

POST-FILTERING FOR STEREO ACOUSTIC ECHO CANCELLATION

Markus Kallinger, Karl-Dirk Kammeyer

University of Bremen, FB 1
Dept. of Communications Engineering
P.O. Box 330 440
D-28334 Bremen, Germany
kallinger@ant.uni-bremen.de

Jörg Bitzer

Houper Digital Audio
Anne-Conway-Str. 1
D-28359 Bremen, Germany
j.bitzer@hda.de

ABSTRACT

The transmission of a spatial acoustical impression is one aim of high-quality video conferencing systems. Therefore, the transmission of stereo speech signals is a major task of such a device. Compared to the mono-case, specific problems arise, if an acoustic echo canceller (AEC) is one component of the transmission device. Its echo attenuation in the receiving room degrades, if spatial statistics in the sending room vary. Simulation results indicate that short echo cancellers are less sensitive regarding this “stereo problem”. However, to achieve sufficient echo attenuation a post-filter should be applied. In this paper, we introduce a new method to design a post-filter for stereo acoustic echo cancellation. A vital task during the design procedure is the estimation of the residual echo’s power spectral density (PSD) at the output of the stereo acoustic echo canceller (Stereo AEC). This estimate is sensitive regarding the “stereo problem”, too. Therefore, we show up an effective way to increase its robustness and give a theoretical foundation. The result is a combination of a short Stereo AEC and a post-filter, which reveals less and shorter degradation of echo attenuation after spatial modifications in the sending room than a long Stereo AEC, which shows the same steady-state performance.

1. INTRODUCTION

Video conferencing systems, which incorporate the transmission of a stereo speech signal, require a signal processing unit to avoid the re-transmission of the far-end speaker’s acoustic echo. AECs represent the optimum solution to this problem in the system theoretical sense. However, specific problems arise in the stereo-case. The minimum mean squared error (MMSE) solution for the canceller’s filters regarding the error signal power at the Stereo AEC’s output is under-determined [1]. Benesty et al. have shown that this non-uniqueness does not necessarily occur in a realistic scenario. However, the MMSE solution still strongly depends on the mouth-to-microphone systems in the sending room. The echo attenuation instantaneously decreases as soon as another far-end speaker starts to talk.

It has been shown that this problem is related to the coherence between the loudspeaker channels. The MMSE solution for each loudspeaker becomes independent of the other, if the coherence vanishes [1]. In this paper we will show that the coherence can be reduced by means of short observation window lengths. In terms of echo cancellation, this suggests the operation of short cancellation filters, which involves lower steady-state echo attenuation.

Therefore, Gustafsson et al. suggested to run a post-filter at the output of a short Stereo AEC [2] to raise the steady-state performance of the combined system. In this paper, we introduce a novel method to estimate the residual echo’s PSD, which is the most important unknown value to design a post-filter. A theoretical motivation for our approach is given, as well. In the next section, we introduce our signal model. In section 3, we explain our method to estimate the residual echo. Simulation results are given in section 4 and section 5 concludes the paper.

2. DESIGN OF A STEREO POST-FILTER

Before we discuss the design procedure for a Wiener post-filter, we introduce a partitioned frequency domain signal model, upon which the presentation in the paper is based. Figure 1 illustrates the arrangement of all investigated sub-systems. The partitioned

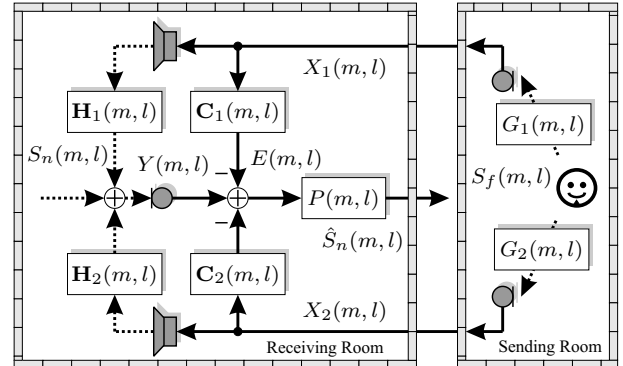


Fig. 1. Partitioned frequency domain signal model of a stereo acoustic echo canceller in front of a post-filter.

