STEREOPHONIC ACOUSTIC ECHO CANCELLATION SYSTEM USING TIME-VARYING ALL-PASS FILTERING FOR SIGNAL DECORRELATION

Murtaza Ali

Wireline Communications Branch, DSPS R & D Center, Texas Instruments Inc. P.O. Box 655474, MS 446 Dallas, TX 75265 e-mail: mali@hc.ti.com, Phone: (972)-995-8033, Fax: (972)-995-6194

ABSTRACT

This paper describes a novel technique for decorrelating the stereo signals in stereophonic acoustic echo cancellation (AEC) systems. At present, most teleconferencing systems use a single full-duplex audio channel for voice communications. However, in order to introduce spatial realism, future teleconferencing systems are expected to have more than one channel (at least stereo with two channels). However, in stereophonic AEC systems, the correlation between the stereo signals does not allow correct identification of the echo path responses. In this paper, we develop a signal decorrelation technique based on time-varying all-pass filtering of the individual stereo signals. Experiments show that this technique does not effect the perception of the stereo signals, but identifies the echo path responses correctly.

1. INTRODUCTION

At present, most teleconferencing systems use a single full-duplex audio channel for voice communications. These systems also make use of an acoustic echo canceller to reduce the undesired echo resulting from the coupling between the loudspeaker and the microphone. To make these systems more lifelike, better and more realistic sound systems are required. High fidelity wide bandwidth (100 to 7000 Hz) voice communication system is now being used. However, in order to introduce spatial realism, more than one channel are needed. Therefore, future teleconferencing systems are expected to have more than one channel (at least stereo with two channels) of full duplex voice communications.

One of the fundamental problem in stereophonic acoustic echo cancellation (AEC) systems is that given the input to the loudspeakers and the output of the microphones in the receiving room, the echo path cannot be determined uniquely [2]-[5]. The problem is due to the correlation between the stereo signals. As a result, any adaptive technique used in stereophonic AEC systems fails to identify the echo path responses correctly. To circumvent this problem, it is necessary to develop techniques to decorrelate the stereo signals at the input to the loudspeakers without affecting stereo perception.

Several techniques have been proposed in the past, e.g., addition of random noise, modulation of signal, decorrelation filters, inter-channel frequency shifting etc. [4] [5]. However, these techniques either do not decorrelate the signals or destroy stereo perception completely. The interleaving comb filtering proposed in [5] only gives partial identification (above 1 kHz) of the echo path responses. Recently, a technique is proposed in [2] based on non-linear processing of the stereo signals. However, as noted by the authors of [2], for tonal signal, the technique based on nonlinearity cannot maintain transparency in perception (changes the pitch perception).

In this paper, we propose a different solution based on timevarying all-pass filtering of the stereo signals. The amount of time-variation allowed is restricted using the psychoacoustic data known as "the just noticeable inter-aural delay" [6] to maintain spatial perception. Our experiments show that this technique decorrelates the signals enough to allow identification of the true echo path responses, while maintaining transparency for speech signals. For a single tone, it introduces small background noise but maintains pitch perception. Since, audio/video conferencing rooms usually have inherent background noise, and noise suppression techniques are usually used in such systems, our technique is wellsuited for such applications.

This paper is organized as follows. Section 2 provides a brief description of a stereophonic teleconferencing system and the associated problem with stereophonic AEC. In Section 3, we describe our new technique for signal decorrelation. Finally, in Section 4, we present experimental evaluation of our proposed technique.

2. STEREOPHONIC ACOUSTIC ECHO CANCELLATION

Fig. 1 shows the configuration of a typical stereophonic echo cancellation system. The transmission room (depicted on the left) has two microphones that pick up the speech signal, x, via the two acoustic paths characterized by the impulse responses, g_1 and g_2 . All acoustic paths are assumed to include the microphone and/or loudspeaker responses. The i^{th} microphone output is then given by (in the frequency domain)

$$X_i(\omega) = G_i(\omega)X(\omega). \tag{1}$$

In this paper, the upper-case letters represent the Fourier transforms of the time-domain signals denoted by the corresponding lower-case letters. The whole system is considered as a discretetime system ignoring any A/D or D/A converter. These signals are presented through the set of loudspeakers in the receiving room (on the right in Fig. 1). Each microphone picks up an echo from each of the loudspeakers. Let h_{ij} be the acoustic path impulse response from the j^{th} loudspeaker to the i^{th} microphone. Then the echos picked up by the microphones in the receiving room are



Figure 1: Configuration of stereophonic echo cancellation systems

given by (in the frequency domain)

$$Y_i(\omega) = \sum_j H_{ij}(\omega) X_j(\omega).$$
 (2)

In the absence of any AEC, the echos y_i 's will be passed back to the loudspeaker in the transmission room and will be recirculated again and again. This will cause multiple echos or may even result in howling instability [5]. Commonly used AEC systems use adaptive finite impulse response (FIR) filters that provide estimates of the echo path responses. The FIR filter coefficients are updated adaptively depending on the input signals to the loudspeaker and the outputs of the microphones.

