Katholieke Universiteit Leuven

Departement Elektrotechniek ESAT-SISTA/TR 04-151

Acoustic Feedback Cancellation for Long Acoustic Paths Using a Nonstationary Source Model[1]

Geert Rombouts[2] [3], Toon van Waterschoot[2], Kris Struyve[4], and Marc Moonen[2]

March 2005
Published in *IEEE Transactions on Signal Processing*, vol. 54, no.9, Sep. 2006, pp. 3426–3434

[1]This report is available by anonymous ftp from *ftp.esat.kuleuven.be* in the directory *pub/sista/vanwaterschoot/reports/04-151.pdf*

[2]K.U.Leuven, Dept. of Electrical Engineering (ESAT), Research group SCD(SISTA), Kasteelpark Arenberg 10, B-3001 Leuven, Belgium, Tel. +32 16 321927, Fax +32 16 321970, WWW: *http://www.esat.kuleuven.be/sista-cosic-docarch*. E-mail: *toon.vanwaterschoot@esat.kuleuven.be*.

[3]This research work was carried out at the ESAT laboratory of the Katholieke Universiteit Leuven and was supported by IWT under Project 020476 “SMS4PA: Sound Management System for Public Address Systems,” the Belgian Federal Science Policy Office IUAP P5/22 (“Dynamical Systems and Control: Computation, Identication and Modelling”) under the Belgian Programme on Interuniversity Attraction Poles, the Flemish Government under the Concerted Research Action GOA-MEFISTO-666 (“Mathematical Engineering for Information and Communication Systems Technology”), and by Research Project FWO G.0233.01 (“Signal Processing and Automatic Patient Fitting for Advanced Auditory Prostheses”). The scientific responsibility is assumed by its authors.

[4]Televic N.V., Leo Bekaertlaan 1, B-8870 Izegem, Belgium, Tel. +32 51 303045, Fax. +32 51 310670, WWW: *http://www.televic.com* E-mail *k.struyve@televic.com*

Acoustic Feedback Cancellation for Long Acoustic Paths Using a Nonstationary Source Model

Geert Rombouts, Toon van Waterschoot, *Student Member, IEEE*, Kris Struyve, and Marc Moonen, *Senior Member, IEEE*

***Abstract*—While several proactive acoustic feedback (Larsen-effect) cancellation schemes have been presented for speech applications with short acoustic feedback paths as encountered in hearing aids, these schemes fail with the long impulse responses inherent to, for instance, public address systems. We derive a new prediction error method (PEM)-based scheme (referred to as PEM-AFROW) which identifies both the acoustic feedback path and the nonstationary speech source model. A cascade of a short- and a long-term predictor removes the coloring and periodicity in voiced speech segments, which account for the unwanted correlation between the loudspeaker signal and the speech source signal. The predictors calculate row operations which are applied to prewhiten the speech source signal, resulting in a least squares system that is solved recursively by means of normalized least mean square or recursive least squares algorithms. Simulations show that this approach is indeed superior to earlier approaches whenever long acoustic channels are dealt with.**

***Index Terms*—Acoustic feedback cancellation, Larsen effectbu.**

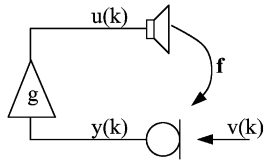

Fig. 1. A microphone-amplifier-loudspeaker-room system.

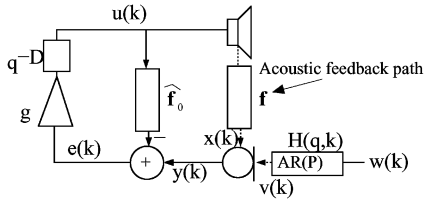

Fig. 2. Acoustic feedback cancellation scheme.

I. Introduction

ACOUSTIC feedback, also referred to as the *Larsen effect* (howling), occurs in microphone-amplifier-loudspeaker-room systems (Fig. 1) when the microphone is too close to the loudspeaker and/or when the amplification is too large. In Fig. 1, g is the amplifier gain, $y(k)$ is the microphone signal, $u(k) = gy(k)$ is the loudspeaker signal, $\mathbf{f}$ is the feedback path impulse response, and $v(k)$ is the (speech) source signal. Acoustic feedback occurs when the loop gain is larger than one at a frequency where the loop phase is a multiple of 2π.

A conventional solution consists of inserting notch filters into the signal path, thus decreasing the loop gain at those frequencies for which the problem arises. Examples of this approach are available in commercial products (Sabine [1], DBX and Behringer [2]). The main disadvantage is that the system is reactive, i.e., the howling phenomenon first needs to be detected before the notch filter can be computed and placed in the loop. In most systems, detection takes more than 0.5 s, hence a "proactive" approach is more desirable. Another disadvantage of the notch-filter-based systems is that they do not suppress the "reverb-like" sound that occurs in a system that is marginally stable.

In this paper, we will focus on single channel acoustic feedback cancellation (AFC) schemes as depicted in Fig. 2 (although multiloudspeaker (beamforming) techniques exist [3], [4]). This type of AFC scheme is inspired by acoustic echo cancellation (AEC) schemes. The filter coefficient vector $\mathbf{f}(k)$ of the acoustic path from the loudspeaker to the microphone

$$F(q,k) = \mathbf{f}(k)^T \mathbf{q} = (f_0 \quad f_1 \quad \dots \quad f_{N-1}) \begin{pmatrix} q^0 \\ q^{-1} \\ \vdots \\ q^{-N+1} \end{pmatrix}$$

is estimated as $\hat{\mathbf{f}}(k)$. Here q^{-1} is the delay operator. The N coefficients of $\hat{\mathbf{f}}(k)$ are copied at regular time instants to the cancellation filter $\hat{\mathbf{f}}_0(k)$. The loudspeaker signal $u(k)$ is filtered by the room impulse response $\mathbf{f}(k)$ and also by the cancellation filter $\hat{\mathbf{f}}_0(k)$. The difference between the cancellation filter output and the microphone signal is the error signal $e(k)$, which should then be equal to the speech source signal $v(k)$ (for a correct model $\hat{\mathbf{f}}(k)$). In Fig. 2, $w(k)$ is the excitation sequence of the source signal and

$$H(q,k) = \frac{1}{1 - a_1(k)q^{-1} - \dots - a_P(k)q^{-P}}$$

Manuscript received March 8, 2005; revised October 11, 2005. This work was supported by IWT under Project 020476 "SMS4PA: Sound Management System for Public Address Systems," the Belgian Federal Science Policy Office IUAP P5/22 ("Dynamical Systems and Control: Computation, Identification and Modelling") under the Belgian Programme on Interuniversity Attraction Poles, the Flemish Government under the Concerted Research Action GOA-MEFISTO-666 (Mathematical Engineering for Information and Communication Systems Technology), and by Research Project FWO G.0233.01 ("Signal Processing and Automatic Patient Fitting for Advanced Auditory Prostheses"). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Behrouz Farhang-Boroujeny.

