere on
top of the screen, for example). As we will see later, a very convenient method using two loudspeakers can accommodate up to four participants. This arrangement can be generalized to create more images. However, it is not clear how many images a participant can conveniently deal with.

Obviously, such hands-free systems need multichannel AECs to reduce echoes that result from the coupling between loudspeakers and microphones in full-duplex communication.\(^1\) Formally, stereo (two-channel) acoustic echo cancellation can be viewed as a straightforward generalization of the single-channel acoustic echo cancellation principle.\(^1\) However, the stereophonic case gives rise to a nonuniqueness problem that does not arise in the single-channel case.\(^1,2\)

Figure 1 shows the configuration for a microphone at the local site, where \(h_1\) and \(h_2\) represent the two echo paths between the two loudspeakers and the microphone. The two reference signals, \(x_1\) and \(x_2\), from the remote sites are obtained by synthesizing stereo sound from the outputs of all the remote single microphones. The nonuniqueness arises because for each remote site, the signals are derived by filtering from a common source.

In the “Interchannel Differences for Synthesizing Stereo Sound” section, we recall the most useful interchannel differences for synthesizing stereo signals \(x_1\) and \(x_2\). In the “Stereo Acoustic Echo Cancellation” section, we discuss the problems of stereo acoustic echo cancellation and possible solutions, as well as how to combine interchannel differences and acoustic echo cancellation. The “Simulations” section presents some results from the simulations.

### Interchannel Differences for Synthesizing Stereo Sound

In the following scenario, we assume that two loudspeakers are positioned symmetrically on each side of the screen and that the conferee is located in front of the screen, close to and approximately centered between the loudspeakers. The location of auditory images in perceptual space is controlled by interchannel intensity and time differences, and it is mediated by the binaural auditory system.

In any discussion of the relationship between interchannel differences and perceptual effects, it is important to maintain a clear distinction between interchannel and interaural differences. If sounds are presented to the two ears by means of headphones, the interaural intensity and time differences \(\Delta I_a\) and \(\Delta t_a\) can be controlled directly. If signals are presented over a pair of loudspeakers, each ear receives both the left-channel and right-channel signals. The left-channel signal arrives earlier and is more intense at the left ear than it is at the right ear and vice versa, so that interchannel intensity and time differences \(\Delta I_c\) and \(\Delta t_c\) influence \(\Delta I_a\) and \(\Delta t_a\). In general, however, interaural intensity and time differences cannot be controlled directly. In addition to the perceptual effects produced by interaural time and intensity differences, localization of sounds emanating from a pair of loudspeakers is also influenced by the precedence effect.\(^3\) When identical or nearly identical sounds come to a listener from two loudspeakers, the sound appears to originate at the loudspeaker from which the sound arrives first.

To arrange the acoustic images, we can manipulate interchannel intensity and time differences, either separately or together. If two identical signals are presented to the two loudspeakers so that there are no interaural differences, the image will be well fused and localized in the median plane. As the interchannel intensity ratio varies from unity, the image will move toward the loudspeaker receiving the more intense signal. If, instead, the interchannel time difference is varied, the image generally will move toward the loudspeaker receiving the leading signal.\(^4,5\)

### Pure Interchannel Intensity Difference

It is well known that the effect of introducing an

---

**Panel 1. Abbreviations, Acronyms, and Terms**

- ADPCM—adaptive differential pulse code modulation
- AEC—acoustic echo canceler
- EIH—Ensemble Interval Histogram
- FAP—fast affine projection
- FRLS—fast recursive least-squares
- IS—Itakura-Saito
- LMS—least-mean-square
- MSE—mean square error
- NL—nonlinear transformation
- RLS—recursive least-squares
interchannel intensity ratio $\Delta I_c$ into signals that are otherwise identical is to move the image away from the median plane toward the loudspeaker receiving the more intense signal. Recent experiments conducted by coauthor Joseph L. Hall for a desktop configuration, as well as previous experiments with conventional loudspeaker placement in a room,\(^6\) indicate that a 20-dB interchannel intensity ratio produces almost complete lateralization.

