ENHANCED DOUBLE-TALK DETECTION BASED ON PSEUDO-COHERENCE IN STEREO

Markus Kallinger and Alfred Mertins

University of Oldenburg
Institute of Physics, Signal Processing Group
26111 Oldenburg, Germany

Karl-Dirk Kammeyer

University of Bremen
Dept. of Communications Engineering
28359 Bremen, Germany

ABSTRACT

This paper aims at double-talk detection for acoustic echo cancellers enabling transmission of speech signals in stereo. Double-talk detectors are affected by the same known non-uniqueness problem as echo cancellers. However, echo cancellers are usually adapted using NLMS-like iterative algorithms. Double-talk detectors being based on pseudo-coherence employ traditional spectral estimation techniques, which involve the use of temporal windows. These windows cause a bias of the coherence between the loudspeaker channels: this bias causes the magnitude squared coherence to be smaller than one. Therefore, the non-uniqueness problem known for echo cancellers does not exist for the mentioned type of double-talk detectors. However, the correlation between the loudspeaker channels provoke a strong bias of the inherent estimation of the echo paths. Consequently, we propose a method to decrease the influence of inter-channel correlations on the reliability of the double-talk detector.

1. INTRODUCTION

Depending on the system orders of the stereo acoustic echo canceller (AEC) and each mouth-room-microphone impulse response (MRMIR) the solution for the stereo AEC may not be unique [1]. However, due to exponentially decreasing but arbitrarily long room impulse responses, a solution for a stereo AEC usually is unique in most realistic scenarios. Unfortunately, it is strongly biased compared to the desired identification of the echo paths in the listening room. The amount of bias depends on the magnitude squared coherence (MSC) between the loudspeaker signals [2]: the closer the MSC gets to one, the stronger is the bias.

A reliable double-talk detector (DTD) based on pseudo-coherence inherently estimates the echo paths, too [3]. However, adaptation is not carried out by an iterative algorithm like the normalized least mean squares (NLMS) algorithm. With this kind of DTD the systems are estimated directly using the Wiener-Hopf equation. For the estimation of the corresponding power and cross-power spectral densities, traditional techniques with temporal windows are used. In Section 2 we show that, when using temporal windows, stereo echo path identification always is unique. Section 3 describes our proposal to enhance the DTD-scheme given in [3]. Simulation results are presented in Section 4 and conclusions in Section 5 finalize the paper.

2. STEREO DOUBLE-TALK DETECTION USING TRADITIONAL SPECTRAL ESTIMATION

In this section we extend the known stereo problem of [1] to the case, where temporal windows for directly estimating power and cross-power spectral densities come into operation. The following investigations are carried out in the z-domain. The basic setup is shown in Fig. 1. Here, only the case with one microphone in the receiving room is examined. However, another microphone for a second transmission channel does not involve additional problems. In the z-domain the error signal is given by

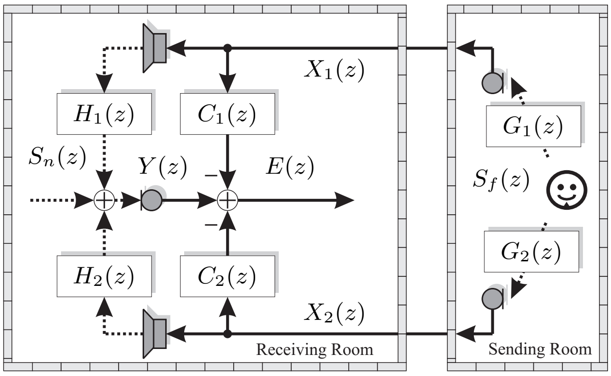

Fig. 1. Block diagram of a stereo acoustic echo canceller.

$$E(z) = Y(z) - \left(\begin{bmatrix} C_1(z) & C_2(z) \end{bmatrix} \begin{bmatrix} X_1(z) & X_2(z) \end{bmatrix}^T \right)$$
$$= Y(z) - \mathbf{C}^T(z)\mathbf{X}(z). \quad (1)$$

