# AN IMPROVED METHOD FOR STEREO ACOUSTIC ECHO CANCELING

*Maximilian Gauger*

Fachgebiet Theorie der Signale, Technische Universität Darmstadt
Merckstr. 25, 64283 Darmstadt, Germany
gauger@nt.tu-darmstadt.de

## ABSTRACT

This paper introduces an improved version of the well-known frequency-domain LMS. We cope with the correlation problem by choosing the step-size according to the correlation factor and by using the novel delay-and-add algorithm which makes use of changes in the transmission room. The combination of these techniques allows to minimize the distortion of the loudspeaker signal.

## 1. INTRODUCTION

Stereo Echo Cancelation is mainly used in two applications. The first application is (video) conference systems, where stereo transmission allows for spatial differentation of the remote speakers, which in turn leads to a feeling of greater reality. This setup usually uses two loudspeakers and two microphones.

The second application involves computer games and command-and-control environments (e.g. voice control in cars). These applications mostly use only one microphone, but also stereo signals on their loudspeaker outputs.

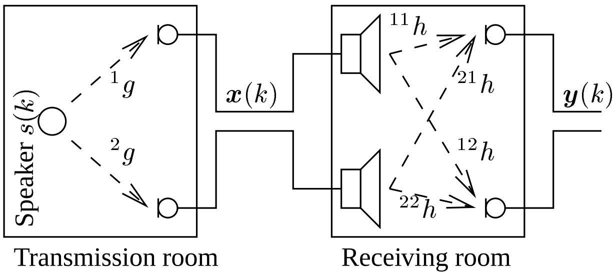

**Fig. 1**. Typical setup of a stereo confererencing system

Contrary to the mono-channel echo cancelation, the stereo case requires the estimation of two or four echo paths per room (see Fig. 1). Not only does this lead to an increased computing load, the convergence of the adaptive algorithm is very much decreased by the fact that the input signals are correlated, i.e. linearly dependent.

To cope with this linear dependency, an artificial distortion of the stereo signal $\boldsymbol{x}(k)$ is used in most approaches to stereo echo canceling[1, 2]. Either the phase or the amplitude can be distorted. To distort the phase, a time-variant allpass filter with random coefficients can be used [1], for the amplitude distortion, a nonlinearity like the half-wave rectifier proposed by Benesty et al. [2] is widely used.

The central issue of these preprocessing approaches is audibility. Any pre-processing which alters the signal will result in some degradation of signal quality. The allpass approach degrades the stereo effect, while the nonlinearity introduces additional noise, which is particularly disturbing when applied to music signals.

The work presented in this paper uses two ways to reduce the amount of distortion required. First, the step size of the adaption is reduced in sections where both channels are highly correlated. Second, the usually unwanted changes in the transmission room are exploited by the new delay-and-add (DNA) algorithm.

## 2. STEREO ACOUSTIC ECHO CANCELING

### 2.1. The Non-Uniqueness Problem

Using the setup shown in Fig. 1, we can express the signal at the receiving room microphone 1 as follows in the $z$ domain:

$$^{1}Y(z) = S(z)\left(\,^{1}G(z)\,^{11}H(z) + \,^{2}G(z)\,^{21}H(z)\right) \qquad (1)$$

$S(z)$ stands for the frequency domain representation of the speaker signal, $^{c}G(z)$ for the transfer function from the speaker to microphone $c$. $\boldsymbol{x}(k)$, $^{ab}H(z)$ is the transmission function from loudspeaker $a$ to microphone $b$ in the receiving room.

To achieve $^{c}\hat{Y}(z) = \,^{c}Y(z)$, we must find a tuple

$$\{\,^{11}\hat{H}(z),\,^{21}\hat{H}(z)\}$$

for which

$$\begin{array}{c} ^{1}G(z)\,^{11}H(z) + \,^{2}G(z)\,^{21}H(z) = \\ ^{1}G(z)\,^{11}\hat{H}(z) + \,^{2}G(z)\,^{21}\hat{H}(z) \end{array} \qquad (2)$$

The existence of such a tuple does not imply that $^{11}\hat{H}(z) = \,^{11}H(z)$ and $^{21}\hat{H}(z) = \,^{21}H(z)$. This effect is called the non-uniqueness problem [3].

