IDENTIFICATION OF UNDERMODELLED ROOM IMPULSE RESPONSES

[1]*Geert Rombouts*, [1]*Toon van Waterschoot*, [2]*Kris Struyve*,[2]*Piet Verhoeve*, [1]*Marc Moonen*

[1]`geert.rombouts@esat.kuleuven.ac.be`
[1]KULeuven/ESAT-SCD, Kasteelpark Arenberg 10, 3001 Heverlee, Belgium
[2]Televic NV, Leo Bekaertlaan 1, 8870 Izegem, Belgium

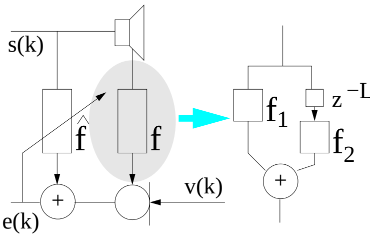

Figure 1: A traditional room impulse response identification scheme fails in the undermodelling case.

Abstract

In several scenarios it is desired to obtain an estimate of only the first part of the room impulse response, e.g. due to computing power restrictions. Room impulse response estimation is also often required in continuous doubletalk situations. In this paper we show that the PEM-AFROW algorithm which has recently been proposed for acoustic feedback cancellation, can be used in these situations to provide a low variance estimate with only a small bias.

1. INTRODUCTION

In this paper, we will focus on scenario's in which it is desired to provide an estimate of the first part (undermodelling) of a room impulse response (RIR), while continuous double talk is present. An example is an acoustic echo canceller followed by a post processor which can remove the residual echo due to the last part of the RIR (typically less energy).
Standard adaptive filtering algorithms will provide a biased and large variance estimate of such a truncated impulse response. In this paper we will show that the most important of both is the large variance, and that by using the PEM-AFROW algorithm [1, 2] which was derived in the context of acoustic feedback cancellation, bias and variance can be reduced. In section 2 a problem statement is given, and in section 3 the PEM-AFROW based approach is introduced. Section 4 provides the simulation results.

2. UNDERMODELLED ROOM IMPULSE RESPONSE

Figure 1 shows a room impulse identification scheme. We assume that the RIR $\mathbf{f} \in \mathbb{R}^N$ varies slowly compared to the statistics of the signals. We assume that the signals involved will be speech signals, and it is known that they can be modelled as time varying AR processes (TVAR). On the other hand, especially in undermodelling scenario's, speech segments which are stationary for a longer period than the length of the modelled part of the impulse response may occur. Hence in the simulations, we use both stationary and time variant AR-models. The near end signal $v(k)$ and the far end signal $s(k)$ are assumed to be independent. In order to describe the undermodelling case, we define

$$\mathbf{f} = \begin{pmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \end{pmatrix},$$

and we constrain

$$\hat{\mathbf{f}} = \begin{pmatrix} \hat{\mathbf{f}_1} \\ 0 \end{pmatrix}$$

with $\mathbf{f_1}, \hat{\mathbf{f_1}} \in \mathbb{R}^M$. The MMSE criterion which is solved in a traditional echo canceller can then be specified as

$$\min \varepsilon\{(\mathbf{s}^T(k) \left(\begin{pmatrix} \mathbf{f_1} \\ \mathbf{f_2} \end{pmatrix} - \begin{pmatrix} \hat{\mathbf{f_1}} \\ 0 \end{pmatrix} \right) + v(k))^2\} = \qquad (1)$$
$$\min \varepsilon\{(\mathbf{s_1}^T(k)\mathbf{f_1} + \mathbf{s_2}^T(k)\mathbf{f_2} - \mathbf{s_1}^T(k)\hat{\mathbf{f_1}} + v(k))^2\}$$

with $\mathbf{s_1}(k)$ a vector containing the first L elements of $\mathbf{s}(k)$, and $\mathbf{s_2}(k)$ the last $N - L$. Define the Hankel-matrices

$$S_i(k) \begin{pmatrix} \mathbf{s}_i(k)^T & \dots & \mathbf{s}_i(1)^T \end{pmatrix}^T \quad i = 1, 2$$

. The microphone signal consists of the signal $S_1(k)\mathbf{f}_1(k)$, and what could be described as 'noise' $n(k)$ for the identification process. The noise consists of the signal $S_2(k)\mathbf{f}_2(k)$ which has a component $n_b(k)$ *in* the column space of $S_1(k)$, and a component *orthogonal* to the column space of $S_1(k)$. The sum of the latter signal and the near end signal $v(k)$ will be called $n_v(k)$, and it will lead to variance on the estimate $\hat{\mathbf{f}_1}$ of $\mathbf{f}_1$, while $n_b(k)$ will lead to a bias on the estimate.