In the stereophonic AEC, there are four echo paths to be identified. We, therefore, need four adaptive filters as shown in Fig. 1. The output of the AEC filters (which can be thought of as an estimated echo) are as follows

$$\hat{Y}_i(\omega) = \sum_j \hat{H}_{ij}(\omega) X_j(\omega).$$

These estimated echos are subtracted from the true echos giving the error signals,

$$E_i(\omega) = Y_i(\omega) - \hat{Y}_i(\omega).$$

These error signals are used to update the filter coefficients. Several techniques are available to calculate the filter updates (e.g., the least means square (LMS), the recursive least square (RLS), the affine projection (AP) algorithms, etc.). All these techniques attempt to minimize these error signals in one way or another.

2.1. The problem of non-uniqueness of solutions

The data available to the echo canceller are the inputs to the loudspeakers, x_i 's, as well as the outputs of the microphones, y_i 's, in the receiving room. The fundamental problem of stereophonic AEC systems is that given this set of data, it is not possible to uniquely determine the echo paths to drive the error, e_i 's to zero (i.e., to eliminate the echos). In order to explain this, let us look at the error in one of the channels (similar analysis can be carried out for the other channels). In the frequency domain, this error is given by

$$E_1(\omega) = \sum_j \left(H_{1j}(\omega) - \hat{H}_{1j}(\omega) \right) G_j(\omega) X(\omega).$$



Figure 2: Configuration of the modified stereophonic echo cancellation systems

Let us assume that somehow, we have been able to achieve perfect echo cancellation, i.e., we have $E_1(\omega) = 0$. Assuming that $X(\omega)$ does not have zeros in the frequencies of interest, the above gives

$$\sum_{j} \left(H_{1j}(\omega) - \hat{H}_{1j}(\omega) \right) G_j(\omega) = 0.$$
(3)

This equation does not imply $H_{1j}(\omega) = \hat{H}_{1j}(\omega)$. Therefore, even if the echo has been driven to zero, we have not necessarily achieved perfect alignment. In other words, the canceller has not necessarily identified the true echo path. In fact, the above equation has infinitely many solutions for $\hat{H}_{1j}(\omega)$. Any adaptation algorithm may lead to any one of these solutions. Note that so long as the conditions in both the transmitting and the receiving rooms are fixed, this does not cause any problem as the echo will remain zero. However, the adaptation technique has to track not only the changes in the receiving room that change the echo path responses, h_{ij} , but also the changes in the conditions in the transmitting room as reflected through changes in g_i . Tracking the conditions in the transmitting room can be specially problematic as g_i may change abruptly and by a large amount (e.g., one speaker stops talking and another speaker starts speaking from a different location).

A detailed discussion of this problem describing several viewpoints can be found in [2]-[5]. Specially, the discussion in [2] provides a better understanding of the above problem both in terms of non-uniqueness and misalignment of the solutions.

3. SIGNAL DECORRELATION

As discussed in Section 2, the reason for non-perfect alignment is that the two signals are correlated due to (1). Thus, in order to solve the problem, we have to find a technique to decorrelate the input signals to the loudspeakers, x_i , in such a way that it does not affect the stereo perception in the receiving room.

The proposed system for the stereophonic echo cancellation system is shown in Fig. 2. Each of the stereo signals is passed through a different all-pass filter denoted by $a_i(n)$. The subscript n is used to indicate that the all-pass filter is time-varying (varying with n).

Rigorously speaking, there is no frequency domain representation of the time-varying filtering operation used in Fig. 2. However, if we assume that $a_i(n)$ does not change much for a given window around time instant *n*, then it is possible to assign a frequency domain transfer function $A(\omega, n)$ to the filtering operation at time instant n. Then the frequency spectra of the output at time instant n can be formally written as

$$egin{array}{rll} Y_i(\omega,n)&=&\sum_j H_{ij}(\omega)A_j(\omega,n)X_j(\omega). \ \hat{Y}_i(\omega,n)&=&\sum_j \hat{H}_{ij}(\omega)A_j(\omega,n)X_j(\omega). \end{array}$$

Then the error in the i^{th} path is

$$E_{i}(\omega, n) = \sum_{j} \left(H_{ij}(\omega) - \hat{H}_{ij}(\omega) \right) A_{j}(\omega, n) G_{j}(\omega) X(\omega)$$