G. Rombouts, T. van Waterschoot, and M. Moonen are with the ESAT Laboratory, Katholieke Universiteit Leuven, 3001 Heverlee, Belgium (e-mail: geert.rombouts@esat.kuleuven.ac.be; toon.vanwaterschoot@esat.kuleuven.ac.be; marc.moonen@esat.kuleuven.ac.be).

K. Struyve is with Televic NV, 8870 Izegem, Belgium (e-mail: k.struyve@televic.com).

Digital Object Identifier 10.1109/TSP.2006.879251

1053-587X/$20.00 © 2006 IEEE

is a time-varying autoregressive (AR) speech model of order P. The coefficient vector of the denominator is $\mathbf{a}(k)$. Finally, the q^{-D} block in Fig. 2 is a forward delay, which is often unavoidable in digital implementations (buffers for analog-to-digital (A/D)/digital-to-analog (D/A) converters, etc.), and which will be exploited further on.

An AEC-like approach has been used for AFC in, e.g., [5] and [6]. In AEC, an important assumption is that the "near end speech signal" $v(k)$ is uncorrelated with the loudspeaker signal $u(k)$, and hence Wiener filtering theory can be applied in order to identify the room impulse response with only the loudspeaker signal as an input to the system. The main complication in AFC compared to AEC is that this assumption is not fulfilled anymore. It is known that a correlation between the speech source signal $v(k)$ and the loudspeaker signal $u(k)$ results in a bias in the identified room impulse response when direct identification is applied (i.e., when the room impulse response is identified as if the system were operating in open loop) [7], [8].

This bias can be removed using the prediction error method (PEM), which incorporates a speech source signal model into the identification procedure [9]. The speech source signal in this system identification setup is to be considered as a "noise" signal, and its parameters have to be estimated along with the room impulse response. This speech model can then be used to prewhiten the speech source signal component in both the loudspeaker signal and the microphone signal. The feedback problem has been studied mainly in the hearing aids context where the feedback path impulse response is less than 5 ms. In this paper, we focus on public address (PA) systems, where the feedback path typically has a much longer impulse response, e.g., up to 500 ms, and hence an alternative approach will be needed.

Speech, although highly nonstationary over longer time periods, is often considered to be stationary during short frames of ca. 20 ms (e.g., 160 samples at 8 kHz). Within these frames, it can be whitened by a cascade of a short-term predictor (STP) and a long-term predictor (LTP). The STP models the filtering operation performed by the vocal tract, and the LTP models the periodic behavior of the glottal excitation sequence during voiced speech. The STP identifies an AR process of order P (e.g., $P = 10$ at 8 kHz). The LTP is an AR process with basically one single nonzero tap at a lag which can vary between 20 and 160 samples at 8 kHz. The room impulse response length for PA applications, on the other hand (unlike for hearing aid applications), can have several thousand of nonzero taps and is mostly stationary over longer periods of time. So it is advantageous to use data windows of several seconds to estimate the room impulse response, over which the speech signal then will be nonstationary. This contrast between the long stationarity period of the long room impulse response and the short stationarity period of the short-term prediction speech model, and the corresponding number of data points that are available to identify each of them, is fundamental to the problem of acoustic feedback cancellation for public address systems.

In [9], the authors derive a two-channel algorithm (2ChAFC), which uses the same data window length for the identification of the speech model as for the room impulse response identification. This is justified for short acoustic paths such as in hearing aids. For PA systems, this approach is suboptimal, as is shown in [10], but it still performs better than so-called direct identification, since it still provides some degree of prewhitening (with an "averaged" speech model).

Another method is described in [11], and its theoretical basis is derived in [9]. We will refer to this algorithm as PEM-AFC. It effectively distinguishes between the estimation window length for the speech model and for the room impulse response model, but it implicitly uses an assumption about the stationarity of the speech model, which is not valid for long acoustic impulse responses as in PA applications. Existing methods also use an incomplete speech model (short-term prediction only), which provides some whitening but does not remove the periodicity due to the glottal excitation, resulting in the loudspeaker signal's still being correlated to the source signal during voiced speech.

In this paper, we introduce a new technique that estimates the speech model over short time windows (over which it is stationary) and the room impulse model over longer time windows (which is necessary because the number of parameters is much larger). The speech model is then not required to be stationary during the complete room impulse response. Our scheme will also include a long-term predictor that models the periodicity in $w(k)$. The difference with the existing schemes like [9] is, aside from this long-term predictor, also the "row transformation" approach. The row transformations allow the whitening to be performed correctly (they can be interpreted as a decorrelation of the speech source component on the right-hand side of a least squares system), and we will show that this scheme indeed outperforms existing methods. These row transformations are required to perform the prewhitening step correctly, hence the algorithm is named *prediction error method based adaptive filtering with row operations* (PEM-AFROW).

This paper is organized as follows. In Section II, we first review the prediction error method, 2ChAFC, and PEM-AFC. In Sections III and IV, we introduce our new procedure. It uses alternating updates of the speech model and the adaptive filter which models the room. In Section VII, we show simulation results, and Section VIII contains the conclusion of this paper.

II. Prediction Error Method

In [9], the so-called prediction error method (known from system identification theory [12]) is applied to the AFC problem. Under the assumption of stationarity, it is proved that, if in a scheme as shown in Fig. 3 the forward delay D is larger than the order P of the AR process which generates the speech signal, the two-channel minimum mean squared error based adaptive filter ($\hat{A}(q,k)$, $\hat{B}(q,k)$) with filter coefficient vectors ($\hat{\mathbf{a}}(k)$, $\hat{\mathbf{b}}(k)$) will converge to the unique solution ($H^{-1}(q,k)$, $F(q,k)H^{-1}(q,k)$).

This allows to introduce a cancellation filter $\hat{F}_0(q,k) = A^{-1}(q,k)B(q,k)$ into the scheme (see Fig. 2). If this is implemented as an infinite impulse response filter, stability monitoring is needed, but it is a trivial modification to perform a deconvolution first, which results in a finite impulse response cancellation filter.