time- and frequency-discrete representation of the involved signals and systems has been chosen to be able to model the system orders of the room impulse response (RIR) $\mathbf{H}(m, l)$, the AEC $\mathbf{C}(m, l)$, and their difference, the system misalignment $\mathbf{D}(m, l)$. m denotes the discrete frequency index, l is the temporal block index. We introduce the following vectors

$$\mathbf{H}(m, l) = \begin{bmatrix} H_{1,0}(m, l) & \cdots & H_{1,L'_H-1}(m, l) \\ H_{2,0}(m, l) & \cdots & H_{2,L'_H-1}(m, l) \end{bmatrix}^T, \quad (1)$$

$$\mathbf{C}(m, l) = \begin{bmatrix} C_{1,0}(m, l) \cdots C_{1,L'_{\text{AEC}}-1}(m, l) 0 \cdots 0 \\ C_{2,0}(m, l) \cdots C_{2,L'_{\text{AEC}}-1}(m, l) 0 \cdots 0 \end{bmatrix}^T, \quad (2)$$

$$\mathbf{D}(m, l) = \mathbf{H}(m, l) - \mathbf{C}(m, l), \quad (3)$$

$$\mathbf{X}(m, l) = [\mathbf{X}_1^T(m, l) \quad \mathbf{X}_2^T(m, l)]^T, \quad (4)$$

$$= \begin{bmatrix} X_1(m, l) & \cdots & X_1(m, l - L'_H + 1) \\ X_2(m, l) & \cdots & X_2(m, l - L'_H + 1) \end{bmatrix}^T. \quad (5)$$

$L_H = L'_H L_{\text{DFT}}$ and $L_{\text{AEC}} = L'_{\text{AEC}} L_{\text{DFT}}$ are the lengths of the echo path impulse responses and the AEC filters, respectively. L_H is a length to model the RIRs in terms of system misalignment estimation as denoted in Figure 2. Actually, the orders of the RIRs are even higher. The residual echo at the output of the Stereo AEC results in

$$\Xi(m, l) = \mathbf{D}^T(m, l) \mathbf{X}(m, l). \quad (6)$$

If we estimate the system misalignment $\mathbf{D}(m, l)$, we can calculate $\Xi(m, l)$ and design a Wiener post-filter

$$\begin{aligned} P(m, l) &= \frac{\hat{\Phi}_{S_n S_n}(m, l)}{\hat{\Phi}_{S_n S_n}(m, l) + \hat{\Phi}_{\Xi \Xi}(m, l)} \\ &= \frac{\hat{\Phi}_{EE}(m, l) - \hat{\Phi}_{\Xi \Xi}(m, l)}{\hat{\Phi}_{EE}(m, l)}. \end{aligned} \quad (7)$$

The estimated PSDs in equation (7) can be gained by the well-known recursive Welch method. The most difficult task of the design procedure remains the estimation of $\mathbf{D}(m, l)$. At high system orders, e.g. if the video conferencing system runs in a reverberant environment, there are two ways to estimate the corresponding system misalignment impulse response $\mathbf{d}(k)$. We can use long DFT lengths L_{DFT} or we can investigate an increased number of partitions at short DFT lengths. Figure 2 illustrates this circumstance (an AEC operates at the first 512 samples).

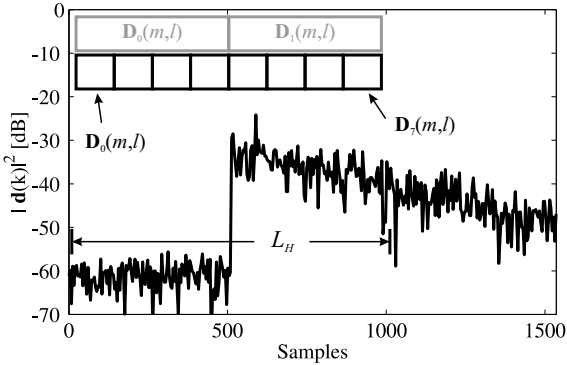


Fig. 2. Time domain illustration of a system misalignment impulse response. An AEC operates at the first 512 samples. The system misalignment can be estimated using long DFT lengths (grey) or an increased number of partitions (black).

