ON SPECTRAL ESTIMATION OF RESIDUAL ECHO IN 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: +49-241-8026960, E-mail: enzner@ind.rwth-aachen.de

ABSTRACT

We present and compare residual echo power spectral estimation techniques. Residual echo arises in hands-free telephony equipment due to insufficient echo cancellation performance. The residual echo power spectral density is very important to control combined residual echo and noise postfiltering. Eventually, we introduce a new partitioned block-adaptive estimation technique delivering considerably improved estimates in reverberant and noisy double talk environments.

1. INTRODUCTION

Due to 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. It has been shown that a combined acoustic echo and noise reduction postfilter substantially improves the performance of the more traditional echo cancellation and noise reduction approach [1]. Postfiltering in turn relies on the residual echo and noise power spectral densities (PSD). The functionality of our full-duplex echo and noise control system is depicted in Figure 1. Since the postfilter is implemented in the frequency domain, all signals are represented by their short term Discrete Fourier Transform (DFT) coefficients at frame index k and normalized discrete frequency index Ω. We denote the DFT coefficients of the microphone signal by

$$Y(\Omega,k) = S(\Omega,k) + N(\Omega,k) + D(\Omega,k)\ , \qquad (1)$$

where $S(\Omega,k)$, $N(\Omega,k)$ and $D(\Omega,k)$ represent clean near speech, background noise, and acoustic echo, respectively. Because of its relatively short length, the echo canceller C yields a robust but insufficient estimate $\hat{D}(\Omega,k)$ of the acoustic echo. Therefore, we apply combined residual echo and noise postfiltering with input signal

$$\begin{aligned} E(\Omega,k) &= Y(\Omega,k) - \hat{D}(\Omega,k) \\ &= S(\Omega,k) + N(\Omega,k) + B(\Omega,k)\ , \qquad (2) \end{aligned}$$

where $B(\Omega,k) = D(\Omega,k) - \hat{D}(\Omega,k)$ is the residual echo signal. In the receiving and sending path of the telephone we have the far end speech $X(\Omega,k)$ and the estimated local speech $\hat{S}(\Omega,k)$, respectively.

2. PREVIOUSLY PROPOSED ESTIMATORS

Residual echo estimation techniques are based on the evaluation of auto and cross PSDs of the signals in Figure 1. For example, the cross PSD of $X(\Omega,k)$ and $E(\Omega,k)$ is written as $\Phi_{XE}(\Omega,k)$, auto PSDs accordingly.

At first, we recapitulate two previously proposed solutions to the estimation problem.

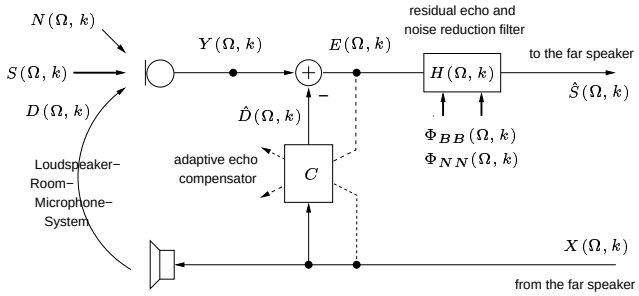

Figure 1: Mobile hands-free telephony environment.

2.1. Coherence Analysis

We assume $B(\Omega,k) = G(\Omega,k)X(\Omega,k)$, where $G(\Omega,k)$ is the residual echo transfer function. Under the assumption of statistically independent $S(\Omega,k)$, $N(\Omega,k)$, and $X(\Omega,k)$, we can write

$$G(\Omega,k) = \frac{\Phi_{XE}(\Omega,k)}{\Phi_{XX}(\Omega,k)} \qquad (3)$$

and obtain $\Phi_{BB}(\Omega,k) = |G(\Omega,k)|^2\,\Phi_{XX}(\Omega,k)$ for the residual echo PSD. This can be expressed equivalently [2] by

$$\Phi_{BB}(\Omega,k) = C_{XE}(\Omega,k)\Phi_{EE}(\Omega,k) \qquad (4)$$

using the magnitude squared coherence function

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

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

The result can be implemented on the basis of Welch's power spectral estimation technique [3], or recursive averaging of periodograms which accounts for the short term stationarity of speech signals ($0 < \alpha < 1$):

$$\hat{\Phi}_{XE}(\Omega,k) = \alpha\,\hat{\Phi}_{XE}(\Omega,k-1) + (1-\alpha)X^*(\Omega,k)E(\Omega,k) \qquad (6)$$

The approach is conceptually clear, however, we observed biased estimates of the residual echo PSD in the case of finite length block processing and due to short term correlations of otherwise independent speech and background noise signals. Therefore, the coherence method still has potentials for considerable improvements with regard to residual echo estimation. This will be shown more detailed in Section 3 and by simulations.