The filters $\hat{\mathbf{b}}(k)$ and $\hat{\mathbf{a}}(k)$ are updated concurrently in the two parallel branches, and so the scheme is referred to as two-channel adaptive feedback canceller (2ChAFC). The estimation data window is the same for both branches, and hence the

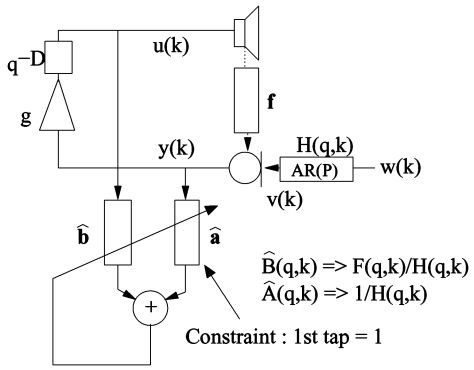

Fig. 3. Identification part of the 2ChAFC scheme [9].

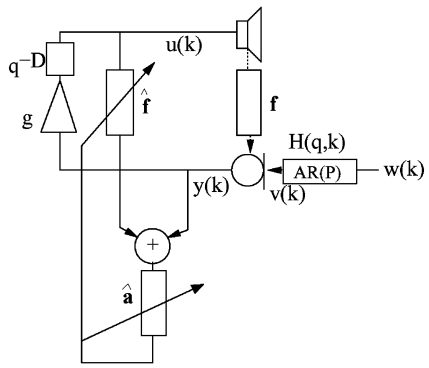

Fig. 4. PEM-AFC scheme [9], [11]. Only the identification part is shown. The estimate $\hat{\mathbf{f}}$ can be used in a scheme like Fig. 2 to perform acoustic feedback cancellation.

scheme provides good results in hearing aid applications, where the speech signal is stationary over the data window length required to estimate of the acoustic path.

PEM-AFC was proposed in [9] and [11]. This technique uses two independently updated adaptive filters $\hat{\mathbf{f}}(k)$ and $\hat{\mathbf{a}}(k)$, as shown in Fig. 4. The estimation of $\hat{\mathbf{f}}(k)$ and $\hat{\mathbf{a}}(k)$ can be performed over different data window lengths [different recursive least squares (RLS) forgetting factors]. In PEM-AFC, $\hat{\mathbf{a}}(k)$ decorrelates the *scalar* output residuals of the adaptive filter $\hat{\mathbf{f}}(k)$. In a stationary setup, where the order of both filters is arbitrary, this is equivalent to a prewhitening operation for the input of the filter $\hat{\mathbf{f}}(k)$. Also this scheme performs well in hearing aid applications, where the speech model does not change over the data window needed to estimate the acoustic impulse response.

As explained in Section I, this assumption does not hold in PA applications. In the next section, we will derive a new method with much better performance for long acoustic paths. It will become clear that PEM-AFC can be viewed as an approximation of this method. The performance gain of the new method in the next section will be shown by simulations.

III. Proposed Algorithm: PEM-AFROW

We model the speech source signal as a time-varying AR (TVAR) signal of order P (see Figs. 1 and 2)

$$v(k) = a_1(k)v(k-1) + \ldots + a_P(k)v(k-P) + b(k)w(k).$$

It is instructive to first assume that $w(k)$, the excitation sequence, is a white noise sequence, and the factor that accounts for the energy variations in the excitation sequence $b(k) = 1$. Further on, we will only assume $b(k)$ to be constant during the length of a frame. In Section IV, we will use a more general model for $w(k)$. We start from the minimization problem

$$\min_{\hat{\mathbf{f}}} \left\| U(k)\hat{\mathbf{f}}(k) - \mathbf{y}(k) \right\| \tag{1}$$

with

$$U(k) = \begin{pmatrix} u(k) & u(k-1) & \cdots & u(k-N+1) \\ u(k-1) & & & \\ \vdots & \vdots & \ddots & \vdots \\ u(0) & 0 & \cdots & 0 \end{pmatrix}$$

$$\mathbf{y}(k) = \begin{pmatrix} y(k) \\ y(k-1) \\ \vdots \\ y(0) \end{pmatrix}.$$

This minimization results in a biased estimate $\hat{\mathbf{f}}(k)$ for $\mathbf{f}(k)$ since $y(k)$ contains the "disturbance" $v(k)$ which is correlated with $u(k)$, i.e.,

$$\mathbf{y}(k) = U(k)\mathbf{f}(k) + \mathbf{v}(k) \tag{2}$$

with

$$\mathbf{v}(k) = \begin{pmatrix} v(k) \\ v(k-1) \\ \vdots \\ v(0) \end{pmatrix}.$$

If we have the finite impulse response filter

$$A(q,k) = H^{-1}(q,k)$$

available at each time instant, with coefficients $\mathbf{a}(k)$, we can apply a prewhitening by forming the matrix

$$A(k) = \begin{pmatrix} 1 & -a_1(k) & \ldots & -a_P(k) & 0 & 0 \\ 0 & 1 & -a_1(k-1) & \ldots & -a_P(k-1) & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

and modifying the minimization problem (1) to

$$\min_{\hat{\mathbf{f}}} \left\| A(k)U(k)\hat{\mathbf{f}}(k) - A(k)\mathbf{y}(k) \right\|. \tag{3}$$

This will effectively whiten the component $v(k)$ in $y(k)$ while performing only row operations on the input matrix $U(k)$, i.e., from (2)

$$A(k)\mathbf{y}(k) = A(k)U(k)\mathbf{f}(k) + \mathbf{w}(k)$$

and thus lead to an unbiased estimate $\hat{\mathbf{f}}(k)$. If we have an estimate $\hat{A}(q,k)$ of $H^{-1}(q,k)$ available at each time instant, with coefficients $\hat{\mathbf{a}}(k) \in \mathbb{R}^{P+1}$, we can apply a prewhitening by forming the matrix

$$\hat{A}(k) = \begin{pmatrix} 1 & -\hat{a_1}(k) & \dots & -\hat{a_P}(k) & 0 & 0 \\ 0 & 1 & -\hat{a_1}(k-1) & \dots & -\hat{a_P}(k-1) & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

and modifying the minimization problem (1) to

$$\min_{\hat{\mathbf{f}}} \left\| \hat{A}(k)U(k)\hat{\mathbf{f}}(k) - \hat{A}(k)\mathbf{y}(k) \right\|. \tag{4}$$

