SYNTHESIZED STEREO COMBINED WITH ACOUSTIC ECHO CANCELLATION FOR DESKTOP CONFERENCING

Jacob Benesty, Dennis R. Morgan, Joseph L. Hall, M. Mohan Sondhi

Bell Laboratories, Lucent Technologies
700 Mountain Avenue, Murray Hill, NJ 07974
Email: { jbenesty,drrm,jlh,mms }@bell-labs.com

ABSTRACT

One promising application in modern communications is desktop conferencing, which can involve several participants over a widely distributed area. Synthesized stereophonic sound will enable a listener to spatially separate one remote talker from another and thereby improve understanding. In such a scenario, we assume we are located in a hands-free environment where the composite acoustic signal is presented over loudspeakers, thus requiring acoustic echo cancellation. In this paper, we explain some of the methods that can be used to synthesize stereo sound and how such methods can be combined efficiently with stereo acoustic echo cancellation in the face of several difficult problems.

1. 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 echos 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 multi-channel 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 mono presentation. Communication with 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 suppose 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 accomodate 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 multi-channel AECs to reduce echos that result from coupling between loudspeakers and microphones in full-duplex communication [1]. Formally, stereo (two-channel) acoustic echo cancellation can be viewed as a simple generalization of the single-channel acoustic echo cancellation principle [1]. However, the stereophonic case gives rise to a non-uniqueness 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.

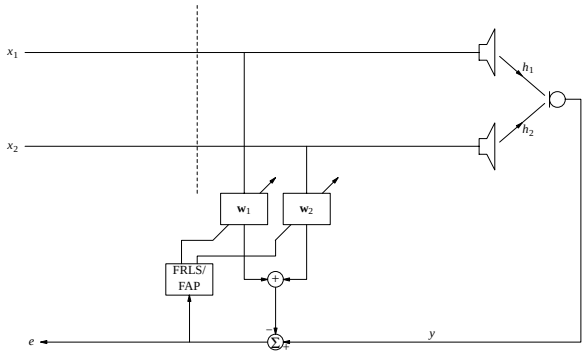

Figure 1: Schematic diagram of stereophonic echo cancellation.

2. 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 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 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 ΔI_{a} and $\Delta \tau_{\mathrm{a}}$ can be controlled directly. If signals are presented over a pair of loudspeakers, each ear receives both the left- and right-channel signals. The left-channel signal arrives earlier and is more intense at the left ear than at the right, and vice versa, so that interchannel intensity and time differences ΔI_{c} and $\Delta \tau_{\mathrm{c}}$ influence ΔI_{a} and $\Delta \tau_{\mathrm{a}}$, but in general interaural intensity and time differences cannot be controlled directly. In addition to perceptual effects produced by interaural time and intensity differences, localization of sounds presented over 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 will in general move toward the loudspeaker receiving the leading signal [4], [5].

2.1. Pure Interchannel Intensity Difference

It is well known that the effect of introducing an interchannel intensity ratio ΔI_{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 J. 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 ΔI_{c}, but separating the different remote talkers would be more difficult.

2.2. Pure Interchannel Time Difference

The nature of the signal plays a more important role in localization and signal separation for interchannel time difference $\Delta \tau_{\mathrm{c}}$ than for interchannel intensity difference ΔI_{c}. For pure tones, the binaural system is insensitive to interaural time differences $\Delta \tau_{\mathrm{a}}$ for frequencies substantially above 1.5 kHz [4], and for lower frequencies, localization of the image is periodic in the interaural time difference $\Delta \tau_{\mathrm{a}}$, with a period equal to the period of the tone. For complex signals with some envelope structure, localization is influenced by both low- and high-frequency interaural time differences. Since, as discussed above, the interaural time difference $\Delta \tau_{\mathrm{a}}$ is indirectly influenced by interchannel time difference $\Delta \tau_{\mathrm{c}}$, it follows that the nature of the signal plays an important role in localization and signal separation for interchannel time difference $\Delta \tau_{\mathrm{c}}$.

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 $\Delta \tau_{\mathrm{c}} = 1$ msec and the acoustic signal from the other remote participant with an interchannel time difference $\Delta \tau_{\mathrm{c}} = -1$ msec. It has been known for a long time [4] that, with headphone presentation, lateralization increases only slightly as interaural time difference increases above 1 msec. Recents 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.

