UNBIASED RESIDUAL ECHO POWER ESTIMATION FOR HANDS-FREE TELEPHONY

Gerald Enzner, Rainer Martin, Peter Vary

Institute of Communication Systems and Data Processing
ind, Aachen University of Technology, D-52056 Aachen, Germany
Phone: +49-241-80-26960, E-mail: enzner@ind.rwth-aachen.de

ABSTRACT

Residual echo arises in hands–free telephony equipment due to insufficient echo canceler convergence, but can be suppressed using a postfilter. The most important control parameter for postfilter adaptation is therefore the residual echo power spectral density (PSD). In this contribution we present and compare residual echo PSD estimation techniques. We introduce a new partitioned block–adaptive estimator delivering unbiased residual echo PSD estimates in strongly reverberant and noisy acoustic environments.

1. INTRODUCTION

In the acoustic environment of mobile hands-free telephones we have to expect low signal–to–noise ratios and considerable acoustic feedback at the local microphone. Under these circumstances an echo canceler alone will not provide sufficient speech quality. Therefore, we apply a combined residual echo and noise reduction postfilter which is implemented in the Discrete Fourier Transform (DFT) domain [1, 2]. The postfilter is controlled by an estimate of the residual echo power spectral density.

The residual echo PSD, however, cannot be directly measured and must be estimated from the available signals. Conventional block oriented approaches [1, 2] with limited DFT length (due to delay and complexity constraints) can only reflect the residual echo within one DFT length of the residual echo impulse response, as illustrated in Figure 1. This leads to a serious bias of the residual echo estimate. Thus, we propose a new unbiased residual echo PSD estimator, based on coherence, which conceptually takes the full length of the residual echo system as well as short–term correlations into account. The idea behind the new approach is to compute the total residual echo PSD as a sum over multiple delayed DFT frames of short length. This leads to the concept of a partitioned residual echo power estimator.

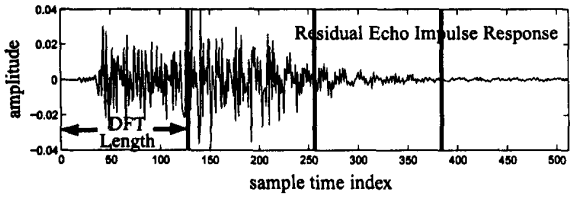

Fig. 1. Partitioned residual echo impulse response.

Figure 2 shows the echo and noise control system with a single loudspeaker and a single microphone. All signals are represented by their Fourier transform, e.g. the microphone signal by

$$Y(\Omega) = S(\Omega) + N(\Omega) + D(\Omega)\,, \tag{1}$$

where $S(\Omega)$, $N(\Omega)$, and $D(\Omega)$ represent clean near speech, background noise, and acoustic echo, respectively.

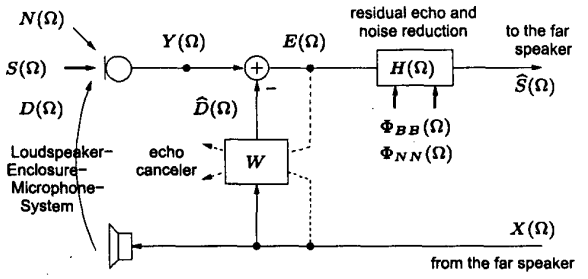

Fig. 2. Combined echo and noise reduction system for mobile hands-free telephony.

During adaptation, the echo canceler W yields a robust but possibly inaccurate estimate $\widehat{D}(\Omega)$ of the acoustic echo. The residual echo $B(\Omega) = D(\Omega) - \widehat{D}(\Omega)$ is then suppressed by the postfilter $H(\Omega)$ with input signal $E(\Omega) = S(\Omega) + N(\Omega) + B(\Omega)$. This is performed, for example, by the Wiener filter

$$H(\Omega) = H_W(\Omega) = \frac{\Phi_{SS}(\Omega)}{\Phi_{SS}(\Omega) + \Phi_{NN}(\Omega) + \Phi_{BB}(\Omega)} \tag{2}$$

relying on estimates of the background noise PSD $\Phi_{NN}(\Omega)$ and the residual echo PSD $\Phi_{BB}(\Omega)$. The background noise PSD can be determined adaptively and accurately by the Minimum Statistics approach [3], whereas the residual echo PSD is obtained from the new partitioned estimator. While the echo canceler converges, the responsibility for acoustic echo control is gradually taken away from the postfilter, in order to maintain the highest near speech quality.