We now introduce the assumption that $\mathbf{a}(k)$ and $b(k)$ are constant during frames of 20 ms, which is a common assumption for speech signals. This means that we rewrite

$$\hat{A}(k) = \begin{pmatrix} 1 & -\hat{a}_i^1 & \dots & -\hat{a}_i^P & 0 & 0 & 0 & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & 0 & 0 & 0 \\ 0 & 0 & 1 & -\hat{a}_i^1 & \dots & -\hat{a}_i^P & 0 & 0 \\ 0 & & 0 & 1 & -\hat{a}_{i-1}^1 & \dots & -\hat{a}_{i-1}^P & 0 \\ 0 & & 0 & 0 & \ddots & \ddots & \ddots & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix} \begin{matrix} \uparrow \\ L \\ \downarrow \\ \uparrow \\ L \\ \downarrow \\ \vdots \end{matrix} \tag{5}$$

and

$$b_i = b(iL) = \ldots = b\left((i+1)L-1\right)$$

with, e.g., $L = 160$ for a sampling rate of 8 kHz and $i = \lceil k/L \rceil$, the first integer larger than or equal to k/L. This means that i is the frame index. We define $\hat{\mathbf{a}}_i = (1 \ -\hat{a}_i^1 \ \dots \ -\hat{a}_i^P)^T$.

We now decouple the nonlinear equations in order to calculate the estimates $\hat{\mathbf{a}}_i$ and the room response $\hat{\mathbf{f}}(k)$ in an alternating fashion. In the first step, an estimate $\hat{\mathbf{f}}(iL-1)$ from the previous frame is used to filter a frame $iL \ldots (i+1)L-1$ of data (20 ms); see Fig. 5. The filter output is subtracted from the corresponding microphone samples, resulting in

$$d(k) = y(k) - \mathbf{u}^T(k)\hat{\mathbf{f}}(iL-1), k = iL \ldots (i+1)L-1$$

with

$$\mathbf{u}(k) = \begin{pmatrix} u(k) \\ \vdots \\ u(k-N+1) \end{pmatrix}.$$

Note that if the estimate $\hat{\mathbf{f}}(iL-1)$ were exact, then $d(k) = v(k)$. Then linear prediction is performed on this $d(k)$, $k = iL \ldots (i+1)L-1$ (Levinson–Durbin algorithm) to find the linear prediction error filter $\hat{\mathbf{a}}_i$ for frame i.

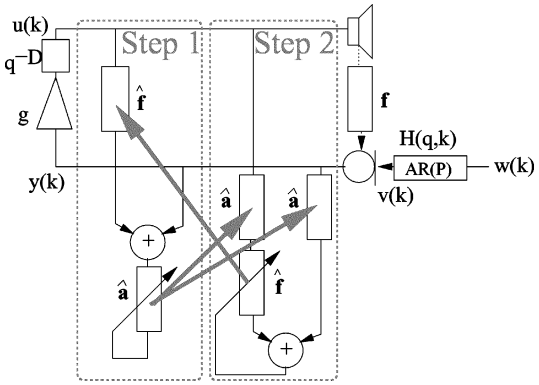

Fig. 5. PEM-AFROW. In the first step, $\hat{\mathbf{a}}$ is estimated on the left-hand side. It is then copied to the right-hand side, where the estimation of $\hat{\mathbf{f}}$ is performed on the same data frame (step 2). Finally, $\hat{\mathbf{f}}$ is copied to the left-hand side and used in the next frame.

In the second step, (4) is solved for $\hat{\mathbf{f}}(k), k = iL \ldots (i+1)L-1$, which is possible since $\hat{\mathbf{a}}_i$ is known from step 1. Note that the window length should be set appropriately. This improves the estimate $\hat{\mathbf{f}}(k)$ for $\mathbf{f}(k)$, since $\mathbf{f}(k)$ is assumed constant or "slow changing" compared to the frame length. Note that the stimulus energy b_i is assumed constant over a frame, and this means that the variance on the estimate $\hat{\mathbf{f}}(k)$ may differ from frame to frame. The two steps can be iterated on the frame. Since none of these two steps will increase

$$\varepsilon\left\{\|\mathbf{e}(k)\|\right\} = \varepsilon\left\{\left\|\hat{A}(k)U(k)\hat{\mathbf{f}}(k) - \hat{A}(k)\mathbf{y}(k)\right\|\right\}$$

with $\varepsilon\{\}$ the expected value operator, the algorithm will converge to a (possibly local) minimum of (4).

In order to reduce the complexity, we will perform only one iteration per frame. The minimization problem (4), with a fixed value of $\hat{A}(k)$, can be solved for $\hat{\mathbf{f}}(k)$ by means of any adaptive filtering algorithm. We have implemented this both using a QR-decomposition-based RLS (QRD-RLS) algorithm and a normalized least mean square (NLMS) algorithm. The input vector is in both cases

$$\mathbf{u}_w^T(k) = \mathbf{a}_i^T \begin{pmatrix} \mathbf{u}^T(k) \\ \vdots \\ \mathbf{u}^T(k-P) \end{pmatrix}, \ i = \lceil k/L \rceil \tag{6}$$

and the desired signal input (right-hand-side sample) is

$$\mathbf{a}_i^T \begin{pmatrix} y(k) \\ \vdots \\ y(k-P) \end{pmatrix}.$$

Since $\mathbf{u}(k)$ is a shifted version of $\mathbf{u}(k-1)$ with one sample prepended and $\mathbf{a}_i$ remains constant during a frame of L samples, $\mathbf{u}_w(k)$ will be a shifted version of $\mathbf{u}_w(k-1)$ with one sample prepended. So inside a frame, only one vector multiplication has

to be performed to calculate $\mathbf{u}_w(k)$. On the other hand, at the start of each frame, a matrix multiplication should be performed to calculate all elements of $\mathbf{u}_w(iL)$ as follows:

$$\mathbf{u}_w^T(iL) = \mathbf{a}_i^T \begin{pmatrix} \mathbf{u}^T(iL+1) \\ \vdots \\ \mathbf{u}^T(iL-P+1) \end{pmatrix}.$$

For real-time implementation, the scheme involves a delay of one frame for the update of $\hat{\mathbf{f}}(k)$, since $\mathbf{a}_i$ can only be calculated at time $k=(i+1)L-1$. Note that this is not a problem since we have assumed that the room impulse response is constant over more than one frame. The delay is effectively implemented as a delay line for the input samples $u(k)$ before they are fed to (6).