2.3. Combined Interchannel Intensity and Time Differences

As discussed in the previous two subsections, the localization of a sound image can be influenced by both the interchannel intensity difference ΔI_{c} and the interchannel time difference $\Delta \tau_{\mathrm{c}}$. 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.

3. STEREO ACOUSTIC ECHO CANCELLATION

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

3.1. Choice of Interchannel Differences for Stereo AEC

In principle, for localization with three remote talkers, the best choice of interchannel difference is ΔI_{c}. But if we want to synthesize a remote talker on the right (resp. left), speech energy will be present only on the right (resp. 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 undesir-

able. 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 have 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 $\Delta\tau_c$ since energy is equally presented to both loudspeakers for all remote talkers. However, in practice, this choice may not be enough 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, $\Delta\tau_c = 1$ msec and the acoustic signal from the other remote participant to both loudspeakers with $\Delta I_c = -6$ dB, $\Delta\tau_c = -1$ msec. With three remote participants, the suggested strategy would be the same for the first two participants with the addition of the third remote participant's microphone signal presented to both loudspeakers with $\Delta I_c = 0$ dB, $\Delta\tau_c = 0$ msec.

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

$$x_i(n) = g_i(n) * s(n),\ i = 1, 2, \tag{1}$$

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

$$\begin{aligned} \mathbf{g_1} &= \left[\begin{array}{ccccc} 1 & 0 & \cdots & 0 & 0 \end{array}\right]^T \\ \mathbf{g_2} &= \left[\begin{array}{ccccc} 0 & 0 & \cdots & 0 & 0.5 \end{array}\right]^T, \end{aligned}$$

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.

3.2. The Nonuniqueness Problem and the Proposed Solution

Let

$$\mathbf{x}_i(n) = \left[\begin{array}{cccc} x_i(n) & x_i(n-1) & \cdots & x_i(n-L+1) \end{array}\right]^T, \quad i = 1, 2,$$

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

$$\mathbf{w}_i = \left[\begin{array}{cccc} w_{i,0} & w_{i,1} & \cdots & w_{i,L-1} \end{array}\right]^T,\ i = 1, 2,$$

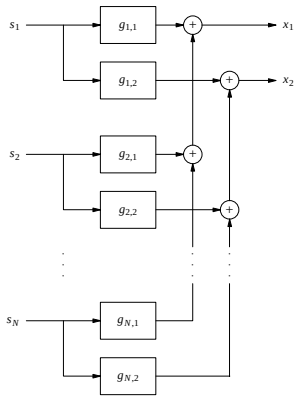

Figure 2: Synthesizing local stereo signals from N remote signals.

be the two model filters. Let

$$\mathbf{R} = E\left\{ \left[\begin{array}{c} \mathbf{x}_1(n) \\ \mathbf{x}_2(n) \end{array}\right] \left[\begin{array}{cc} \mathbf{x}_1^T(n) & \mathbf{x}_2^T(n) \end{array}\right] \right\}$$

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

$$\mathbf{r} = E\left\{ \left[\begin{array}{c} \mathbf{x}_1(n) \\ \mathbf{x}_2(n) \end{array}\right] y(n) \right\}$$

be the cross-correlation vector between the input signals and output (microphone) signal $y(n)$. Then the minimization problem to obtain the model filters $\mathbf{w}_1$ and $\mathbf{w}_2$ leads to the classical Wiener-Hopf equation [2]:

$$\mathbf{R}\left[\begin{array}{c} \mathbf{w}_1 \\ \mathbf{w}_2 \end{array}\right] = \mathbf{r}. \tag{2}$$

We also have the relation [2]

$$\mathbf{x}_1^T(n)\mathbf{g}_2 = \mathbf{x}_2^T(n)\mathbf{g}_1 \tag{3}$$

where $\mathbf{g}_1$ and $\mathbf{g}_2$ are the two impulse responses for synthesizing the stereo signals. This linear relation follows from (1), giving $x_1 * g_2 = s * g_1 * g_2 = x_2 * g_1$.