To calculate the echo canceller coefficients $C_1(z)$ and $C_2(z)$ in the MMSE sense, the signal power $\mathrm{E}\{|E(z)|^2\}$ has to be minimized. Setting

$$\frac{\partial \mathrm{E}\{|E|^2\}}{\partial \mathbf{C}} = -2\mathrm{E}\{\mathbf{X}^* Y\} + 2\mathrm{E}\{\mathbf{X}^*\mathbf{X}^T\mathbf{C}\} \tag{2}$$

to zero delivers the minimum mean square error (MMSE) solution for $\mathbf{C}(z)$:

$$\begin{aligned}\mathbf{C}(z) &= \mathrm{E}\{\mathbf{X}^*(z)\mathbf{X}^T(z)\}^{-1}\,\mathrm{E}\{\mathbf{X}^*(z)Y(z)\}\\ &= \mathbf{R}^{-1}_{XX,\mathrm{st}}(z)\mathbf{\Phi}_{\mathbf{X}Y}(z).\end{aligned} \tag{3}$$

Theoretically, there is no unique solution to this stereo Wiener-Hopf equation. With the help of the power and cross-power spectral densities

$$\Phi_{X_1X_1}(z) = \mathrm{E}\{X_1^*(z)X_1(z)\}\,, \tag{4}$$

$$\Phi_{X_2X_2}(z) = \mathrm{E}\{X_2^*(z)X_2(z)\}\,, \tag{5}$$

and

$$\Phi_{X_1X_2}(z) = \mathrm{E}\{X_1^*(z)X_2(z)\} \tag{6}$$

the stereo correlation matrix is expressed as

$$\mathbf{R}_{XX,\mathrm{st}}(z) = \begin{bmatrix}\Phi_{X_1X_1}(z) & \Phi_{X_1X_2}(z)\\ \Phi_{X_2X_1}(z) & \Phi_{X_2X_2}(z)\end{bmatrix}. \tag{7}$$

The rank of this matrix is one, which can be seen if we replace the cross-power spectral densities using the complex coherence

$$\Gamma_{X_1X_2}(z) = \frac{\Phi_{X_1X_2}(z)}{\sqrt{\Phi_{X_1X_1}(z)\Phi_{X_2X_2}(z)}}. \tag{8}$$

and work out the determinant of $\mathbf{R}_{XX,\mathrm{st}}(z)$, exploiting the fact that the MSC given by $|\Gamma_{X_1X_2}(z)|^2$ equals one. This is true, because $X_1(z)$ and $X_2(z)$ are linearly dependent – they are both generated from $S_f(z)$.

Equation (3) shows a theoretical solution for an arbitrarily long stereo echo canceller that is adapted by a stochastic gradient algorithm. These algorithms do not involve temporal windows, explicitly. However, for DTD based on pseudo-coherence we have to estimate spectral densities as shown in equation (20). This is usually carried out by applying temporal windows on the concerning signals.

For a closer inspection, we take into account the effects of windowing during an estimation method. As an example we illustrate the impact of observation windows on the bias of a true cross-power spectral density (CPSD):

$$\begin{aligned}&\mathrm{E}\{\hat{\Phi}^{\mathrm{Welch}}_{X_1X_2}(z)\}\\ &= \frac{1}{B}\sum_{\kappa=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}\mathrm{E}\{x_1[k]x_2[k+\kappa]\}\,w[k]w[k+\kappa]z^{-\kappa}\\ &= \frac{1}{B}\sum_{\kappa=-\infty}^{\infty} r_{x_1x_2}[\kappa]z^{-\kappa}\sum_{k=-\infty}^{\infty}w[k]w[k+\kappa] &(9)\\ &= \frac{1}{B}\sum_{\kappa=-\infty}^{\infty} r_{x_1x_2}[\kappa]r^E_{ww}[\kappa]z^{-\kappa} &(10)\\ &= \frac{1}{B2\pi j}\oint_T \mathrm{E}\{X_1^*(\nu)X_2(\nu)\}\,\Phi^E_{ww}(\frac{z}{\nu})\nu^{-1}d\nu &(11)\\ &= \frac{1}{B2\pi j}\oint_T \mathrm{E}\{|S_f(\nu)|^2\}\,G_1^*(\nu)G_2(\nu)\Phi^E_{ww}(\frac{z}{\nu})\nu^{-1}d\nu.\end{aligned}$$