$$\mathbf{n}(k) = \underbrace{(S_2(k)\mathbf{f}_2(k))^{//}}_{n_b(k)\text{:Leads to bias}} + \underbrace{(S_2(k)\mathbf{f}_2(k))^{\perp} + \mathbf{v}(k)}_{n_v(k)\text{:Leads to variance}}$$

The bias can be expressed by setting the derivative to $\hat{\mathbf{f_1}}$ of (1) to zero :

$$\varepsilon\{\frac{\partial}{\partial \hat{\mathbf{f_1}}}\} = -2\varepsilon\{\mathbf{s_1}(k)\mathbf{s_1^T}(k)\mathbf{f_1} + \mathbf{s_1}(k)\mathbf{s_2^T}(k)\mathbf{f_2} -$$
$$\mathbf{s_1}(k)\mathbf{s_1^T}(k)\hat{\mathbf{f_1}}\} + \underbrace{\varepsilon\{\mathbf{s_1}(k)v(k)\}}_{=0} = 0$$

Now define $R_{11} = \varepsilon\{\mathbf{s_1}(k)\mathbf{s_1^T}(k)\}$and $R_{12} = \varepsilon\{\mathbf{s_1}(k)\mathbf{s_2^T}(k)\}$. We then have for the bias

$$\hat{\mathbf{f_1}} - \mathbf{f_1} = R_{11}^{-1} R_{12} \mathbf{f_2}. \qquad (2)$$

Note that the bias is zero when $s(k)$ is white noise.

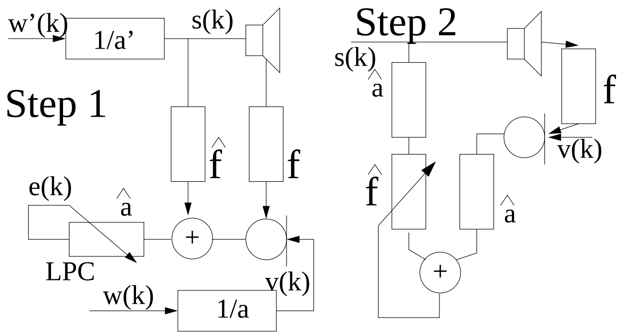

Figure 2: PemAFROW applied for identification of an undermodelled room impulse response.

3. PEM-AFROW BASED APPROACH

In **Figure 2** the PEM-AFROW scheme applied to undermodelled RIR identification is shown. The PEM-AFROW scheme [1, 2] is a prediction error method [3], specifically applied to nonstationary TVAR–signals under the assumption that the plant (the room impulse response) changes slower than the statistics of the signals. It can be used both in open loop [4, 5] and in closed loop [1, 2]. For the simulations which we will perform in this paper, we assume a near end noise signal

$$v(k) = a_1(k)v(k-1) + ... + a_P(k)v(k-P) + w(k),$$

with $w(k)$ a white noise sequence, and similarly, a far end signal

$$s(k) = a'_1(k)v(k-1) + ... + a'_P(k)v(k-P) + w'(k).$$

Define

$$\mathbf{a}(k) = \begin{pmatrix} 1 & -a_1(k) & ... & -a_P & (k) \end{pmatrix}^T$$

and $\mathbf{a}'(k)$ similarly. The coefficients $\mathbf{a}(k)$ and $\mathbf{a}'(k)$ will either be fixed, or changing every 20 msec. The PEM-AFROW algorithm is based on an MSE cost function given as

$$\min_{\hat{\mathbf{a}},\hat{\mathbf{f_1}}} \varepsilon\{\left(\hat{\mathbf{a}}^T(k)\left(S_1(k)\left(\mathbf{f_1} - \hat{\mathbf{f_1}}\right) + S_2(k)\mathbf{f_2} + \mathbf{v}(k)\right)\right)^2\}.$$

In this criterium, $\hat{\mathbf{a}}(k)$ and $\hat{\mathbf{f_1}}(k)$ are estimated in an alternating fashion, and on a frame–by–frame basis. In a first step, $\hat{\mathbf{f_1}}(k)$ is assumed 'correct' and kept fixed, and $\hat{\mathbf{a}}(k)$ will be estimated by linear prediction such that the residual energy is minimized for a frame of data, in which the signals can be assumed to be stationary (20 msec for speech). This means that linear prediction is performed on *the combination* $n(k)$ of the AR–process $v(k)$ and the ARMA-process $S_2(k)\mathbf{f}_2(k)$. In a second step, for the same frame of data, the far end and the microphone signal are prefiltered by the linear prediction error filter $\hat{\mathbf{a}}(k)$, and from these prefiltered signals, the estimate $\hat{\mathbf{f_1}}(k)$ is updated.