Now, if we can achieve perfect echo cancellation by setting $E_i(\omega, n) = 0$, then the above implies

$$\sum_{j} \left(H_{ij}(\omega) - \hat{H}_{ij}(\omega) \right) A_{j}(\omega, n) G_{j}(\omega) X(\omega) = 0.$$

Since the above must be true for all n, i.e., for all variations of $A_j(\omega, n)$ with n, we must have $H_{ij}(\omega) = \hat{H}_{ij}(\omega)$. Thus by using the time varying all-pass filter in the signal path, it is possible to achieve perfect alignment between the adaptive filter and the true echo path. In practice, perfect alignment is not possible due to the finite impulse response of the modeling filters (the adaptive filters) as well as due to the noise present in the signal. However, simulations show that this technique achieves much better identification of the echo paths than was otherwise possible (see section 4).

3.1. Time-varying all pass filter

The system described above, however, must follow certain constraints. First, The signals that are modified through the all-pass filters are played back through the loudspeaker in the receiving room. Therefore, the time-variation of the all-pass filters has to be chosen in such a way that does not alter the stereo perception of the speech. Second, since an adaptive filter will be used to identify the echo path responses, the time-variation of the all-pass filters should be fast enough so that the adaptive technique used cannot track the changes in the all-pass filters. On the other hand, we would like the adaptive technique to be able to track changes in the receiving room. These conflicting requirements show the importance of proper choice of the time-varying all-pass filters. In the following, we discuss one possible choice.

The simplest all-pass filter is a single-order filter that can be described by a single parameter $\alpha_i(n)$. The frequency response of such a system for a given *n* can be written as

$$A_i(\omega, n) = \frac{e^{-j\omega} - \alpha_i(n)}{1 - \alpha_i(n)e^{-j\omega}}$$

Such a filter has several important features, namely

- |A_i(ω, n)| = 1.0, ∀ω and ∀n, i.e., this filter passes all frequencies all the time unattenuated.
- It only changes the phase of each frequency.
- It is completely determined by a single time-varying parameter α_i(n). Thus, the design of the system involves proper choice of α_i(n).

3.2. Choice of $\alpha_i(n)$

In order for the all-pass filter $a_i(n)$ to be stable, the absolute value of $\alpha_i(n)$ must be less than unity. Since, all our signal is real, we have also restricted $\alpha_i(n)$ to be a real value. This also simplifies the filtering operation. $\alpha_i(n)$ is a time-varying parameter. Thus, we need to update $\alpha_i(n)$ at every time instant. The update rule for $\alpha_i(n)$ is as follows

$$\alpha_i(n+1) = \alpha_i(n) + r_i(n),$$

set $\alpha_i(n+1) = \alpha_{i,max}$ if $\alpha_i(n+1) > \alpha_{i,max}$
set $\alpha_i(n+1) = \alpha_{i,min}$ if $\alpha_i(n+1) < \alpha_{i,min}.$ (4)

Here, $r_i(n)$ is an independent and identically distributed (iid) random variable having a uniform probability distribution function (pdf) over the interval $[-R_i, R_i]$. R_i indicates the maximum allowable deviation of $\alpha_i(n)$ from one instant to another. This deviation corresponds to phase jitter introduced by the time-varying all-pass filter for the i^{th} channel. R_i should be made as large as possible to introduce enough signal decorrelation. However, Too large a value of R_i will result in noticeable change in speech perception.

 $\alpha_{i,max}$ and $\alpha_{i,min}$ in (4), represent the the maximum and minimum allowable values of $\alpha_i(n)$. In order to ensure stability, we must have $\alpha_{i,max} < 1$ and $\alpha_{i,min} > -1$. Further restrictions are also required to maintain transparency in speech perception. These restrictions are derived from the data known as "just noticeable inter-aural delay" in psychoacoustics [6]. This data represents the minimum change in the inter-aural time delay between the two ears at a given frequency that causes a noticeable change in the perception of the direction of sound. The all-pass filter changes the phase of each frequency of the input speech. The effect of this phase change is to change the time of arrival of the signal at each frequency at the ears. So, if we limit the phase changes so that the change in the time of arrival for each channel is within the just noticeable inter-aural delay, then spatial perception of stereo signal will not be affected. The just noticeable inter-aural delay varies between 30 μ sec. to 200 μ sec [6]. We have chosen to limit the change in the time of arrival of each frequency within 60 μ sec. This leads to the following values of $\alpha_{i,max}$ and $\alpha_{i,min}$,

$$\alpha_{i,max} = 0$$
 and
 $\alpha_{i,min} = -0.9$ and

Fig. 3 shows the time delay as function of frequency for the two all-pass filters with $\alpha_{i,min} = -0.9$ and $\alpha_{i,max} = 0$. Since, the value of $\alpha_{i,min}$ for the all-pass filters in the two stereo paths are kept within these limits, the resulting inter-aural delay are also within 60 μ sec. Our experiments have shown that this choice leads to good signal decorrelation to allow correct identification of echo path responses and also keeps the stereo perception of speech unchanged.