Once the room impulse response has been identified, the next step is to insert the cancellation filter into the feedback loop scheme by setting $\hat{\mathbf{f}}_0(k) = \hat{\mathbf{f}}(k)$, e.g., at regular time intervals (see Fig. 2). It is important to notice that this obviously influences the adaptation. The input data used for the identification procedure then depend on the current model estimate, which is reminiscent of a nonlinear optimization problem. This dependency is effectively ignored in our implementation (it is also ignored in adaptive control theory [13]). Experiments indicate that updating the cancellation filter regularly is beneficial to the identification process. This can be explained because a time-variant forward path (from microphone to loudspeaker) decreases the correlation between the loudspeaker signal and the speech source signal.

At this point, the difference between PEM-AFC and PEM-AFROW becomes obvious: in PEM-AFROW, the stationarity of the speech model is explicitly assumed in the minimization problem by stating that $\hat{\mathbf{a}}_i$ remains constant during a frame [see (5)]. At the start of each frame, the full input vector $\mathbf{u}_w^T(k)$ is recalculated. In PEM-AFC, this assumption of stationarity is not made for the optimization problem itself (the optimization is decoupled in two completely independent adaptive filters), and the full input vector is never recomputed after a change of $\mathbf{a}(k)$ in PEM-AFC. This can only be justified for short acoustic impulse responses, and (as will be shown later on) PEM-AFC performs much worse than PEM-AFROW for long acoustic impulse responses.

IV. Long Term Prediction

For the TVAR signals studied in the previous section (where $w(k)$ was a white noise sequence), the prewhitening step removes all of the correlation between the loudspeaker signal and the source signal. However, the excitation sequence $w(k)$ for voiced speech is periodical (glottal excitation). Hence the input signal $u(k)$ to the adaptive filter is—due to this periodicity—still correlated with the source signal, even after prewhitening. This in practice leads to a severely biased estimate of the room impulse response (the bias being localized at the excitation harmonics).

A possible solution could be to increase the length (span) of the prediction filter $\mathbf{a}(k)$ so that it becomes longer than the period of the excitation sequence. It would then combine the periodicity together with the AR coefficients for the vocal tract model. The problem here is that the pitch period can be as long as 160 samples, and hence not enough data is then available in a 160-sample window to estimate the $(160+P-1)$ prediction filter coefficients.

A standard approach in speech coding [14] is to cascade an STP of order P (e.g., 12), which models the vocal tract characteristics

$$\mathbf{u}_{sw}^T(k) = \mathbf{a}_i^T \begin{pmatrix} \mathbf{u}^T(k) \\ \vdots \\ \mathbf{u}^T(k-P) \end{pmatrix},\ i = \lceil k/L \rceil \tag{7}$$

with an LTP with only one tap and a lag equal to the pitch period to model the periodicity

$$\begin{aligned} u_{lsw}(k) &= u_{sw}(k) - b_j u_{sw}(k-M_j) \\ j &= \lceil k/L_{ltp} \rceil . \end{aligned}$$

The subscript "sw" means whitened with a short-term predictor, and "lsw" whitened with a cascade of a short- and a long-term predictor. The LTP can be estimated in windows of 20 ms (which is the frame length L of the short-term predictor), with a 10 ms overlap. This means that the LTP model is estimated each 10 ms, corresponding to L_{ltp} samples (at 8 kHz, $L_{ltp}=80$). In order to estimate the LTP, we minimize

$$\begin{aligned} E_j &= \min \varepsilon \left\{ \|u_{lsw}(k)\|^2 \right\} \\ &= \min_{\{M_j, b_j\}} \varepsilon \left\{ \|u_{sw}(k) - u_{sw}(k-M_j) b_j\|^2 \right\}. \end{aligned}$$

The solution follows from

$$\varepsilon \left\{ u_{sw}(k-M_j) u_{sw}(k-M_j) \right\} b_j = \varepsilon \left\{ u_{sw}(k) u_{sw}(k-M_j) \right\}.$$

In order to estimate these values, we define

$$\overline{\mathbf{u}}_{sw}(k) = \begin{pmatrix} u_{sw}(k) \\ u_{sw}(k-1) \\ \vdots \\ u_{sw}(k-L+1) \end{pmatrix}.$$

We can then estimate

$$b_j = \left(\overline{\mathbf{u}}_{sw}^T(k-M_j) \overline{\mathbf{u}}_{sw}(k-M_j) \right)^{-1} \overline{\mathbf{u}}_{sw}^T(k) \overline{\mathbf{u}}_{sw}(k-M_j).$$

The variance of the long-term prediction residual is

$$E_j = \overline{\mathbf{u}}_{sw}^T(k) \overline{\mathbf{u}}_{sw}(k) - \frac{\left(\overline{\mathbf{u}}_{sw}^T(k) \overline{\mathbf{u}}_{sw}(k-M_j) \right)^2}{\overline{\mathbf{u}}_{sw}^T(k-M_j) \overline{\mathbf{u}}_{sw}(k-M_j)}.$$

This variance E_j is evaluated for different values of $M_{j,i} = M_{\min} \ldots M_{\max}$ (the lag), and the parameters (M_j, b_j) that result in the minimum value of E_j are chosen as the predictor for long-term prediction frame j.

In the PEM-AFROW method of Section III and Fig. 5, the speech model will now be a cascade of long- and short-term prediction. The details of the resulting algorithm will be given in Section V.

It is important to note that by applying long-term prediction, the actual order of the speech source model is the lag of the long-term model plus the order of the short-term model, and as stated in [9], to guarantee identifiability, the forward delay must be larger than the order of this model. In practice this delay is approximately 20 ms, but then it does not matter too much where this forward delay is implemented: often a latency D is introduced by buffering after and before the A/D and D/A converters, or even—due to the relatively low velocity of sound waves—this delay appears automatically from the distance between the loudspeaker and the microphone.

In Section III it was mentioned that at frame borders, the whole input vector has to be recalculated by means of a matrix multiplication. It must be noted that when long-term prediction is added to the algorithm, this matrix multiplication has to be performed not only at frame borders of the short-term predictor but also at frame borders of the long-term predictor. The next section will show this in detail.

Note that multiple-tap long-term prediction [14] may be used instead of single-tap long-term prediction. In the algorithmic description of the next section, we have made abstraction of this in order to keep the pseudocode readable. In the algorithm description, we also assumed no overlap between long-term prediction frames. This means that the long-term prediction parameters M_j and b_j are estimated only once per short-term prediction frame (i.e., $i = j$).