Consider the vector $\mathbf{u} = \left[\begin{array}{cc} \mathbf{g}_2^T & -\mathbf{g}_1^T \end{array}\right]^T$. We can verify using (3) that $\mathbf{Ru} = \mathbf{0}_{2L\times 1}$, so $\mathbf{R}$ is not invertible. Then there is no unique solution to the problem and an adaptive algorithm will drive to any one of many possible solutions, which can be very different from the "true" desired solution $\mathbf{w}_1 = \mathbf{h}_1$ and $\mathbf{w}_2 = \mathbf{h}_2$. These nonunique "solutions" are dependent on the impulse responses $\mathbf{g}_1$ and $\mathbf{g}_2$. This, of course, is intolerable because $\mathbf{g}_1$ and $\mathbf{g}_2$ 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 stereophonic AEC is to first preprocess each input signal x_i by the nonlinear transformation [2]

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

where f is a nonlinear function, such as a simple half-wave rectifier. Such a transformation 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 α, 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, LMS type gradient algorithms will be very slow. Therefore, we propose to use a rapidly converging algorithm like the two-channel RLS.

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. A mathematical measure of distance, to be useful in speech processing, has to have a high correlation between its numerical value and the subjective distance judgment, as evaluated on real speech signals [7]. Since 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

$$d_{\mathrm{IS}} = \frac{\mathbf{a}_i^{'T}\mathbf{R}_i\mathbf{a}_i^{'}}{\mathbf{a}_i^{T}\mathbf{R}_i\mathbf{a}_i} - 1 \qquad (5)$$

where $\mathbf{R}_i$ is the Toeplitz autocorrelation matrix of the LPC model $\mathbf{a}_i$ of a speech signal frame $\mathbf{x}_i$ and $\mathbf{a}_i^{'}$ is the LPC model of the corresponding distorted speech signal frame $\mathbf{x}_i^{'}$. 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 with a nonlinearity (half-wave) $\alpha = 0.5$, the IS metric is still small (about 0.03).

We could also use the Ensemble Interval Histogram (EIH) distance (which is based on the EIH model) [8]. The interest in using this distance lies in its capability to mimic human judgement of quality. Indeed, according to [8] EIH is a very good predictor of mean opinion score but only if the two speech observations under comparison are similar enough, 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 1.8×10^{-3}, which is as good as a 32 kb/s ADPCM coder.

A composite diagram of the synthesis, nonlinear transformation, and stereo AEC appears in Fig. 3. This shows the complete local signal processing suite for one conferee. This same setup is reproduced for each conferee, making obvious permutations on the remote signals and choice of synthesis filters. Further details and simulation results can be found in [9].

4. CONCLUSIONS

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 AEC, explaining the main problems and practical solutions for this application.

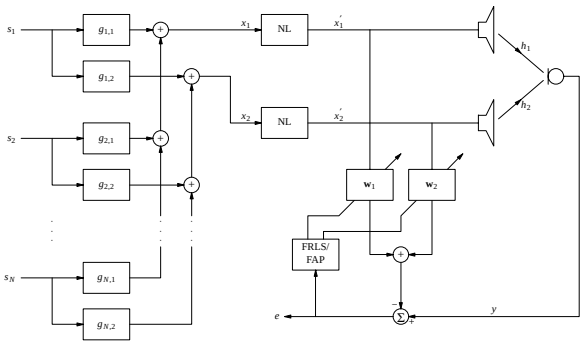

Figure 3: Overall diagram of synthesis, nonlinear transformation, and stereo acoustic echo cancellation.

5. ACKNOWLEDGMENTS

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

6. REFERENCES

[1] 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, No. 8, August 1995, pp. 148-151.

[2] 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, No. 2, pp. 156-165, Mar. 1998.

[3] M. B. Gardner, "Historical background of the Haas and/or precedence effect," *Acoust. Soc. Am.*, vol. 43, pp. 1243-1248, 1968.

[4] N. I. Durlach and H. S. Colburn, "Binaural phenomena," in *Handbook of Perception, Volume IV, Hearing*, E. C. Carterette and M. P. Friedman, Eds., New York: Academic Press, 1978, ch. 10.

[5] F. L. Wightman and D. J. Kistler, "Factors affecting the relative salience of sound localization cues," in *Binaural and Spatial Hearing in Real and Virtual Environments*, R. H. Gilkey and T. R. Anderson, Eds., New Jersey: LEA Publishers, 1997, ch. 1.

[6] J. Blauert, *Spatial Hearing*. Cambridge, MA: MIT Press, 1983, p. 206.

[7] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[8] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Trans. Speech Audio Processing*, vol 2, pp. 115-132, Jan. 1994.

[9] 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.*, to appear.