The remainder of our paper is organized as follows: In Section 2 we will briefly review coherence analysis which was previously proposed for residual echo PSD estimation. Section 3 introduces the new partitioned residual echo estimation concept based on coherence, taking the full length of the residual echo system into account. In order to prove our algorithm, we provide a theoretical analysis of partitioned coherence estimation. Section 4 proposes a statistical correction of the coherence bias due to short–term correlations. Eventually, we will confirm our results by simulations in Section 5.

This work was supported by Nokia Research Center (NRC), Tampere, Finland, and Nokia Mobile Phones (NMP), Bochum, Germany.

2. COHERENCE BASED RESIDUAL ECHO ESTIMATION

The residual echo PSD estimator to be proposed in this paper as well as previously proposed algorithms make extensive use of the coherence function. We therefore briefly review coherence analysis with DFT based implementation.

2.1. Coherence Analysis

The spectrum of the residual echo is given by $B(\Omega) = G(\Omega)X(\Omega)$ where $G(\Omega)$ is the residual echo transfer function and $X(\Omega)$ is the received excitation signal. Assuming statistically independent $S(\Omega)$, $N(\Omega)$, and $X(\Omega)$, we can write

$$G(\Omega) = \frac{\Phi_{XE}(\Omega)}{\Phi_{XX}(\Omega)} \tag{3}$$

by the cross PSD $\Phi_{XE}(\Omega)$ of the signals $X(\Omega)$ and $E(\Omega)$. Consequently, we obtain $\Phi_{BB}(\Omega) = |G(\Omega)|^2 \Phi_{XX}(\Omega)$ for the residual echo PSD. This can be expressed equivalently [2] by

$$\Phi_{BB}(\Omega) = C_{XE}(\Omega)\Phi_{EE}(\Omega) \tag{4}$$

using the magnitude squared coherence function

$$C_{XE}(\Omega) = \frac{|\Phi_{XE}(\Omega)|^2}{\Phi_{XX}(\Omega)\Phi_{EE}(\Omega)} \tag{5}$$

of the signals $X(\Omega)$ and $E(\Omega)$.

2.2. DFT Based Implementation

The Discrete Fourier Transform $E(\Omega_\ell, kR)$ representing $E(\Omega)$ at frame index $k \in \mathbb{Z}$ is obtained from the time–domain signal $e(i)$ at sampling time index i using the window $w_e(i)$ as

$$E(\Omega_\ell, kR) = \sum_{i=0}^{M-1} e(kR + i - R)w_e(i - R)\mathrm{e}^{-j\Omega_\ell i} \tag{6}$$

with frame shift R and the normalized discrete frequency index $\Omega_\ell = 2\pi\ell/M$ for $\ell = 0, 1, \ldots, M-1$. The same notation holds for the DFT coefficients of any other signal under consideration.

Coherence based residual echo estimation, Equations (4) and (5), can now be implemented approximately on the basis of Welch's power spectral estimation technique [4], or by recursive averaging of periodograms which accounts for short–term stationarity of speech signals. The latter one is written with $0 < \alpha < 1$ as

$$\Phi_{XE}(\Omega_\ell, kR) = \alpha \cdot \Phi_{XE}(\Omega_\ell, (k-1)R) + (1-\alpha) \cdot I_{XE}(\Omega_\ell, kR) \tag{7}$$

using the (cross) periodogram

$$I_{XE}(\Omega_\ell, kR) = \frac{X^*(\Omega_\ell, kR)E(\Omega_\ell, kR)}{\sum_{i=0}^{M-1} w_x(i)w_e(i)} \tag{8}$$

between the DFTs $X(\Omega_\ell, kR)$ and $E(\Omega_\ell, kR)$.

