STEREOPHONIC ACOUSTIC ECHO CANCELLATION USING NONLINEAR TRANSFORMATIONS AND COMB FILTERING

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

Stereophonic sound becomes more and more important in a growing number of applications (such as teleconferencing, multimedia workstations, televideo gaming, etc.) where spatial realism is demanded. Such hands-free systems need stereophonic acoustic echo cancelers (AECs) to reduce echos that result from coupling between loudspeakers and microphones in full-duplex communication. In this paper we propose a new stereo AEC based on two experimental observations:(a) the stereo effect is due mostly to sound energy below about 1 kHz and (b) comb filtering above 1 kHz does not degrade auditory localization. The principle of the proposed structure is to use one stereo AEC at low frequencies (e.g. below 1 kHz) with nonlinear transformations on the input signals and another stereo AEC at higher frequencies (e.g. above 1 kHz) with complementary comb filters on the input signals.

1. INTRODUCTION

Stereophonic sound becomes more and more important in a growing number of applications (such as teleconferencing, multimedia workstations, televideo gaming, etc.) where spatial realism is demanded. Such hands-free systems need stereophonic acoustic echo cancelers (AECs) to reduce echos that result from coupling between loudspeakers and microphones in full-duplex communication [1].

Stereophonic acoustic echo cancellation can be viewed as a straightforward generalization of the single-channel acoustic echo cancellation principle [1]. Figure 1 shows this technique for one microphone in the "receiving room" (which is represented by the two echo paths h_1 and h_2 between the two loudspeakers and the microphone). The two reference signals x_1 and x_2 from the "transmission room" are obtained either by two microphones in the case of teleconferencing or by synthesizing stereo sound for localization from the output of a single microphone in the case of desktop conferencing. In both cases, the signals are derived by filtering from a common source, and this gives rise to a non-uniqueness problem that does not arise for the single-channel AEC [1], [2]. As a result, the usual adaptive algorithms converge to solutions that depend on the impulse responses in the (actual or synthesized) *transmission* room. This means that for good echo cancellation one must track not only the changes in the receiving room, but also the changes in the transmission room (for example, when one person stops talking and another person starts).

In [2], [3], we proposed a simple but efficient solution that overcomes the above problem by adding a small nonlinearity (NL) into each channel, as depicted in Fig. 1. The distortion due to the nonlinearity is hardly perceptible, yet it reduces interchannel coherence, thereby allowing reduction of misalignment to a low level. However, the processing load associated with Fig. 1 is exorbitant because the nonlinearity is small and therefore we must use rapidly-converging adaptive algorithms (e.g., two-channel fast recursive least-squares - FRLS). This requirement implies a high level of computational complexity. So a real-time implementation of such a scheme is difficult.

Recently, we presented a suboptimal structure [4], [5]. The principle of this structure (hybrid mono/stereo AEC) is to use stereophonic sound with a stereo AEC at low frequencies (e.g., below 1 kHz) and monophonic sound with a conventional mono AEC at higher frequencies (e.g., above 1 kHz). This solution is a good compromise between the complexity of a full-band stereo AEC and spatial realism. However, in some applications like televideo gaming, full spatial realism may be needed. In the following, we present another structure that converges faster and is easier to implement than the scheme of [3], while retaining more spatial realism than the hybrid scheme of [4], [5].

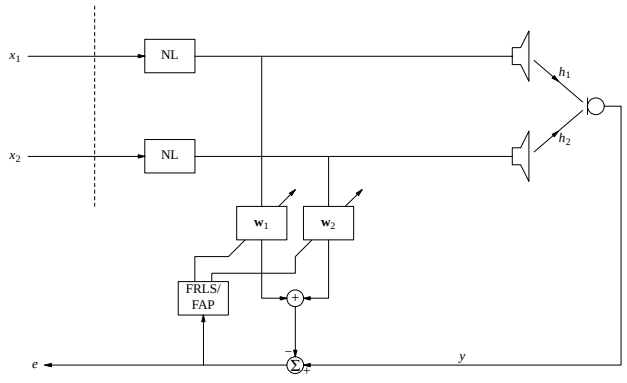

Figure 1: Schematic diagram of stereophonic echo cancellation.