2.2. Virtual Transfer Analysis

According to [1], we assume $B(\Omega,k) \approx F(\Omega,k)D(\Omega,k)$, where $F(\Omega,k) \in \mathbb{R}$ is the real valued virtual transfer function between the acoustic echo and the residual echo. Then, we obtain

$$\Phi_{BB,Virt}(\Omega,k) = \left(\frac{F(\Omega,k)}{1-F(\Omega,k)}\right)^2 \Phi_{\hat{D}\hat{D}}(\Omega,k) \qquad (7)$$

with

$$F(\Omega,k)=\frac{\Phi_{YY}(\Omega,k)-\Phi_{EE}(\Omega,k)-\Phi_{\hat{D}\hat{D}}(\Omega,k)}{\Phi_{YY}(\Omega,k)-\Phi_{EE}(\Omega,k)+\Phi_{\hat{D}\hat{D}}(\Omega,k)}\,. \quad (8)$$

Note that it is not allowed to apply an averaging process directly to the virtual transfer function, neither as time averages nor in the frequency domain. The problem is that the estimate of the virtual transfer function is highly non–ergodic and non–stationary, in contrast to the residual echo PSD which it serves for.

Equations (7) and (8) rely on available auto PSDs only and, hence, the approach is supposed to work also during double talk. However, the above equations do also allow another interpretation if we consider the general formula for the PSD of a signal $Y(\Omega,k)=\hat{D}(\Omega,k)+E(\Omega,k)$:

$$\Phi_{YY}(\Omega,k)=\Phi_{EE}(\Omega,k)+\Phi_{\hat{D}\hat{D}}(\Omega,k)+2\mathrm{Re}\left\{\Phi_{\hat{D}E}(\Omega,k)\right\} \quad (9)$$

Equation (9) holds for estimates obtained from Welch's spectral estimation technique or recursive averaging of periodograms (6), too, since $|Y|=|\hat{D}|+|E|+2\mathrm{Re}\{\hat{D}^*E\}$. Substituting (9) into (7) and (8) delivers the equivalent expression

$$\Phi_{BB,Virt}(\Omega,k)=\frac{\mathrm{Re}^2\left\{\Phi_{\hat{D}E}(\Omega,k)\right\}}{\Phi_{\hat{D}\hat{D}}(\Omega,k)} \quad (10)$$

for the virtual transfer estimator. This could be viewed as an incomplete coherence analysis between $\hat{D}(\Omega,k)$ and $E(\Omega,k)$. In fact, we might even complement this virtual transfer estimator making use of the signal

$$Y'(\Omega,k)=j\cdot\hat{D}(\Omega,k)+E(\Omega,k)\,.$$

On the basis of $Y'(\Omega,k)$ and the general idea behind equation (9) we can compute

$$\mathrm{Im}\left\{\Phi_{\hat{D}E}(\Omega,k)\right\}=\frac{1}{2}\left(\Phi_{Y'Y'}(\Omega,k)-\Phi_{EE}(\Omega,k)-\Phi_{\hat{D}\hat{D}}(\Omega,k)\right) \quad (11)$$

and hereby obtain the extended virtual transfer estimate

$$\begin{aligned}\Phi'_{BB,Virt}(\Omega,k)&=\Phi_{BB,Virt}(\Omega,k)+\frac{\mathrm{Im}^2\left\{\Phi_{\hat{D}E}(\Omega,k)\right\}}{\Phi_{\hat{D}\hat{D}}(\Omega,k)}\\&=C_{\hat{D}E}(\Omega,k)\Phi_{EE}(\Omega,k)\,. \quad (12)\end{aligned}$$

With regard to this interpretation we conclude that the virtual transfer estimator in fact is an incomplete coherence estimator between $\hat{D}(\Omega,k)$ and $E(\Omega,k)$. In particular, we expect similar problems with respect to finite length block processing and local disturbances. This will be shown by simulations.