The approach looks conceptually clear, however, in practice it delivers biased estimates of the residual echo PSD. This is due to insufficient coverage of the loudspeaker–enclosure–microphone system by the DFT length and the results is considerably underestimated residual echo. This is especially true for acoustic environments with large reverberation time and algorithms which use a relatively short echo canceler. Furthermore, short–term correlations of otherwise independent speech, echo, and background noise do always introduce an overestimation of the residual echo. Both effects will be taken into account by the bias–compensated partitioned residual echo (coherence) estimator to be developed in the following Sections.

3. PARTITIONED COHERENCE ESTIMATION

In order to take the full length of the residual echo system into account, while using block processing with limited DFT length, we propose the partitioned residual echo power estimation concept, based on coherence. This will be followed by a theoretic analysis of partitioned coherence estimation in order to validate the approach.

3.1. Partitioned Residual Echo Estimation

The residual acoustic echo in frame $E(\Omega_\ell, kR)$ is obviously correlated with the present and past frames $X(\Omega_\ell, (k-\lambda)R)$ of the excitation signal (corresponding to partitions of the residual echo system). With regard to the exponential decay of a causal residual echo impulse response, we may have to consider only a limited number L of most recent frames

$$X^{(\lambda)}(\Omega_\ell, kR) = X(\Omega_\ell, (k-\lambda)R), \quad 0 \le \lambda \le L-1\,. \tag{9}$$

A partial estimate of the residual echo PSD being due to the individual frame $X^{(\lambda)}(\Omega_\ell, kR)$ of length M is then written as

$$\Phi_{BB}^{(\lambda)}(\Omega_\ell, kR) = C_{X^{(\lambda)}E}(\Omega_\ell, kR)\Phi_{EE}(\Omega_\ell, kR) \tag{10}$$

according to (4). The estimator computes the total residual echo PSD by adding the contributions of several partitions λ

$$\Phi_{BB}(\Omega_\ell, kR) = \sum_{\lambda=0}^{L-1} \Phi_{BB}^{(\lambda)}(\Omega_\ell, kR) \tag{11}$$

where we assumed mutual statistical independence of the excitation frames $X(\Omega_\ell, kR)$. This is not exactly true in the case of speech excitation. Simulations, however, show that the approach can be successfully employed for frame-based acoustic echo suppression (if the DFT length is not extremely short).

Partitioned residual echo power estimation will be further justified by the analysis in the following Section.

3.2. Constraint And Unconstrained Partitioning

We will prove that the proposed algorithm (10) performs exact partitioning of the residual echo system in the case of white noise excitation when using the window functions

$$w_e(i) = \begin{cases} 1 & \text{for } 0 \le i \le M-R-1 \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

and $w_x(i) = w_e(i) + w_e(i+R)$ with $M = 2R$. We further assume that the impulse response $g(i)$ of the residual echo system can be modeled by a causal IIR filter and, thus, the output of the residual echo system is given by the linear convolution

$$e(i) = g(i) * x(i) = \sum_{m=0}^{\infty} g(m)x(i-m) \tag{13}$$

where $x(i)$ is a stationary far end excitation signal with power σ_{xx}^2.

It is straightforward to show that the cross power spectral density $\Phi_{XE}^{(\lambda)}(\Omega_\ell)$ of partition λ is obtained by statistical expectation from the cross periodogram $I_{X^{(\lambda)}E}(\Omega_\ell, kR)$ as

$$\Phi_{XE}^{(\lambda)}(\Omega_\ell) = E\{I_{X^{(\lambda)}E}(\Omega_\ell, kR)\} = \\ = \sum_{p=-(M-1)}^{M-1} \left(r_{xx}(p) * g(p+\lambda R)\right) r_{w_x w_e}(p) e^{-j\Omega_\ell p} , \quad (14)$$

using the auto-correlation $r_{xx}(p)$ of the excitation signal $x(i)$ and the normalized cross-correlation $r_{w_x w_e}(p)$ of the window functions $w_x(i)$ and $w_e(i)$

$$r_{w_x w_e}(p) = \frac{\sum_{i=0}^{M-1} w_x(i) w_e(i+p)}{\sum_{i=0}^{M-1} w_x(i) w_e(i)} . \quad (15)$$

For the windows under consideration and $M = 256$ the cross-correlation function $r_{w_x w_e}(p)$ is shown in Figure 3.