4. EXPERIMENTAL EVALUATION

In order to evaluate the technique, we collected stereo speech samples in our audio laboratory. The audio laboratory was used as the transmitting room. We had two speakers talking alternately in the room when two microphones were used to collect the data. The data were sampled at 16 kHz sampling rate. In one set of data, the speakers were asked to stand still while talking. This was made



Figure 3: Time-delay vs frequency for the two all-pass filters with $\alpha_{i,min} = -0.9$ and $\alpha_{i,max} = 0$

to ensure that the echo path responses remain the same. In another set, they were free to move around the room as they talked into the microphones. We then used our technique to decorrelate the collected stereo signals. We performed informal listening tests by playing the original and the modified stereo signals over both loudspeakers and headphones. All these tests show that the stereo perception of the modified signal is indistinguishable from that of the original.

We simulated the receiving room loudspeaker outputs by convolving the stereo signals using the echo path responses h_{11} and h_{12} . These two echo path responses were obtained using the image method of [1] based on room measurements of one of our conference rooms. The microphone output in the receiving room was simulated by summing up the outputs of these two convolutions. In the above convolutions, we restricted the lengths of the echo path responses to be N = 4096 samples long. We then used the two adaptive filters \hat{h}_{11} and \hat{h}_{12} each of length L = 2048 samples, to identify these echo path responses. We used the fast affine projection technique of order 8 for updating the filter coefficients [4]. Fig. 4 shows the misalignment in dB with time. The misalignment is defined as

$$10 * log_{10} \frac{\|h_{11,1:2048} - \hat{h}_{11}\|_2^2 + \|h_{12,1:2048} - \hat{h}_{12}\|_2^2}{\|h_{11,1:2048}\|_2^2 + \|h_{12,1:2048}\|_2^2}$$

where, the subscript 1 : 2048 is used to indicate that the first 2048 samples of the corresponding echo path responses have been used here. This figure corresponds to the set of data when the transmitting room echo path responses were kept fixed as already described. The dotted line corresponds to the case of original signal and the solid line to the case of modified data using our technique of time-varying all pass filtering.

Since, we have used 'real-world' collected data for the transmitted signals, the situation was not as bad as when simulated data was used. We did not experience sudden jumps, but misalignment settles down at around -14 dB. whereas with our technique of signal decorrelation, the misalignment goes below -20 dB.



Figure 4: Behavior of misalignment with original stereo signal and with stereo signal modified using time-varying all-pass filtering

5. CONCLUSION

Future audio/video conferencing systems are expected to employ stereo audio communication. These systems require stereophonic AEC. This paper describes a new effective but simple technique to decorrelate the stereo signals so that correct identification of the stereophonic path responses is possible. The technique uses a time-varying single-pole all-pass filter in each channel. The timevarying filter parameter is chosen in such a way that it does not effect stereo speech perception but introduces enough decorrelation among the signals in different channels.

6. REFERENCES

- J. Allen and D. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics," *J. Acoust. Soc. Am.*, Vol. 65, No. 4, pp. 943-950, April, 1979.
- [2] J. Benesty, D. R. Morgan and M. M. Sondhi, "A Better Understanding and an Improved Solution to the Problems of Stereophonic Acoustic Echo Cancellation," Preprint, *Proceedings of ICASSP-97*, Vol. 1, pp. 303-306, Munich, Germany, April 21-24, 1997.
- [3] J. Benesty, P. Duhamel and Y. Grenier, "Multi-Channel Adaptive Filtering Applied to Multi-Channel Acoustic Echo Cancellation," Preprint, Submitted to *IEEE Trans. on Signal Processing*, April 1995.
- [4] S. Shimauchi and S. Makino, "Stereo Projection Echo Canceller with True Echo Path Estimation," *Proceedings of ICASSP*-95, pp. 3059-3062, 1995.
- [5] M. M. Sondhi, D. R. Morgan and J. L. Hall, "Stereophonic Acoustic Echo Cancellation– An Overview of the Fundamental Problem," *IEEE Signal Processing Letters*, vol. 2, no. 8, pp. 148-151, August 1995.
- [6] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, Heidelberg, Germany: Springer-Verlag, 1990.