### 2.2. Acoustic Echo Canceling: The LMS Algorithm

To derive the adaptive algorithm used, we start from the well-known Least Mean Squares algorithm (LMS). A (mono) output signal $y(k)$ derived from a (mono) input signal $x(k)$ by convolving it with the impulse response $h(k)$ of the Loudspeaker - Room - Microphone (LRM) system is to be estimated by the adaptive algorithm.

$$\hat{y}(k) \quad = \quad \sum_{l=0}^{N} x(k-l) \cdot \hat{h}(l) \tag{3}$$

where $x(k)$ is the input signal of the LRM system, $\hat{y}(k)$ its estimated output and $\hat{h}(l)$ the estimate of the impulse response.

We can now define the error

$$e(k) = y(k) - \hat{y}(k) \tag{4}$$

with $y(k)$ being the LRM output at time instant $k$.

Using the expectation of the squared error, $E\{e^2(k)\}$, as the cost function, and applying a number of simplifications, we arrive at the adaptive algorithm called (normalized) LMS [4, 5]. It is defined by its update equation:

$$\hat{h}_{k+1}(l) = \hat{h}_k(l) + \mu \frac{1}{\sigma_x^2} e(k) \cdot x(k-l) \tag{5}$$

where $k$ is the time index of the current sample, $l$ the index in the estimated impulse response vector, $\mu$ the step size and $\sigma_x^2$ the average power of the input signal.

### 2.3. The Block LMS

To simplify the calculations, the LMS algorithm can be re-written by not re-calculating $\hat{h}(k)$ at every input sample but rather every $N$ input samples, thus keeping the estimate constant during an entire block of length $N$ (which yields the name Block LMS). The update equation is now modified as follows and calculated only every $N$ samples:

$$\hat{h}_{k+N}(l) = \hat{h}_k(l) + \mu \frac{1}{N\sigma_x^2} \sum_{m=0}^{N-1} e(k-m) \cdot x(k-l-m) \tag{6}$$

This can be interpreted as averaging the update over an entire block and applying it only at the end of the block.

### 2.4. Calculation in the Frequency Domain

The convolution in Eq. 6 can be calculated using fast convolution, i.e. transforming all signals involved to the frequency domain. Setting the length of the adaptive filter equal to the blocklength $N$ and taking care of all constraints that the overlap-save algorithm imposes, we arrive at the following:

$$\begin{aligned}
X_k &= \text{FFT}_{2N}\{\boldsymbol{x}_k\} && (7)\\
Y_k &= \text{FFT}_{2N}\{[\mathbf{0}_N\ \boldsymbol{y}_k\}] && (8)\\
\hat{Y}_k &= X_k \odot H_k && (9)\\
E_k &= Y_k - \hat{Y}_k && (10)\\
\boldsymbol{e}_k &= \text{iFFT}_{2N}\{E_k\} && (11)\\
E_k &= \text{FFT}_{2N}\{[\mathbf{0}_N\ \boldsymbol{e}(N \ldots 2N-1)]\} && (12)\\
\hat{H}_{k+N} &= \hat{H}_k + \mu \cdot X_k^* \odot E_k && (13)
\end{aligned}$$

$\text{FFT}_{2N}$ and $\text{iFFT}_{2N}$ denote the Fast Fourier transform of length $2N$ and its inverse, respectively. $\boldsymbol{x}_k$ and $\boldsymbol{y}_k$ are the vectors of the last input resp. output samples at time instant $k$. $H_k$ and $H_{k+N}$ are the estimates for the room frequency response at the time instances $k$ and $k+N$. $\mu$ is the step size. $\odot$ stands for tap-wise multiplication, $^*$ for the conjugate complex operator.

### 2.5. Improved Step-size Control

As proposed in [5], p. 164, the step sizes can be chosen independently for each frequency bin, turning $\mu$ from the usual scalar into a vector. The highest adaption speed is achieved when $\mu(n) = ||X(n)||^{-2}$, i.e. the squared inverse of the corresponding frequency bin value. It has proven useful to limit this value to make sure the FLMS remains stable.

