A Hands-Free Phone System Based on a Partitioned Frequency-Domain Adaptive Echo Canceler

Pius Estermann
Institute for Signal and Information Processing
Swiss Federal Institute of Technology Zürich
CH – 8092 Zürich Switzerland
Phone: (+41) 1 632 2766   Fax: (+41) 1 632 1208
E-mail: esterman@isi.ee.ethz.ch

August Kaelin
Institute for Signal and Information Processing
Swiss Federal Institute of Technology Zürich
CH – 8092 Zürich Switzerland
Phone: (+41) 1 632 2762   Fax: (+41) 1 632 1208
E-mail: kaelin@isi.ee.ethz.ch

ABSTRACT

Providing means for hands-free conversation is of great interest for industry and is still a current research topic. In this paper, a partitioned frequency-domain adaptive FIR filter is applied in a hands-free phone system to provide echo compensation. It is optimally designed in such a way that it approaches the tracking behavior of the *Recursive Least-Squares* (RLS) algorithm, and it is combined with a new adaptive step-size control in order to cope with varying far-end/local speaker situations. Its performance is demonstrated by means of real speech signals. Assuming a loudspeaker–room–microphone impulse response duration of 3500 taps, an increase in the critical gain of 14dB has been obtained (for each phone) by using an adaptive echo canceler with 1152 taps.

1 INTRODUCTION

Due to their excellent convergence behavior and their computational efficiency, frequency-domain adaptive FIR filters have gained increased attraction [Shy92, Som92]. They are especially promising for real time applications where a large number of coefficients have to be adapted. In this paper, we describe the realization of a robust hands-free phone using a partitioned frequency-domain adaptive FIR filter with a new adaptive step-size control allowing for cancelling of the time-varying coupling between loudspeaker (far-end) signal $x(n)$ and microphone (available) signal $d(n)$, as depicted in Fig. 1. This coupling, including loudspeaker and microphone transfer functions, is referred to as the *loudspeaker–room–microphone system* (LRMS) described by the discrete-time impulse response $h_i(n)$, $0 \leqslant i < \infty$ with n the discrete time variable. The impulse response of a typical LRMS (sampled at 8kHz) is depicted in Fig. 2. In order to cancel most of the coupling, an FIR filter $\hat{h}_i(n)$, $0 \leqslant i < M_C$, of reduced length M_C is adapted to the LRMS. Such an adaptive echo canceler reduces most of the echoes for the far-end speaker. The concept considered here uses an additional *adaptive gain control* (AGC) [Hei95] in order to (i) equalize the time-varying level of the local-speaker signal $w(n)$ (due to his movement in the room), (ii) to suppress local background noise, and (iii) to suppress the remaining residual echoes.

In Section 2 the applied echo canceler is described in some detail. Section 3 continues by giving some implementation details and a design strategy. In section 4, the performance of our echo canceler is discussed by means of its tracking behavior, i.e., the ability to track changes of the unknown LRMS. Finally, Section 5 gives some conclusions.

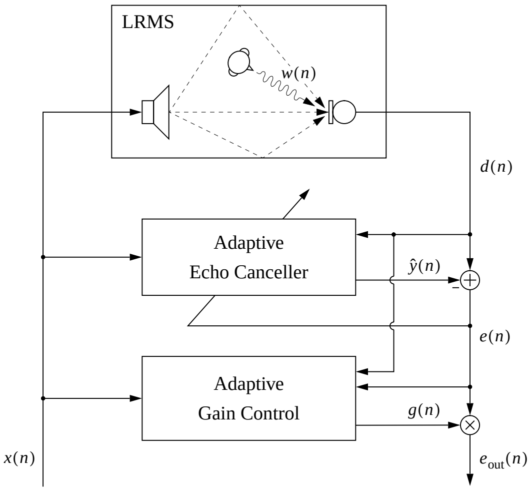

Figure 1: *Basic scheme of the hands-free phone.*

2 ADAPTIVE ECHO CANCELER

In the following we introduce our adaptive echo canceler which is based on the *partitioned frequency-domain least-mean-square* (PFLMS) algorithm [Som92] with a new *adaptive step size control* (ASC), as depicted in Fig. 3. The PFLMS algorithm applies the *overlap-save* technique by segmenting the input signal $x(n)$ and the available signal $d(n)$ into $C-N$ samples overlapping vectors of length C, as given by