2. PRINCIPLE OF THE MIXED STRUCTURE

Suppose that in a given frequency band, all of the signal energy in one channel were removed, leaving only a small amount of uncorrelated background noise. In that case, the coherence would reduce to exactly zero in that band. This motivates the idea of using complementary comb filters [1] to separate frequency components between the left and right channels. This separation permits

the unique identification of relevant portions of the receiving room impulse responses, while at the same time not destroying spatial realism as explained below.

Many experiments show that the dominant stereophonic cues are located below about 1 kHz [6], [7], [8]. Comb filtering below about 1 kHz destroys these cues and degrades localization performance. However, if the comb filtering is restricted to frequencies above 1 kHz, localization performance is almost unimpaired. Note that the "hollow" sound accompanying comb-filtered monophonic representations is greatly reduced under conditions associated with stereophonic presentation [9].

Based on the above psychoacoustical principles, Fig. 2 shows a way to transmit the two microphone signals to the receiving room and also shows a set of stereo AECs matched to these signals. We decompose the two input signals x_1 and x_2 (left and right) into two bands: the low-frequency band (below f_c, where the crossover frequency f_c is on the order of 1 kHz) and the high-frequency band (above f_c). The goal is to process these two bands differently in order to reduce the processing load associated with Fig. 1. In each low-frequency channel we put a nonlinear transformation (NL) to help the adaptive algorithm converge to the "true" solution [3]. This nonlinearity can be larger when used in the low-frequency band alone than when used in the full-band (as in [3]) because the distortion is confined to the low-frequency band; a higher level of this nonlinear transformation implies an improvement of the misalignment convergence rate. In the high-frequency band, the two input signals are filtered by two complementary comb filters (C1 and C2) to allow a unique solution as explained above. A gain factor A is included to adjust the spectral balance.

The above structure is much more efficient than a fullband system, despite the fact that we have two stereo AECs. Indeed, for the low-frequency band, since the maximum frequency is f_c (on the order of 1 kHz), we can subsample the signals by a factor $r = f_s/(2f_c)$, where f_s is the sampling rate of the system. As a result, the arithmetic complexity is divided by r^2 in comparison with a fullband implementation (the number of taps and the number of filter computations per second are both reduced by r). In this case, we can afford to use a rapidly converging adaptive algorithm like the two-channel FRLS [10]. On the other hand, the simple two-channel NLMS algorithm can be used to update the filter coefficients in each non-overlapping high-frequency band; convergence may be slower but this is of little concern since most of the energy in speech is at low frequencies.

Thus, with this proposed structure, we decrease the complexity of the system and increase the convergence rate of the adaptive algorithms, while preserving the stereo effect.

3. ADAPTIVE ALGORITHMS AND SIGNAL TRANSFORMATIONS

In this section, we explain in more detail the adaptive algorithms and signal transformations that are used in each band of the proposed structure. It is possible to have good steady-state echo cancellation even if the adaptive algorithm does not accurately identify the impulse responses h_1 and h_2. However, in such a case, the cancellation will temporarily degrade if the impulse responses in the (actual or synthesized) transmission room change, since the algorithm will have to reconverge [1]. The main goal here is to avoid this problem, and that is why signal transformations are used.