This section showed that we can proceed with the development of residual echo estimators on the basis of equations (4), (5), and (6), possibly using either $X(\Omega,k)$ or $\hat{D}(\Omega,k)$.

3. MULTIPLE–FRAME COHERENCE ANALYSIS

3.1. Problem Description

a) We have already stated in this article before, that we expect a biased coherence estimate (or residual echo estimate) in case of finite length block processing. This is due to the fact that the acoustic echo path may considerably spread the far speech over time before the local microphone picks it up. In car acoustics, for example, we actually have to account for an impulse response of 400 to 800 coefficients at a sampling frequency of 8kHz. In an office environment, the reverberation time can be significantly higher. Typical speech enhancement systems, in contrast, make use of DFT lengths of only 128 or 256 speech samples (due to signal delay and complexity constraints). Therefore, a DFT based block estimator on the basis of equations (4), (5), and (6) will certainly fail in the attempt to reflect the full correlation between the echo compensated signal and the far speech. Thus, one obtains systematically underestimated residual echo. There can be a very distinctive bias in those applications where early echo is already cancelled out by the short and robust echo compensator of length 128 to 256. In this case, the residual echo bias related with (4), (5), and (6) considerably impacts the postfiltering performance.

b) Both local background noise and near speech activity are also decisively influencing the residual echo estimator. Background noise and near speech mean local disturbances from the viewpoint of acoustic echo control. In general, the residual echo estimates will be too high due to short–term correlations between otherwise independent echo, speech and background noise signals. An approximation for the biased coherence estimate (5), which holds for calculations on the basis of Welch's power spectral estimation technique, is given in [4] for stationary signals:

$$\hat{C}\approx C+\frac{1}{N}(1-C)^2\left(1+\frac{2C}{N}\right)\triangleq f_C(C,N) \quad (13)$$

Thereby, N is the number of periodograms used for averaging over time and is related to an equivalent forgetting factor for recursive averaging by $N=(1+\alpha)/(1-\alpha)$. This result allows for the coherence bias correction

$$C\approx f_C^{-1}(\hat{C},N(\alpha)) \quad (14)$$

in the presence of stationary local disturbances.

3.2. Proposed Estimator

The rationale behind the newly proposed multiple–frame residual echo (or coherence) estimator is as follows:

The acoustic echo contained in frame $E(\Omega,k)$ was excited by the far speech frames $X(\Omega,\lambda),\lambda\le k$. With regard to the exponential decay of the room impulse response, we may have to consider only the limited number L of most recent frames $X(\Omega,\lambda),k-L+1\le\lambda\le k$. According to (4) and (14), the residual echo PSD corresponding to $E(\Omega,k)$ and being due to $X(\Omega,\lambda)\triangleq X^{(\lambda)},\lambda\le k$, is written as:

$$\Phi_{BB}^{(\lambda)}(\Omega,k)=f_C^{-1}\left(C_{X^{(\lambda)}E}(\Omega,k),N(\alpha)\right)\Phi_{EE}(\Omega,k) \quad (15)$$

Thereby, we assumed mutual statistical independence of the excitation frames $X(\Omega,k)$. This is approximately true also in the case of speech excitation and DFT lengths of 128 to 256 data points. Formula (15) accounts for stationary local disturbances by the bias correction (14). Note that, strictly speaking, even the acoustic echo due to excitation frames $X(\Omega,k),k\neq\lambda$, represents local disturbances for the estimation of $\Phi_{BB}^{(\lambda)}(\Omega,k)$.

Then, the total residual echo PSD can approximately be calculated as a superposition of the "single–frame" estimates (15):

$$\Phi_{BB,new}(\Omega,k)=\sum_{\lambda=k-L+1}^{k}\Phi_{BB}^{(\lambda)}(\Omega,k) \quad (16)$$

3.3. Related Benefits

The proposed structure for residual echo estimation entails a number of benefits which are briefly discussed as follows:

- Each coherence estimate $C_{X^{(\lambda)}E}(\Omega,k)$ considers an individual PSD $\Phi_{XX}^{(\lambda)}$ of the excitation signal. Thus, we make only weak assumptions with respect to the stationarity of the excitation. This is particularly meaningful for speech excitation in the presence of long reverberation times.
- The bias of each coherence estimate $C_{X^{(\lambda)}E}(\Omega,k)$ is removed individually by the bias correction formula (14).
- Eventually, we observe the freedom to assign individual forgetting factors $\alpha^{(\lambda)}$ to the estimation processes of the coherence

functions $C_{X(\lambda)E}(\Omega,k)$. That is useful, since a reasonable forgetting factor certainly depends on the individual ratio of acoustic echo and local disturbances.