$$\underline{x}[k] = [x((k+1)N-C), x((k+1)N-C+1), \ldots, x((k+1)N-1)]^T \quad (1a)$$

$$\underline{d}[k] = [d((k+1)N-C), d((k+1)N-C+1), \ldots, d((k+1)N-1)]^T \ , \quad (1b)$$

with N the block size and k the block index. Next, the impulse response $\hat{h}_i(n)$, $0 \leqslant n < M_C$, is partitioned into L segments each having length mN with m the partitioning parameter, a positive integer fulfilling $C \geqslant (m+1)N-1$. To update these

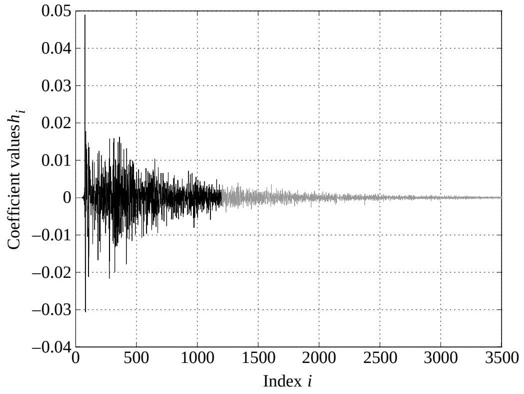

Figure 2: *Example of a typical LRMS impulse response. The adaptively compensated part is depicted in black.*

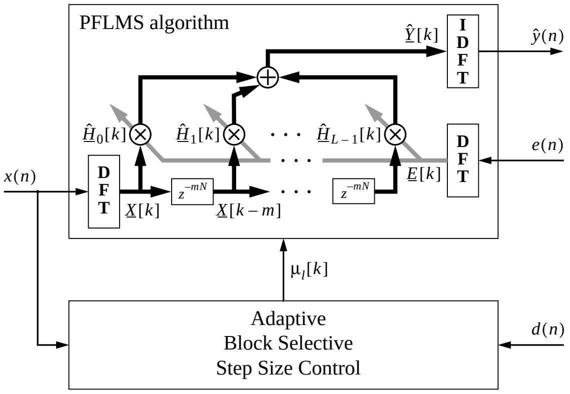

Figure 3: *Scheme of the echo canceler.*

L segments at the discrete time instances $n = kN$, we collect segment l (where $0 \leqslant l < L$) into the following vector:

$$\underline{\hat{h}}_l[k] = \left[\hat{h}_{lmN}(kN), \hat{h}_{lmN+1}(kN), \ldots, \hat{h}_{(l+1)mN-1}(kN), 0, \ldots, 0\right]^T. \tag{2}$$

Padding with zeros has been used to match the length of (2) with the length of the appropriate input vector $\underline{x}[k-lm]$.

The linear convolution $\hat{y}(n) = \hat{h}_i(kN) * x(n)$ can now be written as a sum of cyclic convolutions,

$$\underline{\hat{y}}[k] = \sum_{l=0}^{L-1} \underline{x}[k-lm] \circledast \underline{\hat{h}}_l[k], \tag{3}$$

where the last N values of $\underline{\hat{y}}[k]$ correspond to $\hat{y}(kN)$,$\hat{y}(kN+1),\ldots\hat{y}((k+1)N-1)$.