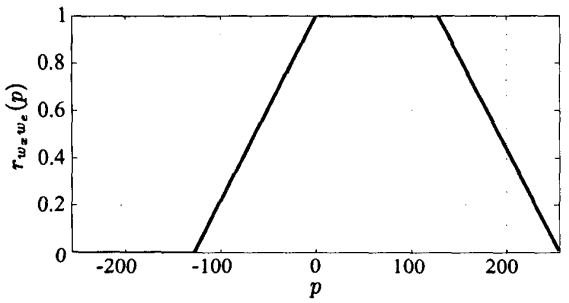

Fig. 3. Window cross-correlation function $r_{w_x w_e}(p)$. $M = 256$.

Since the extent of $r_{xx}(p)$ is much smaller than the extent of $r_{w_x w_e}(p)$ we may approximate

$$\left(r_{xx}(p) * g(p+\lambda R)\right) \cdot r_{w_x w_e}(p) \approx \\ \approx r_{xx}(p) * \left(g(p+\lambda R) \cdot r_{w_x w_e}(p)\right) . \quad (16)$$

Strict equality in (16) will hold for a white noise excitation $r_{xx}(p) = \delta(p)\sigma_{xx}^2$. With the above approximation we have

$$\Phi_{XE}^{(\lambda)}(\Omega_\ell) = \left(\Phi_{XX}(\Omega_\ell) G(\Omega_\ell) e^{j\Omega_\ell \lambda R}\right) * R_{w_x w_e}(\Omega_\ell) \approx \\ \approx \Phi_{XX}(\Omega_\ell) \left(G(\Omega_\ell) e^{j\Omega_\ell \lambda R} * R_{w_x w_e}(\Omega_\ell)\right) \quad (17)$$

where $G(\Omega_\ell)$ is the frequency response of the residual echo system and $R_{w_x w_e}(\Omega_\ell)$ is the cross-power spectrum of the windows.

Moreover, we find that the normalization of the cross power spectral density by $\Phi_{XX}(\Omega_\ell)$ removes the dependence of $\Phi_{XE}^{(\lambda)}(\Omega_\ell)$ on the input signal statistics. The additional constraint operation $DFT\{Q\{IDFT\{\cdot\}\}\}$ yields the spectrum

$$G^{(\lambda)}(\Omega_\ell) = DFT\{Q\{IDFT\left\{\frac{\Phi_{XE}^{(\lambda)}(\Omega_\ell)}{\Phi_{XX}(\Omega_\ell)}\right\}\}\} \quad (18)$$

of a rectangular partition $g(p+\lambda R), p = 0, \ldots, M/2-1, \lambda \in \mathbb{Z}$, of the residual echo system $g(i)$. This can be seen from

$$IDFT\left\{\frac{\Phi_{XE}^{(\lambda)}(\Omega_\ell)}{\Phi_{XX}(\Omega_\ell)}\right\} = g(p+\lambda R) r_{w_x w_e}(p) + \\ + g(p+\lambda R - M) r_{w_x w_e}(p-M), \quad p = 0 \ldots M-1 . \quad (19)$$

in conjuntion with the projection Q which zero–forces the samples for $p = M/2, \ldots, M-1$ and therefore leaves only the signal within the flat–top region of $r_{w_x w_e}(p)$ for further processing.