B is a positive scalar value which compensates the bias introduced by the window $w[k]$ and the succeeding summation. $r^E_{ww}[\kappa]$ denotes the energy auto-correlation function (ACF) of $w[k]$. Finally, we can formulate the impact of the Welch estimation method onto the MSC

$$\begin{aligned}&\mathrm{E}\{|\hat{\Gamma}^{\mathrm{Welch}}_{X_1X_2}(z)|^2\}\\ &= \left|\oint_T \mathrm{E}\{|S_f(\nu)|^2\}\,G_1^*(\nu)G_2(\nu)\Phi^E_{ww}(\frac{z}{\nu})\nu^{-1}d\nu\right|^2\\ &\left(\oint_T \mathrm{E}\{|S_f(\nu)|^2\}\,|G_1(\nu)|^2\Phi^E_{ww}(\frac{z}{\nu})\nu^{-1}d\nu\right.\\ &\left.\oint_T \mathrm{E}\{|S_f(\nu)|^2\}\,|G_2(\nu)|^2\Phi^E_{ww}(\frac{z}{\nu})\nu^{-1}d\nu\right)^{-1}.\end{aligned} \tag{12}$$

For further examinations we carry out the following substitution:

$$\Theta(\nu,z) = \mathrm{E}\{|S_f(\nu)|^2\}\,\Phi^E_{ww}(\frac{z}{\nu})\nu^{-1}. \tag{13}$$

The MSC now amounts to

$$\mathrm{E}\{|\hat{\Gamma}^{\mathrm{Welch}}_{X_1X_2}(z)|^2\} = \frac{\left|\oint_T G_1^*(\nu)G_2(\nu)\Theta(\nu,z)d\nu\right|^2}{\left(\oint_T |G_1(\nu)|^2\Theta(\nu,z)d\nu\right)\left(\oint_T |G_2(\nu)|^2\Theta(\nu,z)d\nu\right)}. \tag{14}$$

Since this expression is smaller or equal to one, we have the inequality

$$\begin{aligned}&\left|\oint_T G_1^*(\nu)G_2(\nu)\Theta(\nu,z)d\nu\right|^2 \le\\ &\left(\oint_T |G_1(\nu)|^2\Theta(\nu,z)d\nu\right)\left(\oint_T |G_2(\nu)|^2\Theta(\nu,z)d\nu\right),\end{aligned} \tag{15}$$

which is a generalized form of the Cauchy-Schwarz inequality with $\Theta(\nu, z)$ being positive, real-valued, and possessing a limited number of nulls in T [4]. Identity is only achieved when $G_1(\nu)$ and $G_2(\nu)$ are linearly dependent, which can be expressed by

$$G_1(\nu) = \beta G_2(\nu) \quad \text{or} \quad G_2(\nu) = \beta' G_1(\nu), \tag{16}$$

where β and β' are scalars. The above equations are only fulfilled if the sending room transfer functions $G_1(z)$ and $G_2(z)$ are equal up to a frequency-independent factor. In general, this is not the case in a real stereo-transmission scenario.

Consequently, the solution for the stereo echo cancellers $C_1(z)$ and $C_2(z)$ is unique, provided that the windows incorporated into a traditional spectral estimation method are considered. However, the stereo ACF matrix $\mathbf{R}_{XX,\mathrm{st}}(z)$ still can be ill-conditioned at certain values of z. This results into a strong bias in terms of the actual room transfer functions $H_1(z)$ and $H_2(z)$.

This insight is of importance, each time the Wiener-Hopf equation is to be solved directly on the basis of estimated correlation functions. This is the case for DTD and residual echo estimation as proposed in [5]. However, it should be noted that this kind of uniqueness is not valid for echo cancellers which employ NLMS-like iterative algorithms.

3. ENHANCED DOUBLE-TALK DETECTION BASED ON PSEUDO-COHERENCE

Using our defined variables in the discrete frequency domain with the frequency index m and the temporal block index l, the pseudo-coherence according to [3] amounts to

$$\gamma_{\mathrm{p}}[l] = \frac{1}{L}\sum_{m=0}^{L-1} \frac{\hat{\mathbf{\Phi}}_{\mathbf{X}Y}^{H}[m,l]\hat{\mathbf{R}}_{XX,\mathrm{st}}^{-1}[m,l]\hat{\mathbf{\Phi}}_{\mathbf{X}Y}[m,l]}{\hat{\sigma}_y^2[l]} \tag{17}$$

with

$$\hat{\sigma}_y^2[l] = \frac{1}{L}\sum_{m=0}^{L-1} \hat{\Phi}_{YY}[m,l]. \tag{18}$$