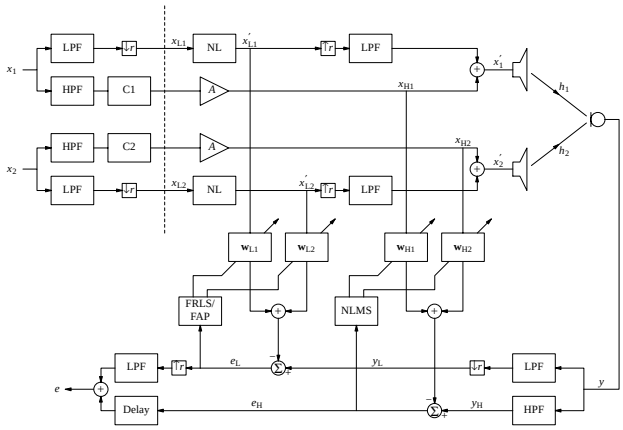

Figure 2: Mixed Stereo acoustic echo canceler.

3.1. Low Band

The best way we know to alleviate the characteristic non-uniqueness of a stereophonic AEC is to first preprocess each input signal $x_{\mathrm{L}i}$ by a nonlinear transformation [2], [3]:

$$x'_{\mathrm{L}i}(n) = x_{\mathrm{L}i}(n) + \alpha f[x_{\mathrm{L}i}(n)], \tag{1}$$

where f is a nonlinear function, such as a simple half-wave rectifier. Such a transformation reduces the inter-channel coherence and hence the condition number of the covariance matrix, thereby greatly reducing the misalignment [3]. 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 for the low-frequency region.

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. A (computationally) fast stabilized version of this algorithm is given in [10] where the total number of operations is $28L$ multiplications and $28L$ additions per sample. We can also use a two-channel affine projection algorithm [11]. All of these algorithms can be implemented in subbands for a real-time application [12].

The misalignment in this band is computed as

$$\varepsilon_{\mathrm{L}} = \frac{\|\mathrm{LPF}\downarrow\{\mathbf{h}_1\} - \mathbf{w}_{\mathrm{L}1}\|^2 + \|\mathrm{LPF}\downarrow\{\mathbf{h}_2\} - \mathbf{w}_{\mathrm{L}2}\|^2}{\|\mathrm{LPF}\downarrow\{\mathbf{h}_1\}\|^2 + \|\mathrm{LPF}\downarrow\{\mathbf{h}_2\}\|^2} \tag{2}$$

where $\mathrm{LPF}\downarrow$ denotes lowpass filtering and downsampling.

3.2. High Band

Two complementary linear-phase comb filters, C1 and C2, of length 256 (see Fig. 3) were designed to operate above about 1 kHz. Each comb filter has approximately two lobes per auditory critical band. The filters were constructed by first designing a prototypical linear-phase FIR filter centered at 4 kHz with a length of 325 points, a pass band of $1/12$ octave, and transition bands of $1/24$ octave. This lobe was frequency scaled to obtain a family of lobes centered at $1/12$-octave intervals from one to six kHz. Each lobe was upsampled to 256 kHz, padded with zeros to equalize

the group delay in all lobes, and downsampled to 16 kHz. Alternate lobes ($1/6$-octave spacing) were added together to produce the complementary comb filters C1 and C2, so that any specified frequency from one to six kHz falls within the passband of either the C1 or the C2 filter. The in-band and out-of-band weighting of the prototypical lobe were specified such that the out-of-band rejection of the C1 or the C2 filter was 50 dB and the ripple of the combined C1 and C2 filters was ± 3 dB. The obtained complementary comb filters were of length 4096, but were truncated to 256 coefficients each in order to reduce the global delay of the system.

For the high-frequency band, we propose to use the two-channel NLMS algorithm (the performance of this class of algorithms should be adequate, since the two comb-filtered input signals are almost completely decorrelated). Here the two-channel NLMS algorithm uses high-pass reference signals $x_{\mathrm{H}1}$ and $x_{\mathrm{H}2}$, and common error signal e_{H}. In practice, this algorithm with the proposed decomposition converges fast since the echo energy is predominant at low frequencies, and therefore the spectral dynamic range is reduced in the highpass signals. We can furthermore use a subband structure for an efficient implementation of the NLMS algorithm.