In obtaining an appropriate frequency-domain description, we now introduce the $C \times C$-point Fourier transform matrix $\boldsymbol{F}$, with elements $F_{ij} = 1/C \exp(-j2\pi/C\, ij)$ and $j = \sqrt{-1}$. For $\underline{x}[k]$ this becomes $\underline{X}[k] = \boldsymbol{F}\underline{x}[k]$. The remaining C-point vectors are assumed to have an equivalent frequency-domain description denoted by their corresponding capital letter. The convolution (3) is now written in the frequency-domain as

$$\underline{\hat{Y}}[k] = \sum_{l=0}^{L-1} \boldsymbol{X}[k-lm]\underline{\hat{H}}_l[k], \tag{4}$$

with the diagonal matrix $\boldsymbol{X}[k]$ having $\underline{X}[k]$ along its diagonal.

Based on these preliminaries the PFLMS algorithm is given as [Est96]

$$\underline{E}[k] = \boldsymbol{F}\boldsymbol{\zeta}^T\boldsymbol{\zeta}\boldsymbol{F}^{-1}\left(\underline{D}[k] - \sum_{l=0}^{L-1}\boldsymbol{X}[k-lm]\underline{\hat{H}}_l[k]\right) \tag{5a}$$

$$\underline{\hat{H}}_l[k+1] = \underline{\hat{H}}_l[k] + \boldsymbol{F}\boldsymbol{\xi}^T\boldsymbol{\xi}\boldsymbol{F}^{-1}\boldsymbol{\mu}_l[k]\boldsymbol{X}^H[k-lm]\underline{E}[k], \tag{5b}$$

for $0 \leqslant l < L$. The projection matrices $\boldsymbol{\zeta}^T\boldsymbol{\zeta}$ and $\boldsymbol{\xi}^T\boldsymbol{\xi}$ are defined as $\boldsymbol{\zeta} = [\boldsymbol{O}_{N\times C-N} \quad \boldsymbol{I}_{N\times N}]$ and $\boldsymbol{\xi} = [\boldsymbol{I}_{mN\times mN} \quad \boldsymbol{O}_{mN\times C-mN}]$. The diagonal step-size matrix $\boldsymbol{\mu}_l[k]$ is defined by its elements

$$\mu_{l,j}[k] = \begin{cases} \eta_l[k]P_{X,j}^{-1}[k-lm] & \text{for } P_{X,j}[k-lm] > c_1\bar{P}_{X,j}[k-lm], \\ 0 & \text{otherwise .} \end{cases} \tag{6}$$

The step-sizes (6) include a normalization to the short-time mean of the powers $X_j[k]$, the elements of $\underline{X}[k]$, given by

$$P_{X,j}[k] = (1-\gamma_1)\left|X_j[k]\right|^2 + \gamma_1 P_{X,j}[k-1], \quad 0 \leqslant j < C, \tag{7}$$

and the adaptive bound

$$\bar{P}_{X,j}[k] = \begin{cases} (1-\gamma_2)\left|X_j[k]\right|^2 + \gamma_2\bar{P}_{X,j}[k-1] & \text{for } \bar{P}_{X,j}[k] > P_{X,j}^{(\min)} \\ P_{X,j}^{(\min)} & \text{otherwise ,} \end{cases} \tag{8}$$

with γ_1 and γ_2 appropriately selected “forgetting” factors ($0 < \gamma_1 < \gamma_2 < 1$) and c_1 a gain factor to be chosen. The frequency selective lower bound $P_{X,j}^{(\min)}$ guarantees that the adaptation is not released during long speech pauses of the far-end speaker.