It should be noted that in case of a white far end signal $s(k)$, in which a conventional echo canceller in an undermodelling setup would perform bias–free, inserting the prediction error filter $\hat{\mathbf{a}}(k)$ of order P, would lead to a bias $Q_1^{-1}Q_2\mathbf{f}_2$, where Q_1 is a band–diagonal matrix with $2P-1$ non–zero diagonals, and Q_2 a matrix with only non-zero elements on a triangle in the first P columns and on the last P rows. If the AR–models used are stable, Q_1^{-1} can be approximated as a band–diagonal matrix too, and the bias will mainly occur in the last P elements of $\hat{\mathbf{f}}_1$. Since P is usually small compared to L, these P elements can be discarded. On the other hand, the structure of the scheme in **Figure 2** allows for $\hat{\mathbf{a}}$ to form an inverse model for the AR–process generating the far end signal, and hence it can also *reduce* the bias. In this setup, $\hat{\mathbf{a}}$ will minimize the linear prediction error energy of $n(k)$, and its effect will depend on the relative energies of the different components of $n(k)$.

The PEM-AFROW based method will — as we will show in section 4 below — effectively reduce the variance on the estimate $\hat{\mathbf{f}}_1(k)$, since the energy of the orthogonal part $\mathbf{n}_v(k)$ of $\mathbf{n}(k)$ is reduced after prefiltering in 'Step 2'.

4. SIMULATIONS

In the simulations, we use an artificial room impulse response with 800 taps (as shown in **Figure 3**), and in each of the experiments, only the first 250 taps of this impulse response will be modelled. From the figure it is clear that a significant amount of energy resides in tap 251 to 800, and (as will become clear in the simulations) without precautions, estimates will be useless. All experiments were performed with least squares identification (batch solution of a least squares system), and *not* with stochastic gradient algorithms (NLMS). In NLMS–type algorithms the variance on the estimate would seem larger because of the excess mismatch which occurs in these algorithms due to the presence of near end signals. This effectively means that the proposed technique even provides a larger improvement on NLMS–type algorithms than on the LS solutions in the simulations. For the simulations, 500 trials were run, and then the bias and variance of the estimate of the modelled part of the room impulse response was calculated. Instead of the variance, we plot the square root of the variance (the standard deviation), because this can directly be compared to the amplitude of the room impulse response.

We first consider a stationary AR far end signal. While speech is of course nonstationary, this scenario is relevant in case of strong undermodelling (less than 20 msec of the impulse response). **Figure 4** shows a simulation where no near end signal is present. In the upper figure the square root of the variance (standard deviation) and the bias on the estimate of the first 250 taps are shown when direct identification (DI) is applied by solving a least squares system, and in the lower figure the result is shown when PEM-AFROW is applied. Without PEM-AFROW the standard deviation is about 0.5, to be compared to the amplitude of the room impulse response, which peaks to only 0.3, see **Figure 3**). The bias is concentrated in the last taps of the modelled part of the room impulse response, because due to the (stable) AR–process for the far end signal, the unmodelled part is less correlated with the new process samples than with the (older) process samples which correspond to the last taps of the modelled part. When PEM-AFROW is applied (lower part of the figure), the standard deviation drops spectacularly to 0.01, and the bias is lowered, but still concentrated in the last taps. This is an interesting result, since these last taps can easily be discarded.