Replacing power spectral densities by their short–time estimates, the above procedure is ideally suited to compute the residual echo power estimate $\Phi_{BB}^{(\lambda)}(\Omega_\ell, kR)$ of partition λ at frame index k:

$$G^{(\lambda)}(\Omega_\ell, kR) = DFT\{Q\{IDFT\left\{\frac{\Phi_{XE}^{(\lambda)}(\Omega_\ell, kR)}{\Phi_{XX}(\Omega_\ell, kR)}\right\}\}\} \quad (20)$$

$$C_{X^{(\lambda)}E}(\Omega_\ell, kR) = \frac{|G^{(\lambda)}(\Omega_\ell, kR)|^2 \Phi_{XX}(\Omega_\ell, kR)}{\Phi_{EE}(\Omega_\ell, kR)} \quad (21)$$

The total residual echo PSD is eventually obtained from Equations (10) and (11), once more assuming white noise excitation.

The coherence estimator in Equation (5) may be viewed as an (approximate) unconstrained version of (20) and (21) with considerably lower complexity.

Another approximate partitioning of the residual echo system which in practice has been proven to be very accurate uses Hann windows $w_e(i) = w_x(i) = 0.5(1 - \cos(2\pi(i + M/2)/M))$, $-M/2 \le i \le M/2 - 1$, and unconstrained coherence estimation with 50% frame overlap.

4. COMPENSATION OF SHORT TERM CORRELATIONS

In general, the residual echo estimate (10) will be too high due to short–term correlations between long–term independent echo and background noise signals. An approximation for the biased coherence $\widehat{C}$ obtained from Equation (5), which holds for estimates on the basis of Welch's power spectral estimation technique, is given in [5] for stationary signals:

$$\widehat{C} \approx C + \frac{1}{N}(1-C)^2 \left(1 + \frac{2C}{N}\right) \triangleq f_C(C) \quad (22)$$

Thereby, C denotes the true coherence and N is the number of periodograms used for averaging over time. N is related to the equivalent forgetting factor α for recursive averaging by $N = (1+\alpha)/(1-\alpha)$.

The proposed mechanism for bias correction directly relies on the inversion of the above formula. Before, however, we have to decrease the variance of the coherence estimate in order to make the bias correction reliable. Therefore we average over several adjacent frequency components of the (cross) PSDs involved in the estimation of the coherence function (5) or (21). Typically this is done non–uniformly with larger subbands at high frequencies, thus avoiding noticeable performance degradation. At the same time, averaging decreases the frequency resolution at which the division in (5) or (21) is carried out and, consequently, performs favorable in terms of computational complexity. The modified coherence estimate is then denoted by $\widetilde{C}_{XE}(\Omega_\ell, kR)$.

Proceeding with the inversion of (22) for each estimator partition λ, e.g. by means of a look–up table, we write the bias corrected version of the partitioned residual echo estimator (11) as

$$\widetilde{\Phi}_{BB}(\Omega_\ell, kR) = \sum_{\lambda=0}^{L-1} f_C^{-1}\left(\widetilde{C}_{X^{(\lambda)}E}(\Omega_\ell, kR)\right)\Phi_{EE}(\Omega_\ell, kR)\,. \tag{23}$$

The algorithm delivers an unbiased residual echo estimate even in the presence of stationary local disturbances. Note that, strictly speaking, also the residual echo of partition $\lambda' \neq \lambda$ represents a kind of local disturbance for estimator partition λ. Eventually, we observe the freedom to assign individual forgetting factors $\alpha^{(\lambda)}$ to the estimation process of each partition, taking individual echo–to–noise ratios into account.

5. SIMULATION RESULTS

5.1. Log–Spectral–Mean

We rate the accuracy of residual echo PSD estimators by the *Log–Spectral–Mean*

$$LSM(kR) = \frac{1}{M}\sum_{l=0}^{M-1} 10\log_{10}\frac{\widehat{\Phi}_{BB}(\Omega_\ell, kR)}{\Phi_{BB}(\Omega_\ell, kR)} \tag{24}$$

of the estimated–to–true residual echo power ratio at frame index k. This is a frame–based bias measure for residual echo estimators, which is ideally zero.

5.2. Measurement of the Residual Echo PSD

Numerical results are compared for three different estimators: the conventional (single–partition) algorithm, Equation (4), the partitioned residual echo estimator, Equation (11), and its bias compensated version, Equation (23), all of them applying exact partitioning according to Equation (21).