The relevant misalignment in this band is computed as

$$\varepsilon_{\mathrm{H}} = \frac{\|\mathrm{HPFC1}\{\mathbf{h}_1 - \mathbf{w}_{\mathrm{H}1}\}\|^2 + \|\mathrm{HPFC2}\{\mathbf{h}_2 - \mathbf{w}_{\mathrm{H}2}\}\|^2}{\|\mathrm{HPFC1}\{\mathbf{h}_1\}\|^2 + \|\mathrm{HPFC2}\{\mathbf{h}_2\}\|^2} \quad (3)$$

where $\mathrm{HPFC1}$ and $\mathrm{HPFC2}$ denote highpass filtering and comb filtering by C1 and C2. Note that this misalignment is computed only in the highpass region and in the passbands of the two complementary comb filters, since there is no energy either at low frequencies or between the teeth.

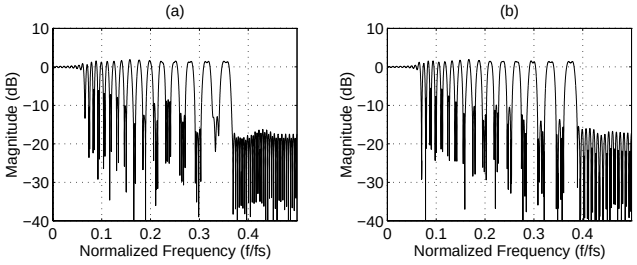

Figure 3: Frequency response of the two complementary linear-phase comb filters. (a) Comb filter C1. (b) Comb filter C2.

3.3. Computational Complexity

Suppose that the length of the adaptive filters necessary to have a good level of echo cancellation in a full-band stereo AEC is equal to L. We already know that the total number of operations per iteration is $28L$ multiplications and $28L$ additions for the two-channel FRLS, and $4L$ multiplications and $4L$ additions for the two-channel NLMS algorithm. Now, suppose this length is taken the same (with respect to the downsampling/upsampling factor r) to have the same level of echo cancellation for the proposed structure of Fig. 2. Then, our structure will require per iteration $28L/r^2 + 4L$ multiplications and $28L/r^2 + 4L$ additions. For example, with a 16 kHz sampling frequency and $f_c \approx 1$ kHz, we take $r = 8$. In this case, the structure of Fig. 2 will need approximately $4.4L$ multiplications and the same number of additions per iteration, to be compared with $28L$ multiplications and $28L$ additions for a full-band stereo AEC with a two-channel FRLS. Thus, the computational complexity is reduced by a factor of about six. In practice, we achieve an even greater reduction since increased room absorption and lower speech energy at higher frequencies permits a reduction of the number of taps used in the high frequency band.

4. SIMULATIONS

We now determine the performance of the proposed structure in Fig. 2 by simulation. The signal source s in the transmission room is a speech signal sampled at 16 kHz. It consists of the following three sentences:

"Bobby did a good deed."
"Do you abide by your bid?"
"A teacher patched it up."

(This is the same speech signal used in [2], [3], [4], [5].) The two microphone signals were obtained by convolving s with two impulse responses g_1, g_2 of length 4096, which were measured in an actual room (HuMaNet I, room B [13]). The microphone output signal y in the receiving room is obtained by summing the two signals $(h_1 * x_1^{'})$ and $(h_2 * x_2^{'})$, where h_1 and h_2 were also measured in an actual room (HuMaNet I, room A [13]) as 4096-point responses, and $x_1^{'}$ and $x_2^{'}$ are the two loudspeaker signals. For all of our simulations, we have used the two-channel NLMS algorithm for the high-frequency band, and taken the length of each of the two adaptive filters $w_{\mathrm{H}1}$ and $w_{\mathrm{H}2}$ to be $L_{\mathrm{H}} = 550$. For the low-frequency band, we have used the two-channel FRLS algorithm [10], with $\lambda = 1 - 1/(10L_{\mathrm{L}})$ and the length of each of the two adaptive filters $w_{\mathrm{L}1}$ and $w_{\mathrm{L}2}$ is $L_{\mathrm{L}} = 256$. We chose a crossover frequency f_c of 900 Hz, and in consideration of the 16 kHz sampling frequency, used a downsampling/upsampling factor $r = 8$. Two 256-tap FIR lowpass (100 - 900 Hz) and highpass (900 - 8000 Hz) filters were designed using the Matlab fir1 routine [4], [5]. Here we use $A = 1$ since the nonlinearity boosts the low frequencies somewhat which moreover tend to add more coherently than high frequencies in a reverberant room.