In **Figure 5**, a stationary near end signal is added. The energy of the stimulus of this signal, $w(k)$ is chosen from an uniform random distribution between -5dB and 0 dB compared to the en-

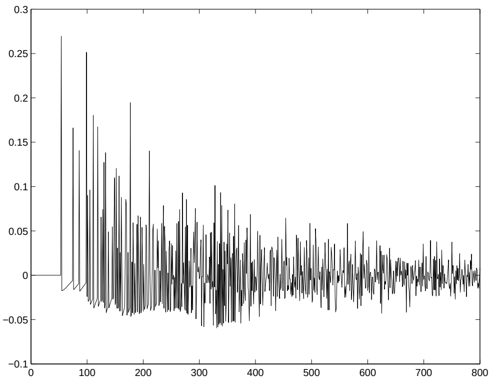

Figure 3: The room impulse response (800 taps) of which only 250 taps are modelled in the experiments.

ergy of the stimulus $w'(k)$ of the far end signal. Similar results are obtained : the bias is (both with and without application of PEM-AFROW) negligible compared to the variance, and concentrated around the last taps of the modelled part. The variance drops because the linear prediction error filter reduces the energy in the uncorrelated part of the microphone signal.

In **Figure 6**, the simulation is repeated for a time variant AR (TVAR) signal for the far end. The AR-coefficients are chosen randomly with a pole radius of 0.83, and kept stationary for about 20 msec (as in speech). In this experiment, no near end signal is added, and only the undermodelling performance is evaluated.

In **Figure** 7, the approach is validated on a real speech signal, with a real time implementation using NLMS adaptive filters. The impulse response was measured independently in silence in order to have a reference. The first 1000 taps of the 5000 taps impulse response are modelled and shown, together with their estimate performed by an NLMS adaptive filter (upper plot) and by PEM-AFROW with NLMS as its adaptive filter (lower plot) (contrarily to the experiments above, the plot only shows one realisation). The estimate by the PEM-AFROW with NLMS algorithm is observed to be much better than the estimate provided by NLMS (direct identification). This is most easily observed in the first 500 taps where the impulse response is zero.

5. CONCLUSION

We have experimentally shown that the PEM-AFROW algorithm can be used to provide estimates of an undermodelled room impulse response with both a low variance and a bias which is concentrated in a few filter taps, which can easily be discarded.

6. ACKNOWLEDGMENTS

This research work was carried out at the ESAT laboratory of the Katholieke Universiteit Leuven, in the frame of Belgian Programme on Interuniversity Attraction Poles, initiated by the Belgian Federal Science Policy Office IUAP P5/22 (‘Dynamical Systems and Control: Computation, Identification and Modelling’), the Concerted Research Action GOA-AMBioRICS, IWT project 040803: ’SMS4PA-II: Sound Management System for Public Address Systems’. The scientific responsibility is assumed by its authors.

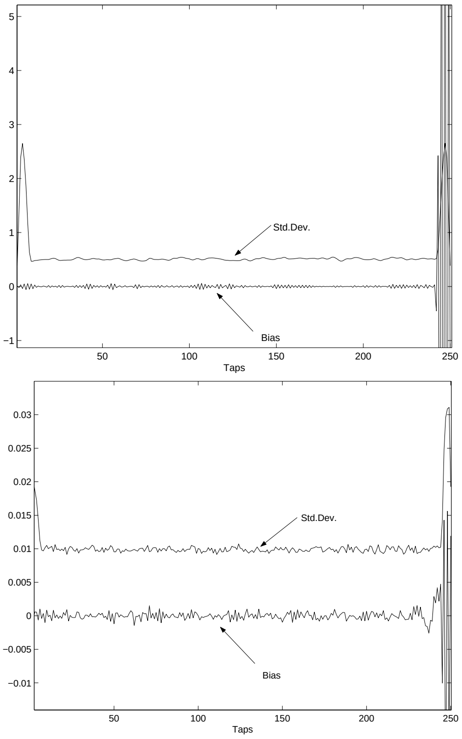

Figure 4: Stationary AR far end, no near end. Upper : direct identification (DI), lower : PEM-AFROW

7. REFERENCES

[1] G. Rombouts, T. van Waterschoot, K. Struyve, and M. Moonen, “Acoustic feedback suppression for long acoustic paths using a nonstationary source model.” Internal Report 05-71, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2004, Accepted for publication in the proceedings of EUSIPCO 2005, ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/ rombouts/reports/PemAFROW.pdf.

[2] G. Rombouts, T. van Waterschoot, K. Struyve, and M. Moonen, “Acoustic feedback suppression for long acoustic paths using a nonstationary source model.” Internal Report 04-151, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2004, ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/ rombouts/reports/PemAFROWFull.pdf.

[3] L. Ljung, *System Identification: Theory for the User*. Englewood Cliffs, New Jersey, USA: Prentice-Hall Inc., 1987.