3.4. Computational Complexity

The newly proposed algorithm basically runs the "single-frame" coherence estimator (4), (5), and (6) L times in parallel in order to deliver an unbiased residual echo estimate. Additionally, we also apply the bias correction (14) L times in parallel to cope with local disturbances. Thus, the complexity of the algorithm is roughly L times as high as for the conventional "single-frame" algorithm. In practice, it turns out that only $L = 3$ or $L = 4$ can considerably improve the residual echo estimate in the case of car acoustics. Furthermore, we might adjust the length of the echo canceller such that we can even omit the first coherence estimator ($\lambda = k$).

The computational complexity of the approach mainly depends on the number of divisions associated with coherence estimation (5). This is, however, originally related with the "single-frame" algorithm. The number of divisions can be reduced by processing averages of frequency components at a time. Given fixed complexity constraints, we strongly recommend to design an unbiased residual echo estimator, if necessary at the cost of a lower frequency resolution.

4. APPLICATIONS

We mention two most important applications where the necessity for reliable residual echo PSDs arises.

4.1. Postfiltering for Joint Acoustic Echo and Noise Control

The underlying application, currently driving the development of residual echo estimation techniques, is the frequency-domain adaptive postfilter for the purpose of combined suppression of residual echo and background noise, as outlined in the introduction of this article.

Spectral weighting

$$\hat{S}(\Omega,k) = H_W(\Omega,k)E(\Omega,k) \tag{17}$$

on the basis of the Wiener rule

$$H_W(\Omega,k) = \frac{\Phi_{SS}(\Omega,k)}{\Phi_{SS}(\Omega,k) + \Phi_{NN}(\Omega,k) + \Phi_{BB}(\Omega,k)} \tag{18}$$

can be viewed as the simplest form of DFT–based speech enhancement, where $\Phi_{BB}(\Omega,k)$ is the residual echo power spectral density and $\Phi_{NN}(\Omega,k)$ is the background noise power spectral density required for the algorithm. Both parameters are crucial for the reliability of the spectral weights $H_W(\Omega,k)$. The background noise PSD can be determined adaptively and accurately by the Minimum Statistics approach [5], whereby the desired noise PSD can be tracked even during speech activity.

Instead of Wiener filtering, we do actually use the more advanced MMSE-LSA spectral weighting algorithm [6] which, however, relies in a similar way on residual echo and noise PSD estimates.

4.2. Frequency-Domain Adaptive Echo Cancellation

We assume $D(\Omega,k) = G'(\Omega,k)X(\Omega,k)$, where $G'(\Omega,k)$ is an acoustic echo transfer function. We track the least-mean-square approximation to $G'(\Omega,k)$ by the unconstrained fast LMS algorithm

$$\hat{G}'(\Omega,k+1) = \hat{G}'(\Omega,k) + \mu(\Omega,k)\, X^*(\Omega,k)E(\Omega,k) \tag{19}$$

with stepsize factor $\mu(\Omega,k)$ and error signal

$$E(\Omega,k) = Y(\Omega,k) - \hat{G}'(\Omega,k)X(\Omega,k) \tag{20}$$

according to [3]. With regard to the least-mean-square criterion

$$\mathcal{E}\left\{\left|\hat{G}'(\Omega,k) - G'(\Omega,k)\right|^2\right\} \quad \rightarrow \quad \min \tag{21}$$

we obtain the optimal stepsize factor (compare [7])

$$\mu(\Omega,k) = \frac{\Phi_{BB}(\Omega,k)}{\Phi_{BB}(\Omega,k) + \Phi_{EE}(\Omega,k)} \cdot \frac{1}{\Phi_{XX}(\Omega,k)}\,, \tag{22}$$

which in turn depends on an accurate estimate of the residual echo PSD $\Phi_{BB}(\Omega,k)$. For the derivation of the stepsize, we assumed $\mathcal{E}\left\{|X(\Omega,k)|^4\right\} = 2\Phi_{XX}^2(\Omega,k)$, other calculations are straightforward.