Figures 4 and 5 show the mean square error (MSE) of the NLMS algorithm (high frequencies), the MSE of the FRLS algorithm (low frequencies), the MSE of the combined signal, and the misalignment for each AEC in its respective band. The misalignment in each band was computed as in (2) and (3). In Fig. 4, there are no nonlinear transformations on the two input signals and no comb filters separating the high band. We can notice how bad the two misalignments are (lower right panel). In Fig. 5 we use a half-wave rectifier with $\alpha = 1.0$ for the nonlinear transformation [4], [5]. With this value, there is little audible degradation of the original signal and the stereo effect in the low-frequency band is not affected. (We determined from informal listening tests that $\alpha = 1.0$ for the low-band was as innocuous as $\alpha = 0.3$ previously used for the full band case reported in [3].) For the high-frequency band, we use the two complementary comb filters of Fig. 3. In this case, the misalignment is greatly reduced. Note that the high band misalignment is still decreasing after 4 seconds due to slow convergence of the NLMS algorithm. However, this is less of a concern than in the low band where most of the energy is concentrated for speech.

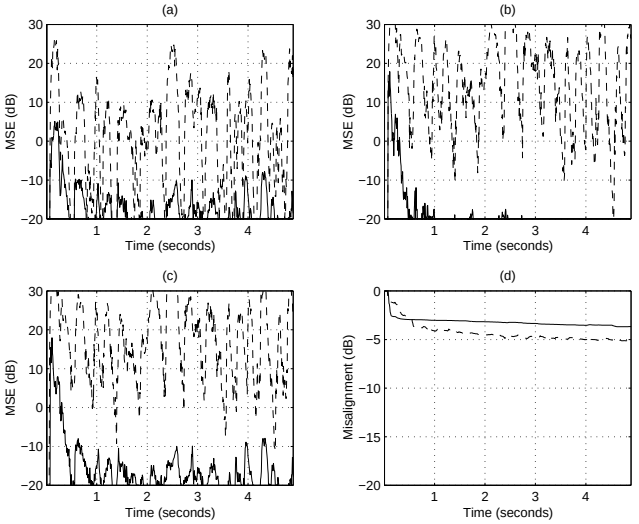

Figure 4: Performance of the mixed stereo AEC using the NLMS algorithm with $L_{\mathrm{H}} = 550$ at high frequencies without C1 and C2, and the FRLS algorithm with $L_{\mathrm{L}} = 256$ and $\alpha = 0$ at low frequencies. (a)-(c) MSE (–) as compared to original echo level (– –) at high frequencies (a), low frequencies (b), and combined (c). (d) misalignment of stereo AEC at low frequencies (–) and stereo AEC at high frequencies (– –).