If there are two remote conferees, experiments with headphones conducted in our laboratory suggest that interchannel intensity difference may be the best choice for desktop conferencing in terms of auditory localization and signal separation. The suggested strategy is to present the acoustic signal from one remote participant to one loudspeaker and the acoustic signal from the other remote participant to the other loudspeaker. With three remote participants, the suggested strategy would be the same for the first two participants, with the acoustic signal from the third remote participant presented equally to both loudspeakers. Thus, communication with good localization and signal separation among four conferees (one local plus three remote) appears to be feasible. The number of participants could be increased by using finer gradations of $\Delta I_c$, but separating the different remote talkers would be more difficult.

**Pure Interchannel Time Difference**

The nature of the signal plays a more important role in localization and signal separation for interchannel time difference $\Delta t_c$ than for interchannel intensity difference $\Delta I_c$. For pure tones, the binaural system is insensitive to interaural time differences $\Delta t_a$ for frequencies substantially above 1.5 kHz.\(^4\) For lower frequencies, localization of the image is periodic in the interaural time difference $\Delta t_a$, with a period equal to the period of the tone. For complex signals with some envelope structure, localization is influenced by both

Figure 1.
*Schematic diagram of stereophonic echo cancellation.*
low-frequency and high-frequency interaural time differences. Since, as discussed above, the interaural time difference $\Delta \tau_a$ is indirectly influenced by interchannel time difference $\Delta \tau_c$, it follows that the nature of the signal plays an important role in localization and signal separation for interchannel time difference $\Delta \tau_c$.

If there are two remote talkers, a suggested strategy for localization with interchannel time difference is to present the acoustic signal from one remote participant with an interchannel time difference $\Delta \tau_c = 1$ ms and the acoustic signal from the other remote participant with an interchannel time difference $\Delta \tau_c = -1$ ms. It has been known for a long time that, with headphone presentation, lateralization increases only slightly as interaural time difference increases above 1 ms.4,7 Recent experiments with desktop loudspeakers, as well as previous experiments with conventional loudspeaker placement in a room,6 show much the same effect. We do not know how interchannel time difference specifically affects the cocktail party effect.

**Combined Interchannel Intensity and Time Differences**

As discussed in the previous two subsections, both the interchannel intensity difference $\Delta I_c$ and the interchannel time difference $\Delta \tau_c$ can influence localization of a sound image. To a certain extent and within limits, these two types of interchannel differences are tradable in the sense that the same localization can be achieved with various combinations of the two variables. For example, one can achieve roughly the same image position with an amplitude shift, a time shift, or an appropriate combination of the two (time-intensity trading). Furthermore, under some conditions, intensity difference and time difference can be used to reinforce each other to provide a larger shift than is achievable by either one alone.

**Stereo Acoustic Echo Cancellation**

Integrating both synthesized stereo sound and stereo acoustic echo cancellation is not easy. The effectiveness of stereo acoustic echo cancellation will depend on the way the stereo sound is synthesized. Moreover, a problem of nonuniqueness is expected in the minimization problem because for any one remote participant, the two synthesized stereo signals $x_1$ and $x_2$ come from the same source.

**Choice of Interchannel Differences for Stereo Acoustic Echo Cancellation**

In principle, for localization with three remote talkers, the best choice of interchannel difference is $\Delta I_c$. But if we want to synthesize a remote talker on the right (respectively left), speech energy will be present only on the right (respectively left) loudspeaker, so we will be able to identify only one impulse response (from this loudspeaker to the microphone) and not the other one. From an acoustic echo cancellation point of view, this situation is highly undesirable. For example, if the remote talker on the right stops talking and the remote talker on the left begins, the adaptive algorithm will have to reconverge to the corresponding acoustic path because, in the meantime, it will have “forgotten” the other acoustic path. Therefore, the adaptive algorithm will have to track the different talkers continually, reconverging for each one, so the system will become uncontrollable—especially in a nonstationary environment (changes of the acoustic paths) and in double-talk situations. As a result, we will experience degraded echo cancellation much of the time.

The solution to this problem is that, for each remote talker, we must have some energy on both loudspeakers to be able to maintain identification of the two impulse responses between loudspeakers and the microphone. Thus, the optimal choice of interchannel difference from an acoustic echo cancellation point of view is pure $\Delta \tau_c$ because energy is equally presented to both loudspeakers for all remote talkers. However, in practice, this choice may not be sufficient for good localization. Therefore, combined $\Delta I_c / \Delta \tau_c$ seems to be the best compromise between good localization and echo cancellation.