3. CALCULATION OF THE SYSTEM MISALIGNMENT

In this chapter, we explain our method to estimate the system misalignment between the echo paths and the Stereo AEC filters. First,

we want to attain the stereo Wiener solution to be able to calculate $\mathbf{D}(m, l)$ by means of the signals $E(m, l)$, $X_1(m, l)$ and $X_2(m, l)$. The near-end speaker $S_n(m, l)$ is a part of the error signal $E(m, l)$ and acts as an interference during the estimation of $\mathbf{D}(m, l)$. Since this problem has already been addressed [3] and the stereo problem can be discussed independently, we assume that $S_n(m, l) = 0$.

$$\begin{aligned} \Xi(m, l) &= \mathbf{D}^T(m, l) \mathbf{X}(m, l) \\ \mathbf{X}^*(m, l) \Xi(m, l) &= (\mathbf{X}^*(m, l) \mathbf{X}^T(m, l)) \mathbf{D}(m, l), \\ E\{\mathbf{X}^*(m, l) \Xi(m, l)\} &= \mathbf{R}_{XX}(m, l) \mathbf{D}(m, l), \\ \mathbf{D}(m, l) &= \mathbf{R}_{XX}^{-1}(m, l) \Phi_{X\Xi}(m, l). \end{aligned} \quad (8)$$

$E\{\cdot\}$ is the expectation operator. A unique solution for this calculation only exists, if the correlation matrix $\mathbf{R}_{XX}(m, l)$ can be inverted. Thus, we take a closer look at its structure. $\mathbf{R}_{XX}(m, l)$ has the dimension $2L'_H \times 2L'_H$ and is assembled by the auto- and cross-correlation matrices of the signal vectors $\mathbf{X}_1(m, l)$ and $\mathbf{X}_2(m, l)$

$$\mathbf{R}_{XX}(m, l) = \begin{pmatrix} \mathbf{R}_{X_1 X_1}(m, l) & \mathbf{R}_{X_1 X_2}(m, l) \\ \mathbf{R}_{X_2 X_1}(m, l) & \mathbf{R}_{X_2 X_2}(m, l) \end{pmatrix}. \quad (9)$$

3.1. Coherence between the loudspeaker channels

In order to get a deeper insight into the behaviour of the stereo correlation matrix $\mathbf{R}_{XX}(m, l)$, we assume that both loudspeaker signals $X_1(m, l)$ and $X_2(m, l)$ are not correlated in the temporal direction:

$$E\{X_{1/2}^*(m, l - i) X_{1/2}(m, l - k)\} = 0, \quad \forall i \neq k. \quad (10)$$

The correlation matrix $\mathbf{R}_{XX}(m, l)$ results into the form

$$\mathbf{R}_{XX}(m, l) = \begin{pmatrix} \Phi_{X_1 X_1}(m, l) \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Phi_{X_1 X_1}(m, l - L'_H + 1) \\ \Phi_{X_2 X_1}(m, l) \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Phi_{X_2 X_1}(m, l - L'_H + 1) \\ \Phi_{X_1 X_2}(m, l) \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Phi_{X_1 X_2}(m, l - L'_H + 1) \\ \Phi_{X_2 X_2}(m, l) \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Phi_{X_2 X_2}(m, l - L'_H + 1) \end{pmatrix}. \quad (11)$$

Therefore, each partition of $\mathbf{D}(m, l)$ ($0 \leq i \leq L'_H - 1$) can be computed separately

$$\mathbf{D}_i(m, l) = \mathbf{R}_{i,XX}^{-1}(m, l) \Phi_{i,X\Xi}(m, l), \quad (12)$$

$$\mathbf{D}_i(m, l) = [D_{1,i}(m, l) \quad D_{2,i}(m, l)]^T, \quad (13)$$

$$\mathbf{R}_{i,XX}(m, l) = \begin{pmatrix} \Phi_{X_1 X_1}(m, l - i) & \Phi_{X_1 X_2}(m, l - i) \\ \Phi_{X_2 X_1}(m, l - i) & \Phi_{X_2 X_2}(m, l - i) \end{pmatrix}, \quad (14)$$

$$\Phi_{i,X\Xi}(m, l) = E\{[X_1^*(m, l - i) \Xi(m, l) \quad X_2^*(m, l - i) \Xi(m, l)]^T\}. \quad (15)$$

At each partition, we have to carry out a 2×2 matrix inversion.