### 2.6. Extension to the Stereo Case

For the stereo case, we assume that two loudspeakers and two microphones are present, making it necessary to estimate four different impulse responses (more precisely, transfer functions when working in the FFT domain). The error signal in the time domain can be expressed as follows:

$$\begin{aligned}
{}^1e(k) &= {}^1y(k) - \textstyle\sum_{l=0}^{N-1} {}^1x(k-l) \cdot {}^{11}h(l) \\
&\qquad - \textstyle\sum_{l=0}^{N-1} {}^2x(k-l) \cdot {}^{21}h(l) && (14)\\
{}^2e(k) &= {}^2y(k) - \textstyle\sum_{l=0}^{N-1} {}^1x(k-l) \cdot {}^{12}h(l) \\
&\qquad - \textstyle\sum_{l=0}^{N-1} {}^2x(k-l) \cdot {}^{22}h(l) && (15)
\end{aligned}$$

${}^ce(k)$ stands for the error in channel $c$, $x(k)$ and $y(k)$ are indexed in the same manner. ${}^{ab}h(l)$ is tap $l$ of the impulse response relating the input channel (i.e. loudspeaker) $a$ to the output channel (i.e. microphone) $b$.

When using the standard LMS, four update equations are necessary to estimate all four impulse responses. The update equation for ${}^{ab}\hat{h}$ reads:

$${}^{ab}\hat{h}_{k+1}(l) = {}^{ab}\hat{h}_k(l) + \mu \frac{1}{\sigma_{x_1}^2 + \sigma_{x_2}^2} {}^be(k) \cdot {}^ax(k-l) \tag{16}$$

Applying this relation to the frequency domain algorithm shown in 2.4, using the frequency-selective step size from

2.5, and multiplying with another step-size factor $\alpha$ introduced in section 3, our final adaptive algorithm reads:

$$\begin{aligned}
{}^{c}X_k &= \mathrm{FFT}_{2N}\{\,{}^{c}\boldsymbol{x}_k\} && (17)\\
{}^{c}Y_k &= \mathrm{FFT}_{2N}\{[\mathbf{0}_N\ {}^{c}\boldsymbol{y}_k\} && (18)\\
P_k &= {}^{1}X_k \odot {}^{1}X_k^* + {}^{2}X_k \odot {}^{2}X_k^* && (19)\\
{}^{1}\hat{Y}_k &= {}^{1}X_k \odot {}^{11}H_k + {}^{2}X_k \odot {}^{21}H_k && (20)\\
{}^{2}\hat{Y}_k &= {}^{1}X_k \odot {}^{12}H_k + {}^{2}X_k \odot {}^{22}H_k && (21)\\
{}^{c}E_k &= {}^{c}Y_k - {}^{c}\hat{Y}_k && (22)\\
{}^{c}\boldsymbol{e}_k &= \mathrm{iFFT}_{2N}\{\,{}^{c}E_k\} && (23)\\
{}^{c}E_k &= \mathrm{FFT}_{2N}\{[\mathbf{0}_N\ {}^{c}\boldsymbol{e}_k(N\ldots 2N-1)]\} && (24)\\
{}^{ab}H_{k+1} &= {}^{ab}H_k + \alpha_k\,{}^{b}E_k \odot {}^{a}X_k^* \oslash P_k && (25)
\end{aligned}$$

$P_k$ are the binwise powers of the input signal (the sum is used for normalization as proposed in [6]). $\oslash$ denotes binwise division. This algorithm is almost identical to the one presented in [7].

### 2.7. Performance of the Stereo FLMS Algorithm

The stereo FLMS algorithm shown above works well for uncorrelated input signals, as expected and confirmed by simulations.

For correlated input signals, however, further measures must be taken to achieve proper convergence and to avoid results like the one seen in the upper curve in Fig. 2.