If there are two remote talkers, a strategy for good localization and echo cancellation would be to present the acoustic signal from one remote participant to both loudspeakers with $\Delta I_c = 6$ dB, $\Delta \tau_c = 1$ ms and the acoustic signal from the other remote participant to both loudspeakers with $\Delta I_c = -6$ dB, $\Delta \tau_c = -1$ ms. With three remote participants, the suggested strategy would be the same for the first two participants with Bell Labs Technical Journal • July–September 1998 151
the addition of the third remote participant’s microphone signal presented to both loudspeakers with $\Delta I_c = 0$ dB, $\Delta T_c = 0$ ms.

Thus, for any remote participant’s microphone signal $s$, the contribution to the local synthesized stereo signals is written

$$x_i(n) = g_i(n) * s(n), \quad i = 1, 2,$$

where $*$ denotes convolution and $g_i$ are the impulse responses for realizing the desired $\Delta I_c/\Delta T_c$. For example, with the above suggested 6-dB and 1-ms values, a talker on the left, say, would be synthesized with

$$g_1 = [1 \ 0 \ \ldots \ 0 \ 0]^T,$$
$$g_2 = [0 \ 0 \ \ldots \ 0 \ 0.5]^T,$$

where the number of samples of delay in $g_2$ corresponds to 1 ms.

**Figure 2** shows how the signals from $N$ remote conferees are combined to produce the local synthesized signals $x_1$, $x_2$. Each $g_{j,1}$, $g_{j,2}$ pair is selected as exemplified above to locate the acoustic image in some desired position. There is some flexibility as to where the synthesis function is located, and the most efficient deployment will depend on the particular system architecture. These considerations, however, are beyond the scope of this paper.

**The Nonuniqueness Problem and the Proposed Solution**

Let

$$X_i(\eta) = \left[ x_i(\eta) \ x_i(\eta-1) \ \ldots \ x_i(\eta-L+1) \right]^T, \quad i = 1, 2,$$

be the two $L$-dimensional vectors formed from samples of the stereo channel input signals, and

$$W_i = \left[ w_{i,0} \ w_{i,1} \ \ldots \ w_{i,L-1} \right]^T, \quad i = 1, 2,$$

be the two model filters. Let

$$R = E\left\{ X_1(\eta) X_1^T(\eta) \ X_2(\eta) X_2^T(\eta) \right\},$$

be the covariance matrix of the input signals, where $E$ denotes mathematical expectation, and let the vector

$$r = E\left\{ X_1(\eta) y(\eta) \ X_2(\eta) y(\eta) \right\},$$

be the cross-correlation vector between the input signals and output (microphone) signal $y(n)$. Then, the minimization problem to obtain the model filters $W_1$ and $W_2$ leads to the classical Wiener-Hopf equation

$$R \begin{bmatrix} W_1 \\ W_2 \end{bmatrix} = r.$$  

We also have the relation

$$X_i^T(n) g_2 = X_2^T(n) g_1,$$

where $g_1$ and $g_2$ are the two impulse responses for synthesizing the stereo signals. This linear relation follows from (1), giving

$$x_1 * g_2 = s * g_1 * g_2 = x_2 * g_1.$$
on the impulse responses $g_1$ and $g_2$. This, of course, is intolerable because $g_1$ and $g_2$ can change instantaneously—for example, as one remote conferee stops talking and another begins.

The best way we know to alleviate the characteristic nonuniqueness of a stereo AEC is to first preprocess each input signal $x_i$ by the nonlinear transformation (see Figure 3)\(^2,9\)

\[ x'_i = x_i(n) + \alpha f[x_i(n)] , \tag{4} \]

where $f$ is a nonlinear function, such as a simple half-wave rectifier. A transformation like this reduces the interchannel coherence and hence, the condition number of the covariance matrix, thereby greatly reducing the misalignment.\(^2\) Because the two input signals $x_1$ and $x_2$ are almost (or can be exactly) the same, it is important to use two different nonlinear functions. For example, we can use a positive half wave for $x_1$ and a negative half wave for $x_2$. With a reasonably small value of $\alpha$, this distortion is hardly audible in typical listening situations and does not affect stereo perception. Thus, we include this kind of transformation in the stereo AEC.