To take a closer look at $\mathbf{R}_{i,XX}(m, l)$ we define the time- and frequency-discrete coherence

$$\Gamma_{X_1X_2}(m, l) = \frac{\Phi_{X_1X_2}(m, l)}{\sqrt{\Phi_{X_1X_1}(m, l)\Phi_{X_2X_2}(m, l)}}. \quad (16)$$

Therefore, the partial auto-correlation matrix according to equation (14) results in¹

$$\mathbf{R}_{i,XX} = \begin{pmatrix} \Phi_{X_1X_1} & \Gamma_{X_1X_2} \sqrt{\Phi_{X_1X_1}\Phi_{X_2X_2}} \\ \Gamma_{X_1X_2}^* \sqrt{\Phi_{X_1X_1}\Phi_{X_2X_2}} & \Phi_{X_2X_2} \end{pmatrix}. \quad (17)$$

It can easily be seen that this matrix is singular, if $|\Gamma_{X_1X_2}| = 1$, which is the case in a stereo setup. This holds true in a theoretical scenario, because the signals $X_1(m, l)$ and $X_2(m, l)$ result from the same source signal $S_f(m, l)$ via a linear convolution [4]. However, in real-world applications data windows of finite length come into operation, which involve a certain bias at the estimation of the spectral density functions. In equation (20) and in Figure 3, we illustrate that this bias causes the coherence's magnitude to decrease.

The windows for the estimation of each spectral density $w_1(k)$ and $w_2(k)$ are rectangular and their lengths are L_{DFT} and $2L_{\text{DFT}}$, respectively. Asymptotically bias free estimates of the spectral densities for L_{DFT} samples of the corresponding correlation functions are possible by this choice of windows (shown for auto correlation functions in [5]). A cross spectral density² can be estimated according to Welch

$$\begin{aligned} \mathbb{E} \left\{ \hat{\Phi}_{X_1X_2}^{\text{Welch}}(m) \right\} &= \\ &= \frac{1}{2L_{\text{DFT}}} \sum_{k=0}^{2L_{\text{DFT}}-1} \sum_{\kappa=-k}^{2L_{\text{DFT}}-1-k} \mathbb{E} \left\{ x_1(k)x_2(k+\kappa) \right\} \\ &= \frac{1}{2L_{\text{DFT}}} \sum_{\nu=0}^{2L_{\text{DFT}}-1} \Phi_{X_1X_2}(\nu) \Phi_{W_1W_2}^E(m-\nu). \end{aligned} \quad (18)$$

$$\quad (19)$$

Auto spectral densities are calculated with exactly the same windows accordingly. $\Phi_{W_1W_2}^E(m)$ is the energy spectral density of the window functions. The frequency-discrete spectra $X_1(m, l)$ and $X_2(m, l)$ are calculated by means of $S_f(m, l)$ and the transfer functions $G_1(m, l)$ and $G_2(m, l)$ as denoted in Figure 1. The Welch estimation of the magnitude squared coherence (MSC) results in

$$\begin{aligned} \mathbb{E} \left\{ \left| \hat{\Gamma}_{X_1X_2}^{\text{Welch}}(m) \right|^2 \right\} &= \\ &= \frac{\mathbb{E} \left\{ \left| \hat{\Phi}_{X_1X_2}^{\text{Welch}}(m) \right|^2 \right\}}{\mathbb{E} \left\{ \hat{\Phi}_{X_1X_1}^{\text{Welch}}(m) \right\} \mathbb{E} \left\{ \hat{\Phi}_{X_2X_2}^{\text{Welch}}(m) \right\}} \\ &= \frac{\left| \sum_{\nu} (G_1^*(\nu)G_2(\nu)) \Phi_{W_1W_2}^E(m-\nu) \right|^2}{\sum_{\nu} |G_1(\nu)|^2 \Phi_{W_1W_2}^E(m-\nu) \sum_{\nu} |G_2(\nu)|^2 \Phi_{W_1W_2}^E(m-\nu)}. \end{aligned} \quad (20)$$

$\Phi_{S_fS_f}(m)$ has already been cancelled. Without the observation windows $w_1(k)$ and $w_2(k)$ the MSC would be one. The simulation results shown in Figure 3 confirm our expectations. The coherence decreases, when shorter windows come into operation.