V. Algorithm Description

In order to apply row updates, the input vector of the adaptive filter $\mathbf{u}_{lsw}$ (which consists of the residual signal of the cascade of the short- and long-term predictors) and the vector from which this is computed, namely, $\mathbf{u}_{sw}$, have to be recalculated whenever the short-term predictor changes (which happens at each frame border). When the vector $\mathbf{u}_{sw}$ has been recalculated M samples ago (M is the lag of the long-term predictor), $\mathbf{u}_{lsw}$ has to be recomputed completely (by applying the long-term predictor). At all other time instants $\mathbf{u}_{sw}$ and $\mathbf{u}_{lsw}$ are shifted versions of the previous $\mathbf{u}_{sw}$ and $\mathbf{u}_{lsw}$, with only one new element in front.

If the long-term prediction data windows overlap, the vector $\mathbf{u}_{lsw}$ has to be recomputed whenever the long-term predictor changes.

A threshold T was included in order to account for A/D, D/A, or amplifier noise. Without this threshold, the algorithm's coefficients could drift during silence. A complete pseudocode description is given in Algorithm 1.

Algorithm 1 PEM-AFROW. The function 'Adapt (coefficients, inputvector, desired signal sample)' is an adaptive filtering algorithm, e.g., RLS or NLMS. The function 'LongTermPredictor (short term prediction residual vector)' performs the operations which are explained in Section IV.

For each sample $u(k)$, $y(k)$, $k = 0, 1, \ldots$:

$\mathbf{u}(k) = [u(k); \overline{\mathbf{u}}(k-1)]$

$\mathbf{y}(k) = [y(k); \overline{\mathbf{y}}(k-1)]$

// $\overline{\cdot}$ means all components but last

// $\mathbf{u}(k)$, $\mathbf{y}(k)$ have length P

$d(k) = y(k) - \hat{\mathbf{f}}_0^{\,T}(k) * [u(k); u(k-1) \ldots u(k-N+1)]$

// $d(k)$ is the compensator output

$\mathbf{d}(k) = [d(k); \overline{\mathbf{d}}(k-1)]$

// $\mathbf{d}(k)$ has length L.

// Short term whitening

$u_{sw}(k-L) = \mathbf{u}^T(k-L)\mathbf{a}$

$\mathbf{u}_{sw}(k-L) = [u_{sw}(k-L); \overline{\mathbf{u}}_{sw}(k-L-1)]$

$y_{sw}(k-L) = \mathbf{y}^T(k-L)\mathbf{a}$

// Long term whitening

$u_{lsw}(k-L) = u_{sw}(k-L) - b u_{sw}(k-L-M)$

$y_{lsw}(k-L) = y_{sw}(k-L) - b y_{sw}(k-L-M)$

$\mathbf{u}_{lsw}(k-L) = [u_{lsw}(k-L); \overline{\mathbf{u}}_{lsw}(k-L-1)]$

if $(\mathbf{u}_{lsw}^T(k-L)\mathbf{u}_{lsw}(k-L) > T)$

    // Energy threshold T: adapt or not

    $\hat{\mathbf{f}}(k+1) = \mathtt{Adapt}(\hat{\mathbf{f}}(k), \mathbf{u}_{lsw}(k-L), y(k-L))$; // With Rls or Nlms

if $k = tL - 1$ with t integer

    $(\mathbf{a}, \mathbf{r_{stp}}) = \mathtt{LevinsonDurbin}(\mathbf{d}(k))$

    $(M, b) = \mathtt{LongTermPredictor}(\mathbf{r_{stp}})$

    // Recalc. whole NLMS input vector

    for i = 0: N − 2

        $u_{sw,i}(k-L) = \mathbf{u}^T(k-L-i).\mathbf{a}$

        $u_{lsw,i}(k-L) = u_{sw,i}(k-L) - u_{sw,i}(k-L-M).b$

    endfor

endif

if $k - M = tL - 1$ with t integer

    // Recalc LTP on full input vector

    for i = 0: N − 2

        $u_{lsw,i}(k-L+1) = u_{sw,i}(k-L+1) - u_{sw,i}(k-L-M+1).b$

    endfor

endif

if DecideToUpdateCompensator (k)

    $\hat{\mathbf{f}}_0(k) = \hat{\mathbf{f}}(k)$

else

    $\hat{\mathbf{f}}_0(k) = \hat{\mathbf{f}}_0(k-1)$

endif

VI. Complexity

The complexity is evaluated when the algorithm is operated with an NLMS adaptive filter. In these complexity expressions, a multiplication and an addition are counted as two separate floating-point operations. A "search range" $M_{\min}$ to $M_{\max}$ has to be specified for the lag of the long-term predictor (at 8 kHz, typically $M_{\min} = 20$, $M_{\max} = 160$). The complexity depends on these parameters through

$$dM = M_{\max} - M_{\min}.$$

For the complexity calculation, we assume one tap long-term prediction, and we also assume that the frames do not overlap. Since at each frame border the full NLMS input vector is recalculated, the complexity per sample is

$$8(N+P) + 4dM + 5 + \frac{(2P+4)N + 4P^2 - 5P + 15}{L}$$

floating-point operations. The algorithm was implemented in C++ on a Pentium III 1 GHz PC without any specific optimization effort and runs in real time with $M_{\min} = 40$, $M_{\max} = 320$, $N = 2000$, $P = 24$, $L = 320$ at 16 kHz sampling rate, with long-term prediction overlap of 160 samples.

VII. Simulation Results

In Fig. 6, we first compare the PEM-AFROW approach to the decoupled PEM-AFC approach for short paths, as they occur in, e.g., hearing aids. In the figure, the average of the error norm

$$\left\| \mathbf{f}(k) - \hat{\mathbf{f}}(k) \right\|$$

for seven different speech signal inputs is plotted as a function of time. In order to provide an idea of the variance, also the convergence curves of the individual realizations are shown. Note that only the identification performance (in closed loop) is shown, which means that the cancellation filter $\hat{\mathbf{f}}_0$ (see Fig. 2) is *not* inserted into the scheme during adaptation. If the filter would be inserted, the performance will generally be better, since inserting the filter corresponds in fact to inserting an uncorrelated signal into the loop. The acoustic path has 50 taps. We use NLMS for both approaches, since in a practical implementation this would be the adaptive algorithm of choice (due to complexity constraints). The NLMS step size is 0.04 for PEM-AFROW, while PEM-AFC, which uses a modified NLMS algorithm and hence a different definition of the step size, was tuned to give equal initial convergence speed (step size 0.03). This allows us to make a fair comparison of the resulting bias/variance of the solution. Note that the performance of all algorithms is dependent of the energy ratio (signal-to-noise ratio) of the loudspeaker component arriving on the microphone versus the source signal arriving on the microphone (the source signal should thus be interpreted as "noise"). The simulations shown here were performed with a value of -5 dB