Since convergence to the unique solution depends on the small nonlinear term, least-mean-square (LMS)-type gradient algorithms will be very slow. Therefore, we propose using a rapidly converging algorithm like the two-channel recursive least-squares (RLS) algorithm. The total number of operations is $28L$ multiplications and $28L$ additions per sample in a (computationally) fast, stabilized version of this algorithm.\(^8\) We can also use a two-channel affine projection algorithm.\(^10,11\) All these algorithms can be implemented in subbands for a real-time application.\(^12\)

In general, a distortion of the type in (4) could be expected to produce objectionable distortion. However, for speech signals, the distortion is barely perceptible for the following three reasons. First, the distorted signal $x'_i$ depends only on the instantaneous value of the original signal $x_i$ so that during periods of silence no distortion is added. Second, the periodicity remains unchanged. Third, for voiced sounds, the harmonic structure of the signal induces “self masking” of the harmonic distortion components.

A subjectively meaningful measure to compare $x_i$ and $x'_i$ is not easy to find. To be useful in speech processing, a mathematical measure of distance must have a high correlation between its numerical value and the subjective distance judgment, as evaluated on real speech signals.\(^13\) Because many psychoacoustic studies of perceived sound differences can be interpreted in terms of differences of spectral features, measurement of spectral distortion can be argued to be reasonable both mathematically and subjectively.

A very useful distortion measure is the Itakura-Saito (IS) measure, given as

\[ d_{IS} = \frac{a_i^T R a_i^*}{a_i^T R a_i} - 1 , \tag{5} \]

where $R_i$ is the Toeplitz autocorrelation matrix of the LPC model $a_i$ of a speech signal frame $x_i$, and $a_i^*$ is the LPC model of the corresponding distorted speech signal frame $x'_i$. Many experiments in speech recognition show that if the IS measure is less than about 0.1, the two spectra that we compare are perceptually nearly identical. Simulations show that even with a half-wave nonlinearity with $\alpha$ as high as 0.5, the IS metric is small (about 0.03).

We could also use the Ensemble Interval Histogram (EIH) distance, which is based on the EIH model.\(^14\) The interest in using this distance lies in its capability to mimic human judgement of quality. Indeed, EIH is a very good predictor of the mean opinion score but only if the two speech observations under comparison are similar enough,\(^14\) which is the case here. Then, this measure should be a good predictor of the speech signal degradation when nonlinear distortions are used. Simulations show that with a
nonlinearity (half-wave) \( \alpha = 0.5 \), the EIH distance is \( 1.8 \times 10^{-3} \), which is as good as a 32-kb/s adaptive differential pulse code modulation (ADPCM) coder.

**Figure 4** is a composite diagram of the synthesis, nonlinear transformation, and stereo AEC. It shows the complete local signal processing suite for one conferee. This same setup is reproduced for each conferee, making obvious permutations on both the remote signals and the choice of synthesis filters.

**Simulations**

By using simulation, we first show the convergence of the adaptive algorithm **without** and **with** a nonlinearity when a remote conferee talks for a while, then stops talking, and another remote conferee begins talking. To demonstrate this effect more clearly, we use two independent white noises as the two remote sources and we synthesize the stereo effect using \( \Delta I_c = 6 \text{ dB}, \Delta \tau_c = 1 \text{ ms for source 1 and } \Delta I_c = -6 \text{ dB}, \Delta \tau_c = -1 \text{ ms for source 2.} \)

The microphone output signal \( y(n) \) in the receiving room is obtained by summing the two convolutions \((h_1 \ast x_1)\) and \((h_2 \ast x_2)\), where \( h_1 \) and \( h_2 \) were measured in an actual room as 4,096-point responses. The length of the adaptive filters \( w_1 \) and \( w_2 \) is taken as \( L = 1,000 \). An additional white noise with a 40-dB signal-to-noise ratio is added to the microphone signal \( y(n) \). The sampling frequency is 8 kHz. For all our simulations, we used the two-channel fast recursive least-squares (FRLS) algorithm. Source 1 talks first for 3 s. After 3 s, source 1 stops talking and source 2 begins talking for 2.5 s.
Figure 5 shows the behavior of the mean square error (MSE) without a nonlinearity while Figure 6 shows the behavior of the MSE with a nonlinearity (positive and negative half wave, $\alpha = 0.5$). For the purpose of smoothing the curves, error samples are averaged over 128 points. We can see that the adaptive algorithm is very sensitive to changes in the remote talkers without a nonlinearity. When changes occur, the algorithm diverges and must reconverge again. However, with a nonlinearity, the adaptive algorithm is much more robust.