¹To keep this equation readable, the argument (m, l) is omitted, here.

²We have assumed stationary signals $x_1(k)$ and $x_2(k)$. Thus, the block index l can be omitted, here.

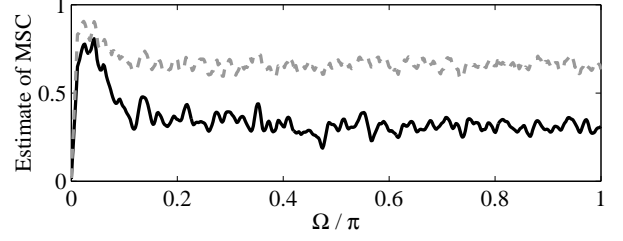


Fig. 3. Estimated coherence as a function of the frequency Ω using exemplary transfer functions $G_1(m, l)$ and $G_2(m, l)$ according to equation (20). At the grey, dashed coherence a window length of $L_{\text{DFT}} = 1024$ was used; at the black solid line a window length of 128.

3.2. Partitioned calculation of the system misalignment

So far, it seems very advantageous to use short observation windows with a raised number of partitions to estimate systems with long impulse responses. However, these estimates are biased by increased additive interferences. To illustrate this circumstance, we examine the case of two partitions for the estimation of $\mathbf{D}_m(m, l)$ in a mono setup.

$$\mathbf{D}_m(m, l) = [D_{m,0}(m, l) \ D_{m,1}(m, l)]^T \quad (21)$$

$$\mathbf{X}_m(m, l) = [X(m, l) \ X(m, l-1)]^T \quad (22)$$

$$\Xi(m, l) = \mathbf{D}_m^T(m, l)\mathbf{X}_m(m, l) \quad (23)$$

$$\hat{D}_{m,0}(m, l) = \frac{\mathbb{E} \{ X^*(m, l)\Xi(m, l) \}}{\mathbb{E} \{ |X(m, l)|^2 \}} \quad (24)$$

$$= D_{m,0}(m, l) + \frac{D_{m,1}(m, l)\mathbb{E} \{ X^*(m, l)X(m, l-1) \}}{\mathbb{E} \{ |X(m, l)|^2 \}} \quad (25)$$

The fraction in equation (25) acts as an additive interference on the estimate of $D_{m,0}(m, l)$. However, since the system misalignment does not vary too quickly, the estimates of its partitions $D_{m,i}(m, l)$ can be smoothed recursively. This measure reduces the influence by each interfering fraction during the estimation of $D_{m,0}(m, l)$ and $D_{m,1}(m, l)$. If we lower the DFT-length and raise the number of partitions to obtain a reduced MSC between the loudspeaker channels in the stereo case, we increase the number of interfering fractions at the same time. At very short window lengths ($L_{\text{DFT}} \leq 64$) the bias increases quickly. However, simulation results have shown that $L_{\text{DFT}} = 128$ at a sampling frequency $f_s = 8$ kHz is a good compromise between low coherence between the loudspeaker channels and hardly biased estimates of the residual echo.

4. SIMULATION RESULTS

All investigations have been carried out with simulated RIRs (generated using the well-known image method [6]) at a lengths of 4096 samples with a reverberation time of $\tau_{60} = 400$ ms.

In a first step, we want to examine the robustness of the estimated residual echo power against modifications in the sending room. Therefore, we have used white noise for the excitation signal $S_f(m, l)$. The sending room transfer functions $G_1(m, l)$ and $G_2(m, l)$ were modified at sample 40,000. The receiving room transfer functions $\mathbf{H}_1(m, l)$ and $\mathbf{H}_2(m, l)$ were modified at sample 16,000. The AECs $\mathbf{C}_1(m, l)$ and $\mathbf{C}_2(m, l)$ were omitted for

this study. We can observe that the estimated residual echo power using a window length of 512 decreases more drastically (see Figure 4, black, dashed-dotted line) than using a window size of 128 (black solid line). The expectations according to section 3.1 are fulfilled.