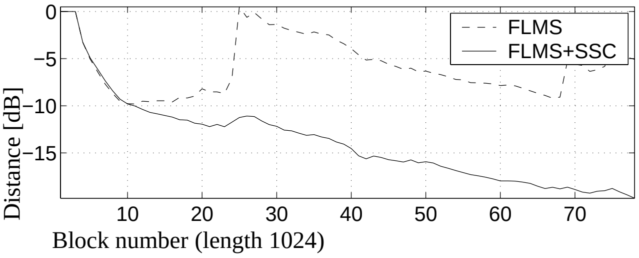

**Fig. 2**. System distance (input: Recorded stereo signal, solo singer, no room changes, $f_s = 16000$ Hz)

## 3. ADAPTIVE STEP-SIZE CONTROL

As pointed out by Gänsler et al. [8], the degree of correlation is directly related to the misalignment of the adaptive filter. While their approach adapts the factor of the used nonlinearity to achieve a constant level of correlation, we propose to use the measured correlation to adapt the step size, thus slowing down or halting the adaption in case of high correlation which might lead to misalignment.

This is where the factor $\alpha$ used in Eq. 25 comes into play. When making $\alpha$ small during blocks showing a high degree of correlation between the channels, misalignment due to correlation should decrease.

The correlation coefficient in block $\boldsymbol{x}_k$ is defined as:

$$\gamma_x = \frac{{}^{1}\boldsymbol{x}_k^T\,{}^{2}\boldsymbol{x}_k}{\sqrt{\left({}^{1}\boldsymbol{x}_k^T\,{}^{1}\boldsymbol{x}_k\right)\left({}^{2}\boldsymbol{x}_k^T\,{}^{2}\boldsymbol{x}_k\right)}} \qquad (26)$$

Since $\gamma_x$ is in $[-1, 1]$ and $\alpha_k$ must be in $[0, 1]$ to ensure stability, a reasonable choice is:

$$\alpha_k = 1 - \gamma_x^2 \qquad (27)$$

When using this step-size factor, the misalignment problem improves heavily; most of the erroneous adaption exhibited in the upper curve of Fig. 2 around block number 25 disappears without any modification of the signal. We now obtain the convergence behavior of the lower curve in Fig. 2.

Without any modification of the input signal and without complex processing, we achieve an adaption improvement of at least 10 dB.

## 4. THE DELAY-AND-ADD (DNA) ALGORITHM

### 4.1. The Algorithm

Very early in the development of SAEC algorithms, Shimauchi and Makino [9] made use of the fact that the impulse responses of the transmission paths from the speaker to the microphones in the transmission room (denoted as ${}^{1}g$ and ${}^{2}g$ in Fig. 1) are time-variant. This will hold for all cases in which a real speaker is recorded in stereo.

If the impulse responses in the receiving room are almost constant, we can exploit the fact that the problem is linear:

We define

$$\begin{aligned}
{}^{c}\xi^{M}(k) &= {}^{c}x(k) + {}^{c}x(k-M) && (28)\\
{}^{c}\upsilon^{M}(k) &= {}^{c}y(k) + {}^{c}y(k-M) && (29)
\end{aligned}$$

If $\boldsymbol{h}_k \approx \boldsymbol{h}_{k-M}$, we can write:

$${}^{c}\upsilon^{M}(k) \approx \sum_{a=1}^{2}\sum_{l=0}^{N-1} {}^{a}\xi^{M}(k-l)\cdot {}^{ac}h_k(l) \qquad (30)$$

Both $\xi$ and $\upsilon$ are easily calculated from given data without any need to change the transmitted signal.

The non-uniqueness pointed out in section 2.1 is now reduced, provided that there is at least some slight change in the impulse responses of the transmission room, ${}^{1}g(l)$ and ${}^{2}g(l)$.

This approach will of course produce horribly wrong results if applied while a room change in the receiving room takes place. If this happens, the assumption ${}^{c}\boldsymbol{h}_k \approx^{c} \boldsymbol{h}_{k-M}$ no longer holds true and the misalignment becomes larger than before.

To avoid this, we propose a shadow-filter approach. The DNA algorithm is run in the background; if the ERLE of the DNA filter is significantly better than the one of the normal filter, its coefficients are copied to the normal filter.

### 4.2. Experimental Results

If applied to the signal used for the adaptive step-size control experiments (a solo singer), the result shows to be little different from the one obtained without DNA, although the convergence curve looks more stable (see Fig. 3).