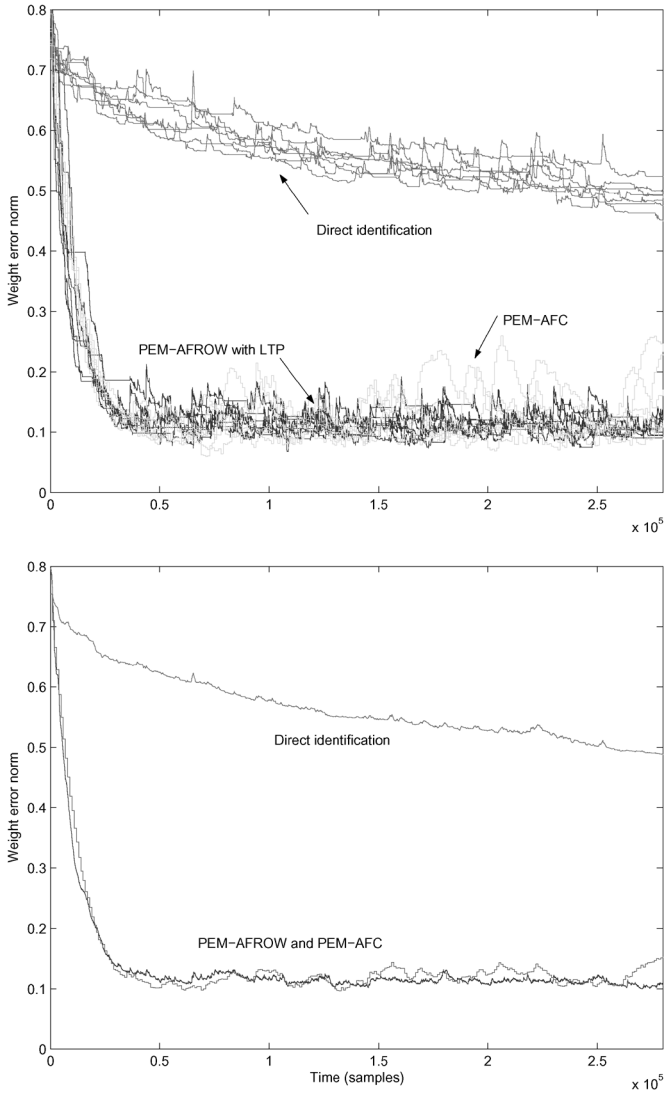

Fig. 6. Closed-loop identification for short acoustic paths (50 taps). Upper figure: weight error norm versus time for seven realizations with different speech signals for PEM-AFROW, PEM-AFC, and direct identification. Lower figure: the averages of the convergence curves of the realizations. PEM-AFROW performs equally well as PEM-AFC for short paths (as expected). PEM-AFROW with LTP does not perform better than without LTP.

for this ratio, which was the edge of stability for this setup, and hence certainly relevant. Experiments show a similar performance difference between the algorithms for other ratios. The short time prediction frame length is 160 samples, the long-term prediction frame overlap is 80 samples, and the minimum and maximum lag for the long-term predictor are 20 and 160, respectively. The sampling frequency is 8 kHz. The speech model order in PEM-AFROW and PEM-AFC is ten. The forward delay is 200 taps in both PEM-AFROW and PEM-AFC (note that the PEM-AFC version of [11] does not explicitly incorporate a forward delay, but the theoretical analysis of [9] shows that this is required for correct performance, hence we added it to the system). PEM-AFC and PEM-AFROW are seen to yield comparable performance, which is to be expected since the assumption of stationarity of the speech model during the data window

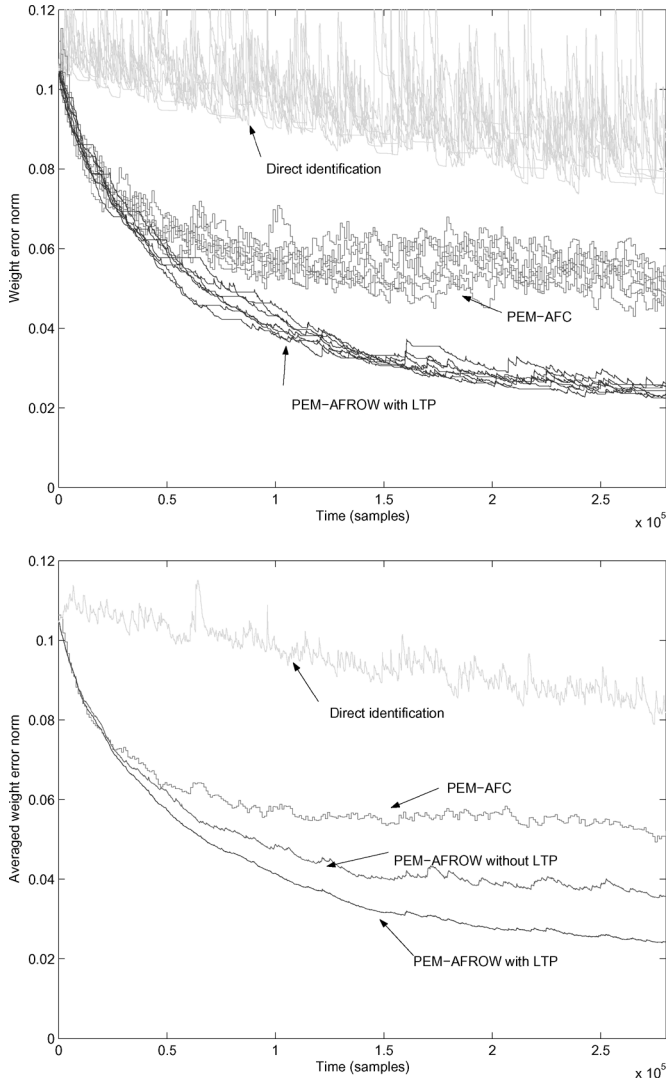

Fig. 7. Identification for long acoustic paths (1000 taps). Upper figure: weight error norm versus time for seven realizations with different speech signals for PEM-AFROW, PEM-AFC, and direct identification. Lower figure: the averages of the weight error norms for the realizations. In the left figure, PEM-AFROW without LTP is not shown to preserve the clarity of the figure. PEM-AFROW clearly outperforms the others.

for acoustic path identification is fulfilled in this experiment. The difference with PEM-AFROW including LTP is insignificant (and hence this curve is not plotted). Direct identification is seen to give poor results (i.e., when the room impulse response is identified as if the system were operating in open loop).