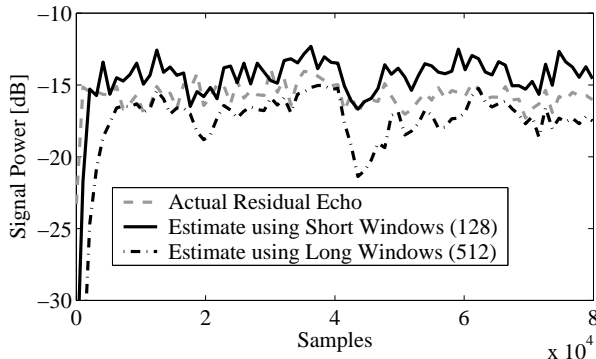


Fig. 4. Actual and estimated residual echo powers as a function of time.

Figure 5 shows the echo return loss enhancement (ERLE) as a function of time at three different setups with a white noise excitation $S_f(m, l)$ in the sending room. For the adaptation of the AEC, we have used a PFB LMS algorithm [7], which was extended for the stereo application. The two dashed curves illustrate its performance. When a different speaker position is switched on (sample 40,000) the AEC with 1536 coefficients at each filter loses more than 50% of its echo attenuation. The 512 coefficients AEC degrades at only 20%, which is a result of the lower MSC due to its shorter observation window. The post-filter (solid line) helps to raise the short AEC’s ERLE to that of the long one. The combined system with a Stereo AEC and a post-filter still suffers from the “stereo problem” but it does so to a clearly smaller extent. As in the mono case, a stereo post-filter converges much faster than an AEC.

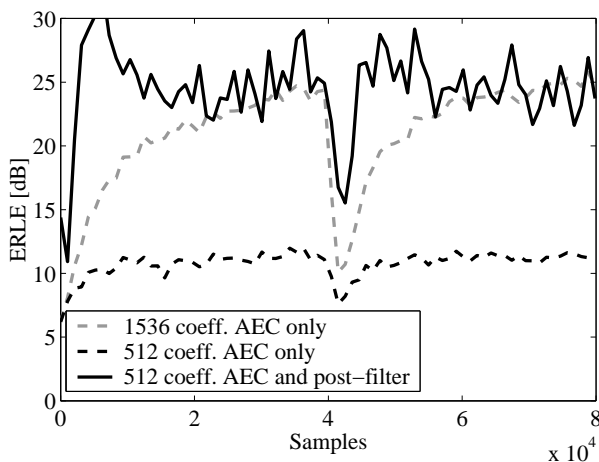


Fig. 5. ERLE as a function of time.

5. CONCLUSIONS

In this paper, we have investigated theoretical foundations for the application of a post-filter for Stereo AEC. As with echo cancellers, there is a non-uniqueness problem, which leads to failure of the echo suppression, when another speaker in the sending room starts to talk. However, we have shown a simple criterion for the design of a stereo post-filter, which makes it more robust against spatial modifications in the sending room. The robustness is based on the reduced coherence between the loudspeaker channels resulting from the bias, which is introduced by short data observation windows. In addition, a post-filter converges much faster than an AEC. Therefore, it represents a powerful extension to stereo acoustic echo cancellation.

ACKNOWLEDGEMENT

The author would like to thank Dr. K. U. Simmer for numerous helpful discussions!

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] S. Gustafsson and F. Schwarz, “A Postfilter for Improved Stereo Acoustic Echo Cancellation,” in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Pocono Manor, Pennsylvania, Sep 1999, pp. 32–35.
- [3] M. Kallinger, J. Bitzer, and K. D. Kammeyer, “Multi-Microphone Residual Echo Estimation,” in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Hong Kong, China, Apr 2003.
- [4] G.C. Carter, Knapp C. H., and A. H. Nuttall, “Estimation of the magnitude squared coherence function via overlapped fast fourier transform processing,” *IEEE Trans. on Audio and Electroacoustics*, vol. 21, no. 4, pp. 337–344, Aug 1983.
- [5] L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Englewood Cliffs, Prentice Hall, 1978.
- [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] D. W. E. Schobben, G. P. M. Egelmeers, and P. C. W. Sommen, “Efficient Realization of the Block Frequency Domain Adaptive Filter,” in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, April 1997, pp. 2257–2260.