In the second part of these simulations, we show the effectiveness of our nonlinear transformation method on the coherence and misalignment with synthetic stereo sound using actual speech signals. A stereo signal was synthesized with $\Delta t_c = 0$ dB, $\Delta \tau_c = 0$ ms using a speech signal sampled at 8 kHz. Thus, the two signals $x_1$ and $x_2$ presented at the loudspeakers in the receiving room are identical (this is the most difficult example). Here, we use again the positive and negative half-wave rectifiers in Figure 3 for $x_1$ and $x_2$ with $\alpha = 0.5$. With this value, no audible degradation of the original signal occurs. Everything else is the same as in the previous example.

Figure 7 shows the behavior of the coherence magnitude and the misalignment (for the former, we use the MATLAB* spectrum function and take the square root of the computed magnitude-squared
Figure 6.
Behavior of the mean square error (MSE) with nonlinearity (positive and negative half-wave, $\alpha = 0.5$); other conditions same as in Figure 5.

Figure 7.
Behavior of the coherence magnitude (a) and misalignment (b) with synthetic stereo sound and $\alpha = 0.5$. 
coherence). The coherence magnitude samples are averaged over 100 points. We note that the coherence is reduced, implying a great improvement of the misalignment. Without the nonlinear transformation, the signals $x_1$ and $x_2$ are identical, so the coherence magnitude is equal to 1 at all frequencies and the misalignment is not much less than unity (0 dB).

**Conclusion**

In this paper, we have presented the most useful ways to synthesize stereo sound for multiparticipant desktop conferencing. We have also shown how this can be combined with stereo acoustic echo cancellation, explaining the main problems and practical solutions for this application.

**Acknowledgment**

We thank B. H. Juang of Bell Labs for encouraging us to write this paper.

*Trademark*

MATLAB is a registered trademark of The MathWorks, Inc.

**References**


(Manuscript approved August 1998)

**JACOB BENESTY** is a member of technical staff in the Acoustics and Speech Research Department at Bell Labs in Murray Hill, New Jersey. He has an M.S. degree in microwaves from the Pierre And Marie Curie University and a Ph.D. in control and signal processing from ORSAY University, both in Paris, France. Dr. Benesty is responsible for signal processing techniques applied to electroacoustic systems.
DENNIS R. MORGAN is a distinguished member of technical staff in the Acoustics and Speech Research Department at Bell Labs in Murray Hill, New Jersey. He has a B.S. degree from the University of Cincinnati in Ohio, as well as M.S. and Ph.D. degrees from Syracuse University in New York, all in electrical engineering. Dr. Morgan’s research interests are in adaptive signal processing techniques applied to electroacoustic systems. He has authored numerous journal publications and is coauthor of the book *Active Noise Control Systems: Algorithms and DSP Implementations*. Since 1995, Dr. Morgan has served as associate editor for IEEE Transactions on Speech and Audio Processing.

JOSEPH L. HALL is a distinguished member of technical staff in the Acoustics and Speech Research Department at Bell Labs in Murray Hill, New Jersey. He received a B.A. degree in physics from Williams College in Williamstown, Massachusetts. In addition, he has S.B., S.M., and Ph.D. degrees in electrical engineering, all from the Massachusetts Institute of Technology in Cambridge. His research interests are in the area of auditory psychophysics. A fellow of the Acoustical Society of America, Dr. Hall formerly served on the society’s executive council and was an associate editor of the Journal of the Acoustical Society of America.

M. MOHAN SONDHI is a distinguished member of technical staff in the Acoustics and Speech Research Department at Bell Labs in Murray Hill, New Jersey. He has a B.S. in physics from Delhi University in India, a D.I.I.Sc. in communications engineering from the Indian Institute of Science in Bangalore, as well as M.S. and Ph.D. degrees in electrical engineering from the University of Wisconsin in Madison. Dr. Sondhi’s research interests include speech signal processing, echo cancellation, and the modeling of auditory, speech, and visual processing by humans. He has authored or coauthored more than 95 technical papers, and he serves on the editorial boards of two technical journals. Dr. Sondhi is a former IEEE associate editor and guest scientist at three institutions. He is also a Bell Labs Fellow and a recipient of the IEEE Sumner Award.