L denotes the length of the employed DFT. All spectral densities are estimated using first-order recursive low pass filters, e. g.

$$\hat{\Phi}_{YY}[m,l] = \alpha\hat{\Phi}_{YY}[m,l{-}1]{+}(1{-}\alpha)\left|Y[m,l]\right|^2. \tag{19}$$

The matrix $\hat{\mathbf{R}}_{XX,\mathrm{st}}[m,l]$ becomes singular if the sending room impulse responses are too short [1]. Thus, additional measures would have to be taken to calculate $\gamma_{\mathrm{p}}[l]$. However, according to insights from Section 2 we know that a Welch estimate of $\mathbf{R}_{XX,\mathrm{st}}[m,l]$ has full rank and is invertible.

Simulations have shown that this stereo ACF can be very ill-conditioned at certain frequencies. Therefore, we propose to assign those frequencies low weight while assigning high weight to sub-bands with a low MSC. The spectrally weighted formulation of the pseudo-coherence

$$\gamma_{\mathrm{wp}}[l] = \frac{1}{L}\sum_{m=0}^{L-1} W[m,l]\frac{\hat{\mathbf{\Phi}}_{\mathbf{X}Y}^{H}\hat{\mathbf{R}}_{XX,\mathrm{st}}^{-1}\hat{\mathbf{\Phi}}_{\mathbf{X}Y}}{\hat{\sigma}_{y,\mathrm{w}}^2[l]} \tag{20}$$

and its normalization

$$\hat{\sigma}_{y,\mathrm{w}}^2[l] = \frac{1}{L}\sum_{m=0}^{L-1} W[m,l]\hat{\Phi}_{YY}[m,l] \tag{21}$$

should lead to increased robustness against modifications in the sending room, provided that the window $W[m,l]$ is appropriately designed.

A basis for a weighting rule can be found by investigating a formulation for the AEC coefficients with the help of equation (3)

$$\mathbf{C} = \frac{1}{\Phi_{X_1X_1}\Phi_{X_2X_2}(1-|\Gamma_{X_1X_2}|^2)}\mathrm{adj}\{\mathbf{R}_{XX,\mathrm{st}}\}\,\mathbf{\Phi}_{\mathbf{X}Y}$$

with

$$\mathrm{adj}\{\mathbf{R}_{XX,\mathrm{st}}(z)\} = \begin{bmatrix} \Phi_{X_2X_2}(z) & -\Phi_{X_1X_2}(z) \\ -\Phi_{X_2X_1}(z) & \Phi_{X_1X_1}(z) \end{bmatrix}.$$

When the MSC gets close to one, it plays a dominant role in the determinant, which runs close to zero. The window $W[m,l]$ can be assigned to the reciprocal of its appearance in the determinant to suppress the influence of deteriorated frequency bins:

$$W_{\mathrm{MSC}}[m,l] = 1 - \left|\hat{\Gamma}_{X_1X_2}[m,l]\right|^2. \tag{22}$$

Simulation results have shown that the square root of the above expression delivers good results, too:

$$W_{\mathrm{SRMSC}}[m,l] = \sqrt{1 - \left|\hat{\Gamma}_{X_1X_2}[m,l]\right|^2}. \tag{23}$$

4. SIMULATION RESULTS

Our simulation environment can be described using the following parameters: we used simulated room impulse responses at lengths of 4096 coefficients each and a reverberation time of 400 ms in the receiving room. They were generated using the well-known image method [6]. The sending room impulse responses were limited to a length of only 100 coefficients. At the same time, the length of each AEC-filter amounted to 1536. We used the partitioned frequency block least mean squares (PFBLMS) algorithm [7] for adaptation. The DFT-length L for calculating the pseudo-coherence was 1024. The constant α for smoothing was chosen to correspond to a "reverberation time" of 400 ms.

Fig. 2 shows the pseudo-coherence as a function of time. If no weighting window is applied, we can observe a small dip at sample 28,000, where the sending room impulse responses were changed. Using $W_{\mathrm{MSC}}[m,l]$ the pseudo-coherence decreases too much. With the weighting according to equation (23) we can hardly observe any influence of the stereo problem. This is in contrast to the behavior of the echo return loss enhancement (ERLE) measure, as will be shown below.