[4] T. van Waterschoot, G. Rombouts, and M. Moonen, “On the performance of decorrelation by prefiltering for adaptive feedback cancellation in Public Address systems,” in *Proceedings of the 2004 IEEE Benelux Signal Processing Symposium (SPS 2004)*, (Hilvarenbeek, The Netherlands), pp. 167–170, Apr. 2004. Available from ftp://ftp.esat.kuleuven.ac.be/pub/sista/vanwaterschoot/ reports/report2004-24.ps.gz.

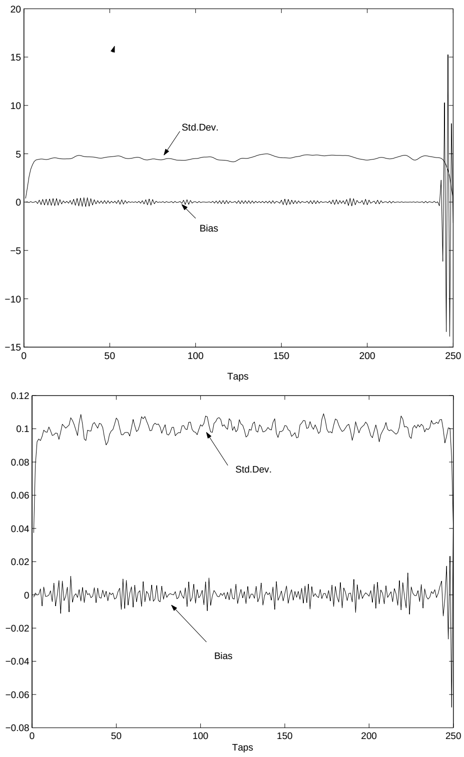

Figure 5: Stationary AR far end and near end. Upper : DI, lower : PEM-AFROW

[5] T. van Waterschoot, G. Rombouts, K. Struyve, and M. Moonen, “Acoustic echo cancellation in the presence of continuous double-talk,” in *Transactions of The first annual IEEE BENELUX/DSP Valley Signal Processing Symposium (SPS-DARTS 2005), Het Provinciehuis, Antwerp, Belgium*, IEEE Benelux/DSP Valley, April 2005.

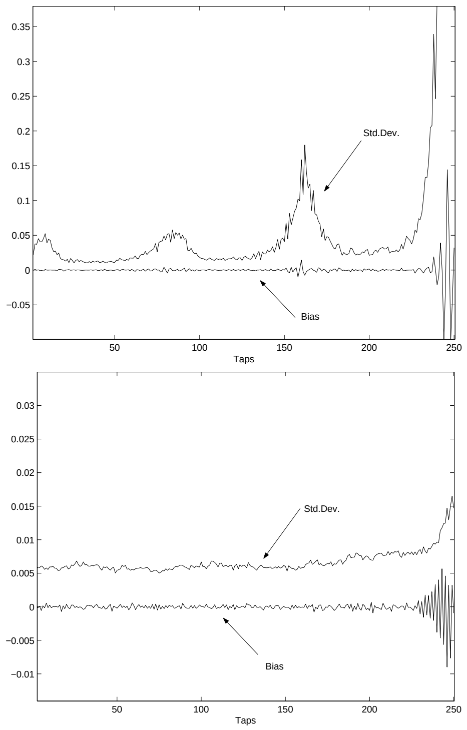

Figure 6: TVAR far end, no near-end. Upper : DI, lower : PEM-AFROW.

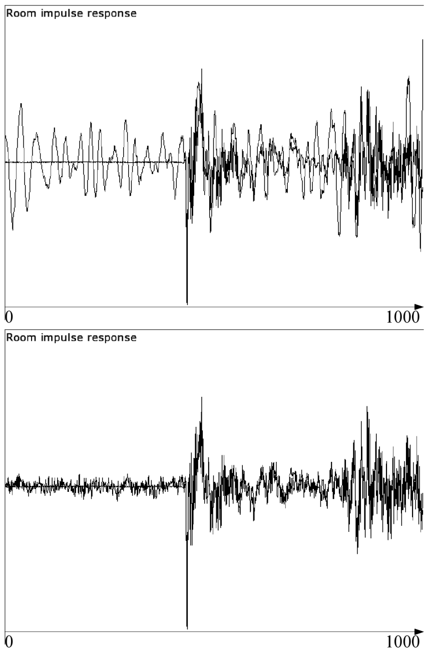

Figure 7: Upper : undermodelled (1000 out of 5000 taps) identification with NLMS. Lower : with PEM-AFROW.