For the simulation, we use a stationary white noise excitation $X(\Omega)$, various levels of local speech $S(\Omega)$, and car background noise $N(\Omega)$. The acoustic echo is generated by means of a fixed car impulse response of 512 coefficients, the first 128 coefficients being canceled nearly ideally by a fixed echo compensator with 128 taps. The DFT length for residual echo estimation is $M = 256$. With regard to the short–term stationarity of speech, we chose the forgetting factor $\alpha = 0.8$ for the single-partition estimator. For the partitioned estimator we use $L = 4$ single–partition coherence estimators running in parallel, where the corresponding forgetting factors were individually chosen as $\alpha^{(0)} = 0.8$, $\alpha^{(1)} = 0.8$, $\alpha^{(2)} = 0.9$, and $\alpha^{(3)} = 0.9$.

Figure 4 shows the results for three different acoustic environments: In the first 300 signal frames, no local speech nor background noise contributes to the microphone signal, thus, acoustic echo only. In frames 300 to 600 we added local car background noise at the echo–to–noise ratio of -6 dB. Eventually, in frames 600 to 900, we have simulated a double talk situation at the speech–to–noise ratio of 0 dB and car background noise at the same level as before.

From Figure 4 we observe that the conventional (single partition) estimator (4) does not completely reflect the residual echo. The bias of the estimator is most severe without local speech or noise. In the presence of local background noise and speech activity, the approach achieves better performance only because of the additional bias introduced by short–term correlations. We further

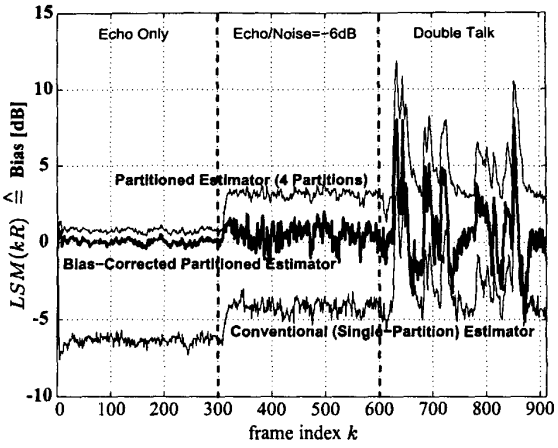

Fig. 4. Bias for constrained partitioned residual echo estimators.

observe that the partitioned residual echo estimator (11) achieves nearly unbiased residual echo estimates if there is neither speech nor background noise present. This is due to the full coverage of the residual echo impulse response of length 512 by $L = 4$ estimator partitions. However, we can see the bias of this method in the presence of local noise (due to short–term correlations). This is circumvented by the additional coherence bias correction applied to the partitioned residual echo estimator in (23). The latter algorithm delivers unbiased estimates with regard to any acoustic environment. Of course, the variance of the estimator still depends on the local echo–to–noise/speech ratio. Note, however, that in the presence of background noise the estimate is not required to be as accurate as in noise-free conditions. Hence, we conclude that the partitioned residual echo estimator with individual bias correction in each partition λ delivers consistently excellent results for the application of residual echo postfiltering.

6. CONCLUSIONS

We derived an unbiased residual echo PSD estimator by taking the full length of the residual echo impulse response as well as the bias due to short–term correlations into account. The estimator delivers unbiased results in all kinds of noisy and reverberant acoustic environments and is therefore ideally suited to control postfiltering for residual echo suppression in hands-free telephony.

7. REFERENCES

[1] S. Gustafsson, R. Martin, and P. Vary, "Combined acoustic echo control and noise reduction for hands-free telephony," *Signal Processing*, vol. 64, pp. 21–32, 1998.

[2] C. Beaugeant, *Réduction de Bruit et Contrôle d'Echo pour les Applications Radiomobiles*, Ph.D. thesis, University of Rennes 1, 1999.

[3] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 5, July 2001.

[4] S. Haykin, "Adaptive Filter Theory," Prentice Hall, 1996.

[5] G. C. Carter, "Coherence and time delay estimation," *Proceedings of the IEEE*, vol. 75, pp. 236–255, 1987.