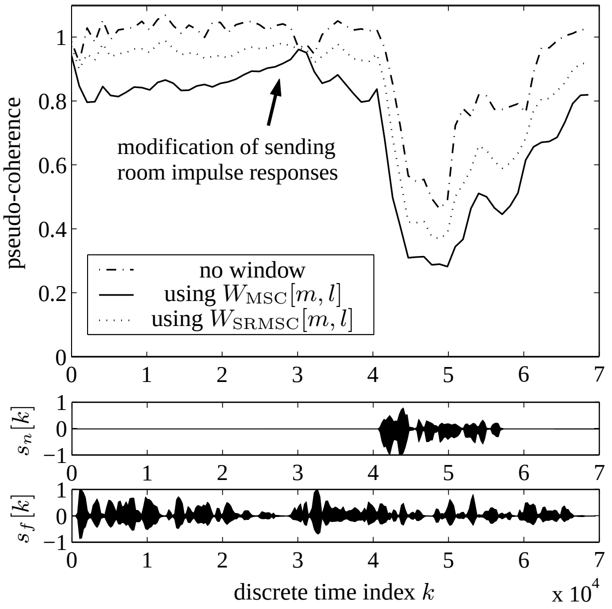

Fig. 2. Pseudo-coherence as a function of time.

Fig. 3 shows the ERLE as a function of time for the same near-end and far-end speech signals as in Fig. 2. At the point of modifying the sending room impulse responses (sample 28,000) we can observe a much more distinct dip than with the pseudo-coherence. Therefore, we can state that DTD based on traditional estimation methods using temporal windows on the signals is more robust against problems in stereophonic echo cancellation environments than the AEC itself.

5. CONCLUSIONS

In this contribution we have addressed the theory concerning stereo acoustic echo cancellation. A result is that the non-uniqueness problem does not exist when temporal windows come into operation. This is the case when power

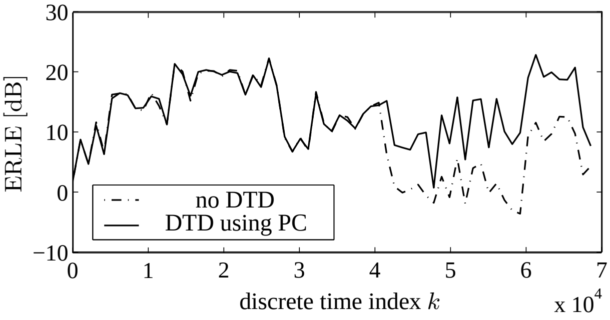

Fig. 3. Echo return loss enhancement as a function of time.

spectral densities are estimated using Welch's well-known method. Double-talk detection using pseudo-coherence is an application, which is based on traditional spectral estimation. Simulation results demonstrate the reduced impact of performance-degrading modifications of impulse responses in the sending room on double-talk detection compared to echo cancellation itself. Moreover, we have shown an effective way to increase the double-talk detection's robustness against modifications of impulse responses in the sending room.

6. REFERENCES

[1] 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. on Speech and Audio Processing*, vol. 6, no. 2, pp. 156–165, Mar 1998.

[2] T. Gänsler and J. Benesty, "New Insights into the Stereophonic Echo Cancellation Problem and an Adaptive Nonlinearity Solution," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 5, pp. 257–267, Jul 2002.

[3] T. Gänsler and J. Benesty, "A Frequency-Domain Double-Talk Detector Based on a Normalized Cross-Correlation Vector," *Signal Processing*, vol. 81, pp. 1783–1787, 2001.

[4] L. E. Franks, *Signal Theory*, Prentice Hall, London, 1969.

[5] M. Kallinger, J. Bitzer, and K. D. Kammeyer, "Post-Filtering for Stereo Acoustic Echo Cancellation," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Kyoto, Japan, Sep 2003.

[6] J. B. Allen and D. A. Berkley, "Image Method for Efficiently Simulating Small–Room Acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, 1979.

[7] J.-S. Soo and K. Pang, "Multidelay Block Frequency Domain Adaptive Filter," *IEEE Trans. on Acoustics Speech and Signal Processing*, vol. 38, pp. 373–376, Feb 1990.