Using a multi-source input (choir music), the DNA shows its advantages, however. It allows to use a rather small nonlinearity factor (0.1) and still to arrive at the results that the conventional FLMS algorithm only reaches using a factor twice as large (0.2), causing much greater annoyance and distortion. This is illustrated in Fig. 4.

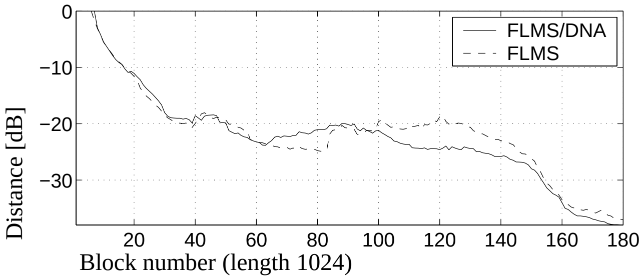

**Fig. 3**. System distance (setup identical to Fig. 2). Comparison of conventional FLMS and FLMS/DNA. $M = 3000$.

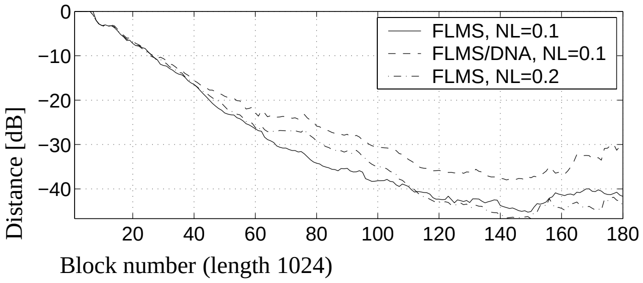

**Fig. 4**. System distance (Choir music). Comparison of conventional FLMS and FLMS/DNA. $M = 3000$, nonlinearities applied.

## 5. CONCLUSIONS

We have presented the stereo echo canceling problem in detail and shown two novel approaches to solve the non-uniqueness problem. Both of them try to avoid or at least minimize any audible pre-processing. The first method simply slows down adaption whenever misalignment becomes probable. The second modification attempts to make use of changes in the transmission room by using old and current signals simultaneously.

The next steps will definitely be to look deeper into the optimal step size in case of a given correlation factor and to improve the DNA algorithm such that it shows an improvement for all types of signals.

## 6. REFERENCES

[1] M. Ali, "Stereophonic acoustic echo cancellation system using time-varying all-pass filtering for signal decorrelation," in *ICASSP-98*, 1998, vol. 6, pp. 3689–3692.

[2] J. Benesty, D. Morgan, and M. Sondhi, "A better understanding and an improved solution to the problems of stereophonic echo cancellation," in *ICASSP-97*, 1997, vol. 1, pp. 299–302.

[3] M. Sondhi, D. Morgan, and J. Hall, "Sterophonic acoustic echo cancellation - an overview of the fundamental problem," *IEEE Signal Processing Letters*, vol. 2, no. 8, August 1995.

[4] Simon Haykin, *Adaptive Filter Theory*, Prentice Hall, Upper Saddle River NJ, 1996.

[5] George Moschytz and Markus Hofbauer, *Adaptive Filter*, Springer, Berlin, 2000.

[6] S. Shimauchi, Y. Haneda, S. Makino, and Y. Kaneda, "New configuration for a stereo echo canceller with nonlinear pre-processing," in *ICASSP-98*, 1998, vol. 6, pp. 3685–3688.

[7] J. Benesty and D. Morgan, "Frequency-domain adaptive filtering revisited, generalization to the multi-channel case, and application to acoustic echo cancellation," in *ICASSP-00*, 2000, vol. 2, pp. 789–792.

[8] T. Gänsler and J. Benesty, "New insights into the stereophonic acoustic echo cancellation problem and an adaptive nonlinearity solution," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 257–267, 2002.

[9] S. Shimauchi and S. Makino, "Stereo projection echo canceller with true echo path estimation," in *ICASSP-95*, 1995, vol. 5, pp. 3059–3062.