5. RESULTS

5.1. Instrumental Measures

For the purpose of instrumental evaluation of residual echo estimation techniques, we make use of a set of frame-oriented *Log–Spectral–Distance* measures. Based on the DFT block length K_Ω, we consider the *Log–Spectral–Mean*

$$M(k) = \frac{1}{K_\Omega}\sum_\Omega 10\log_{10}\frac{\hat{\Phi}_{BB}(\Omega,k)}{\Phi_{BB}(\Omega,k)} \tag{23}$$

of the ratio of the estimated and the true residual echo power. Furthermore, we evaluate the *Log–Spectral–Mean* of the ratio of the estimated and the true spectral Wiener weights (18) required for residual echo and background noise suppression, given $\Phi_{SS}(\Omega,k)$ and $\Phi_{NN}(\Omega,k)$:

$$M_W(k) = \frac{1}{K_\Omega}\sum_\Omega 10\log_{10}\frac{\hat{H}_W(\Omega,k)}{H_W(\Omega,k)} \tag{24}$$

The latter $M_W(k)$ actually measures the reliability of the residual echo estimate with respect to its impact on spectral weighting. For example, in the presence of strong local speech activity, the residual echo estimate does not need to be as accurate as for a speech pause (in order to determine good spectral weights).

Eventually, we consider the *Log–Spectral–Distance*

$$D_W(k) = \sqrt{\frac{1}{K_\Omega}\sum_\Omega\left(10\log_{10}\frac{\hat{H}_W(\Omega,k)}{H_W(\Omega,k)}\right)^2} \tag{25}$$

between the estimated and the true spectral Wiener weights.

5.2. Numerical Results

We compare numerical results for the estimation techniques under consideration. In particular, we present residual echo estimates on the basis of white noise excitation $X(\Omega,k)$ under various levels of local speech $S(\Omega,k)$ and car background noise $N(\Omega,k)$. The acoustic echo is generated artificially by means of a car impulse response cut to 512 coefficients, the first 128 coefficients being cancelled ideally by the echo compensator. The DFT length for residual echo estimation is 256 data points with 50% frame overlap and Hann windowing. With regard to the short term stationarity of speech, we chose the forgetting factor $\alpha = 0.8$. The number of excitation frames involved with the multiple–frame coherence estimator is $L = 4$, the corresponding forgetting factors were individually chosen as $\alpha^{(k)} = 0.9$, $\alpha^{(k-1)} = 0.8$, $\alpha^{(k-2)} = 0.8$, $\alpha^{(k-3)} = 0.9$

Figure 2 depicts three different states of local disturbances: In the first 300 signal frames, there is no local contribution to the microphone signal, thus, only acoustic echo. In frames 300 to 600 we added local car background noise, and in frames 600 to 900, there is car background noise and local speech present (double talk). The corresponding true coherence level $C_{XE}(\Omega,k) = \Phi_{BB}(\Omega,k)/\Phi_{EE}(\Omega,k)$, which is depicted only in Figure 3 and 4 for clarity, indicates those situations.

From Figure 2 we observe that both the "single-frame" coherence estimator (4) and the extended virtual transfer estimator (12) do not completely reflect the residual echo. Note that the virtual

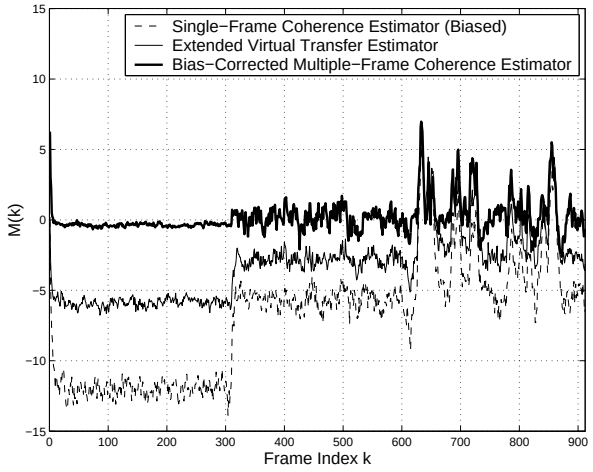

Figure 2: Bias of various residual echo estimators.