For longer paths, as in PA applications, PEM-AFC performance decreases with path length, and for, e.g., $N = 1000$, its behavior is certainly better than the direct identification behavior, as shown in Fig. 7, but a large bias remains. Also here the shown convergence curve of PEM-AFC was tuned to have equal initial convergence to PEM-AFROW. PEM-AFROW, on the other hand, does perform well also for long paths. The poor performance of PEM-AFC is to be attributed to the stationary speech model assumption, which is not fulfilled for long paths.

VIII. Conclusion

We have introduced a new algorithm, referred to as PEM-AFROW, which allows for acoustic feedback cancellation in setups with long acoustic paths. It uses a speech source model with short- and long-term prediction. Not only the howling phenomenon is suppressed but also the reverberation-like sounds, which indeed become audible in the region of marginal stability. The main difference with existing schemes is twofold. First, our algorithm incorporates a long-term prediction filter which removes periodicity in the short-term speech signal residual; second, we do not assume stationarity of the speech signal over the length of the data window on which the acoustic path is identified.

PEM-AFROW hence performs very well for long acoustic paths, contrarily to existing algorithms that were developed for short path applications. For short acoustic paths, the performance of PEM-AFROW is equal to that of the existing methods. Thanks to the low complexity, the algorithm can easily be implemented in real time.

References

[1] M. P. Lewis, T. J. Tucker, and D. M. Oster, "Method and apparatus for adaptive audio resonant frequency filtering," U.S. Patent 5 245 665, Sep. 1993.

[2] R. Porayath and D. J. Mapes-Riordan, "Acoustic feedback elimination using adaptive notch filter algorithm," U.S. Patent 5 999 631, Dec. 1999.

[3] C. P. Janse and H. J. W. Belt, "Sound reinforcement system having an echo suppressor and loudspeaker beamformer," Patent WIPO WO 03/010996 A3, Jun. 24, 2002.

[4] G. W. Elko and M. M. Goodwin, "Beam dithering: acoustic feedback control using a modulated-directivity loudspeaker array," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1993, vol. 1, pp. 173–176.

[5] S. Kamerling, K. Janse, and F. van der Meulen, "A new way of acoustic feedback suppression," *AES*, 1998.

[6] C. P. Janse and P. A. A. Timmermans, "Signal amplifier system with improved echo cancellation," U.S. Patent 5 748 751, May 1998.

[7] M. G. Siqueira and A. Alwan, "Steady-state analysis of continuous adaptation in acoustic feedback reduction systems for hearing-aids," *IEEE Trans. Speech Audio Process.*, vol. 8, pp. 443–453, Jul. 2000.

[8] J. Hellgren and U. Forssell, "Bias of feedback cancellation algorithms in hearing aids based on direct closed loop identification," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 906–913, Nov. 2001.

[9] A. Spriet, I. Proudler, M. Moonen, and J. Wouters, "Adaptive feedback cancellation in hearing aids with linear prediction of the desired signal," *IEEE Trans. Signal Process.*, vol. 53, no. 10, pp. 3749–3763, Oct. 2003.

[10] T. van Waterschoot, G. Rombouts, and M. Moonen, "On the performance of decorrelation by prefiltering for adaptive feedback cancellation in public address systems," in *Proc. 2004 IEEE Benelux Signal Process. Symp. (SPS 2004)*, Hilvarenbeek, The Netherlands, Apr. 2004, pp. 167–170 [Online]. Available: ftp://ftp.esat.kuleuven.ac.be/pub/sista/vanwaterschoot/reports/04-24.pdf

[11] R. Leber and W. Schaub, "Circuit and method for the adaptive suppression of an acoustic feedback," U.S. Patent 6 611 600, 2003.

[12] L. Ljung, *System Identification: Theory for the User*. Englewood Cliffs, NJ: Prentice-Hall, 1987.

[13] L. Ljung and S. Soderstrom, *Theory and Practice of Recursive Identification*. Cambridge, MA: MIT Press, 1983, ch. 7.

[14] R. P. Ramachandran and P. Kabal, "Pitch prediction filters in speech coding," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, pp. 467–477, Apr. 1989.



Geert Rombouts was born in Turnhout, Belgium, in 1973. He received the Master's degree in electrical engineering and the Ph.D. degree in engineering (Toegepaste Wetenschappen) from the Katholieke Universiteit Leuven, Belgium, in 1997 and 2003, respectively.

Currently, he is a Postdoctoral Researcher with the ESAT-SCD laboratories, Katholieke Universiteit Leuven.



Toon van Waterschoot (S'04) was born in Lier, Belgium, on June 11, 1979. In 2001 he graduated received the master's degree in electrical engineering from Katholieke Universiteit Leuven, Belgium, in 2001, where he is currently pursuing the Ph.D. degree.

He spent a year as a Teaching Assistant with the Hogere Zeevaartschool Antwerpen, Belgium. His doctoral research project is on signal processing applied to audio and acoustics. His current research is focused on the development of digital signal-processing algorithms for acoustic echo and feedback cancellation.



Kris Struyve received the Ph.D. degree in telecommunications engineering from University Ghent, Belgium.

Prior to joining Televic in 2003, he was a Transport Network Expert for Hermes Europe Railtel and Tellium. His present research activities include networked audio/video technologies.



Marc Moonen (M'94–SM'06) received the electrical engineering degree and the Ph.D. degree in applied sciences from Katholieke Universiteit Leuven, Belgium, in 1986 and 1990, respectively.

Since 2004, he has been a full Professor in the Electrical Engineering Department, Katholieke Universiteit Leuven, where he is currently heading a research team of 16 Ph.D. candidates and postdoctorals working in the area of signal processing for digital communications, wireless communications, DSL, and audio signal processing. He has been a European Association for Signal, Speech and Image Processing (EURASIP) AdCom Member since 2000. He is currently Editor-in-Chief of the *EURASIP Journal on Applied Signal Processing* (2003–2005). He is a member of the Editorial Board of *Integration: The VLSI Journal* and the *EURASIP Journal on Wireless Communications and Networking*.

Prof. Moonen received the 1994 KU Leuven Research Council Award, the 1997 Alcatel Bell (Belgium) Award (with P. Vandaele), and the 2004 Alcatel Bell (Belgium) Award (with R. Cendrillon). He was a 1997 Laureate of the Belgium Royal Academy of Science. He was Chairman of the IEEE Benelux Signal Processing Chapter (1998–2002). He has been a member of the Editorial Board of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—PART II: ANALOG AND DIGITAL SIGNAL PROCESSING (2002–2003). He is a member of the Editorial Board of *IEEE Signal Processing Magazine*.