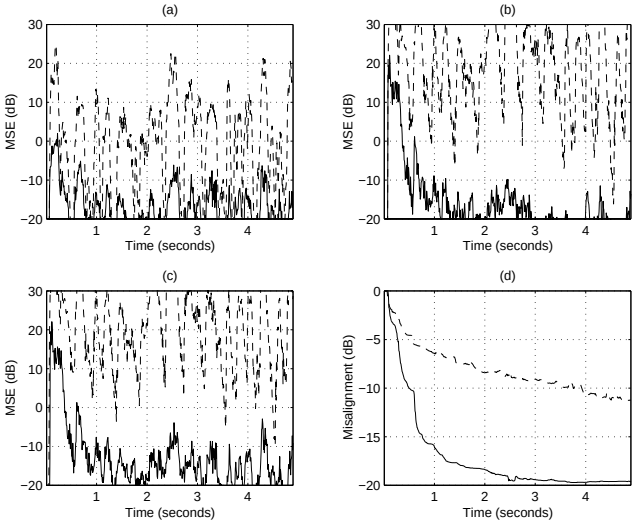

Figure 5: Performance of the mixed stereo AEC using the NLMS algorithm with $L_{\mathrm{H}} = 550$ at high frequencies with C1 and C2, and the FRLS algorithm with $L_{\mathrm{L}} = 256$ and $\alpha = 1.0$ at low frequencies. (a)-(c) MSE (–) as compared to original echo level (– –) at high frequencies (a), low frequencies (b), and combined (c). (d) misalignment of stereo AEC at low frequencies (–) and mono AEC at high frequencies (– –).

5. CONCLUSION

Thanks to new findings in psychoacoustics, we proposed a new structure (Fig. 2) to reduce the computational complexity associated with the structure of Fig. 1. We combined two different effective means for reducing the misalignment that exploit some simple psychoacoustical principles of stereo sound. This structure is an extension to the one that we recently proposed in [4], [5] and can be a possible solution to an application like televideo gaming.

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 problems of stereophonic acoustic echo cancellation," in *Proc. IEEE ICASSP*, 1997, pp 303-306.

[3] 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," to appear in *IEEE Trans. Speech Audio Processing*.

[4] J. Benesty, D. R. Morgan, and M. M. Sondhi, "A hybrid mono/stereo acoustic echo canceler," to appear in *Proc. IEEE ASSP Workshop Appls. Signal Processing Audio Acoustics*, 1997.

[5] J. Benesty, D. R. Morgan, and M. M. Sondhi, "A hybrid mono/stereo acoustic echo canceler," submitted for publication in *IEEE Trans. Speech Audio Processing*.

[6] W. A. Yost, F. L. Wightman, and D. M. Green, "Lateralization of filtered clicks," *J. Acoust. Soc. Am.*, vol. 50, pp. 1526-1531, 1971.

[7] F. L. Wightman and D. J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *J. Acoust. Soc. Am.*, vol. 91, pp. 1648-1661, Mar. 1992.

[8] R. M. Stern and G. D. Shear, "Lateralization and detection of low-frequency binaural stimuli: Effects of distribution of internal delay," *J. Acoust. Soc. Am.*, vol. 100, pp. 2278-2288, Oct. 1996.

[9] P. M. Zurek, "Measurements of binaural echo suppression," *J. Acoust. Soc. Am.*, vol. 66, pp. 1750-1757, Dec. 1979.

[10] J. Benesty, F. Amand, A. Gilloire, and Y. Grenier, "Adaptive filtering algorithms for stereophonic acoustic echo cancellation," in *Proc. IEEE ICASSP*, 1995, pp. 3099-3102.

[11] J. Benesty, P. Duhamel, and Y. Grenier, "A multi-channel affine projection algorithm with applications to multichannel acoustic echo cancellation," *IEEE Signal Processing Lett.*, Vol. 3, pp. 35-37, Feb. 1996.

[12] S. Makino, K. Strauss, S. Shimauchi, Y. Haneda, and A. Nakagawa, "Subband stereo echo canceller using the projection algorithm with fast convergence to the true echo path," in *Proc. IEEE ICASSP*, 1997, pp. 299-302.

[13] D. A. Berkley and J. L. Flanagan, "HuMaNet: an experimental human-machine communications network based on ISDN wideband audio," *AT&T Tech. J.*, vol. 69, pp. 87-99. Sept./Oct. 1990.