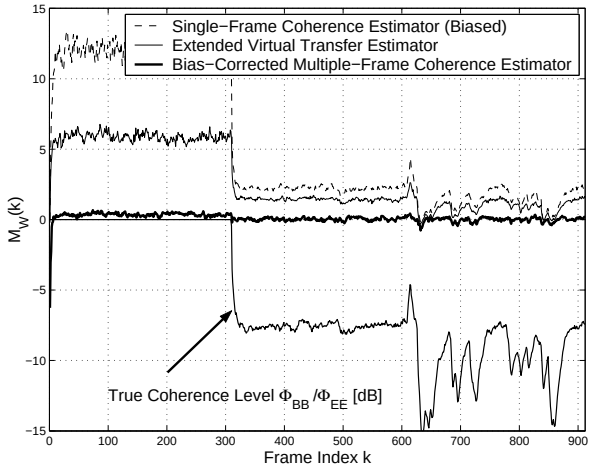

Figure 3: Bias of the Wiener weights due to different estimators.

transfer estimator delivers comparatively better results only because the estimated echo $\hat{D}(\Omega, k)$ is basically a delayed version of the excitation $X(\Omega, k)$. The bias of both estimators is most severe when there are no local disturbances. In the presence of local background noise and speech activity, both methods achieve better performance only because of the additional bias (short–term correlations) introduced in this case.

In contrast, Figure 2 shows that the multiple–frame coherence estimator achieves nearly unbiased residual echo estimates for any kind of local acoustic environment. In particular, during local pause we obtain an unbiased estimate due to the multiple–frame residual echo superposition (16). And in the case of local background noise (and speech), the coherence bias correction (14) avoids the occurrence of considerable overestimates.

The performance of the multiple–frame coherence estimator becomes even more favorable from Figure 3. Here, we can see that the estimation errors due to nonstationary local disturbances (compare Figure 2) do hardly ever impact the construction of spectral Wiener weights from residual echo estimates. Hence, we conclude that the multiple–frame coherence estimator delivers consistently excellent results for the application of postfiltering.

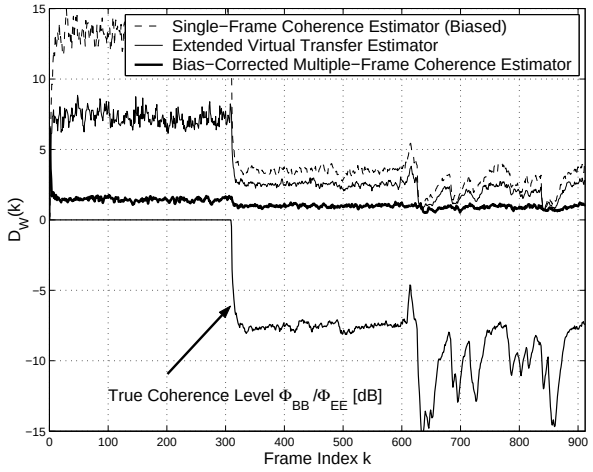

Figure 4: Variance of the Wiener weights for different estimators.

Eventually, Figure 4 proves that the multiple–frame coherence estimator also introduces the smallest "variance" $D_W(k)$ compared with other estimation techniques. The potentials for combined acoustic echo and noise postfiltering are obvious.

5.3. Auditive Results

The estimation techniques under consideration were simulated in our acoustic echo and noise control setup (see Figure 1) using synthetic and real world data. Using the multiple–frame coherence estimator, the spectral postfiltering algorithm achieved very high residual echo attenuation, while preserving very good speech quality during double talk.

CONCLUSIONS

We proposed the new block–partitioned (multiple–frame) residual echo estimator based on recursive averaging of periodograms and coherence bias correction. The estimator delivers unbiased results with regard to various acoustic environments.

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*. PhD thesis, University of Rennes 1, 1999.

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

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

[5] R. Martin, "Spectral Subtraction Based on Minimum Statistics." Proc. EUSIPCO-94, Edinburgh, pp. 1182-1185, September 12-16, 1994.

[6] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 33, pp. 443–445, April 1985.

[7] B. Nitsch, "A frequency-selective stepfactor control for an adaptive filter algorithm working in the frequency domain," *Signal Processing*, vol. 80, pp. 1733–1745, 2000.