In order to consider the varying far-end/local speaker activities, we propose to use a segment-dependent switching factor $s_l[k]$ which controls the freezing of the adaptation. Based on the energy of the input signal vector

$$P_x[k] = \underline{x}^T[k]\underline{x}[k], \tag{9}$$

and the relevant energy of the available signal vector

$$P_d[k] = \underline{d}^T[k]\boldsymbol{\zeta}^T\boldsymbol{\zeta}\underline{d}[k], \tag{10}$$

it is given as follows:

$$s_l[k] = \begin{cases} 1 & \text{for } P_x[k-lm] > c_2\bar{P}_x[k-lm] \;\cap \\ & \qquad P_x[k-lm] > r_l P_d[k] \\ 0 & \text{otherwise ,} \end{cases} \tag{11}$$

with

$$\bar{P}_x[k] = \begin{cases} (1-\gamma_3)P_x[k] + \gamma_3\bar{P}_x[k-1] & \text{for } \bar{P}_x[k] > P_x^{(\min)} \\ P_x^{(\min)} & \text{otherwise .} \end{cases} \tag{12}$$

This switching factor allows the adaptation of the lth segment only if (i) its appropriate input vector $\underline{x}[k-lm]$ contains sufficient energy, if compared to the averaged energy (12) multiplied by the gain factor c_2 [Mar96, p. 87], and (ii) if this input energy exceeds the relevant energy of the available signal vector $\underline{d}[k]$ multiplied by the threshold parameter r_l.

Relation (11) reliably freezes the adaptation if a strong local speaker is present. If no local speaker is present, the echo canceler can be adapted with maximum step size $\eta = \eta_{\max}$. Assuming a weak local speaker (or background noise) the tracking behavior can be considerably improved if the adaptation is not frozen, but performed with a reduced step size $\eta = c_3\eta_{\max}$.

If the correlation between the input signal $x(n)$ and the available signal $d(n)$ exceeds a certain value, the likelihood of the presence of a weak local speaker is low. We therefore use an appropriate correlation measure as proposed in [Hei95]. It is computed by evaluating

$$\rho_{xd}[k,v] = \frac{\displaystyle\sum_{i=0}^{I-1} x((k+1)N-1-i-v)d((k+1)N-1-i)}{\displaystyle\sum_{i=0}^{I-1} |x((k+1)N-1-i-v)d((k+1)N-1-i)|} \tag{13}$$

in the range $0 < v \leqslant V$ with V and I selected appropriately for a given LRMS. Its maximum value

$$\rho_{xd}[k] = \max_{v}\left(|\rho_{xd}[k,v]|\right) , \tag{14}$$

together with (11) is now used to determine the normalized step size as

$$\eta_l[k] = \begin{cases} s_l[k]\eta_{\max} & \text{for } \rho_{xd}[k] > \rho_0 \\ s_l[k]c_3\eta_{\max} & \text{otherwise ,} \end{cases} \tag{15}$$

with ρ_0 to be chosen experimentally.

Finally we mention that if the number of segments L, is increased, the projection $\boldsymbol{F}\boldsymbol{\xi}^T\boldsymbol{\xi}\boldsymbol{F}^{-1}$ in (5b) becomes the dominant part of the computational complexity of the PFLMS algorithm. Assuming a reasonable (normalized) step-size $\eta_{\max}$, it is sufficient to apply the projection only every Lth block index k. In doing so, no recognizable distortion of the tracking behavior has been observed.

3 IMPLEMENTATION

The adaptive echo canceler, as described in Section 2, has been implemented together with an AGC based on [Hei95]. It is noted that, compared to [Hei95], the computational complexity of the AGC has been reduced drastically by incorporating a similar block processing technique as discussed above for the echo canceler [MvH96].

The PFLMS algorithm (5) has several parameters which have to be optimally selected for a given LRMS. The block-size parameter N determines the minimum additional delay introduced in the signal path. A small N results in an increased computational complexity of the echo canceler.

The DFT length C, corresponding to the length of the input vectors, and usually a power of two, should be selected so that on the one hand the minimum computational complexity is obtained (small C value), and on the other hand so that the speech signal $x(n)$ is "sufficiently" decorrelated (large C value). This "sufficient" decorrelation results in a tracking behavior which approaches the one of the RLS algorithm [EK95] in a given LRMS environment. With $C \geqslant (m+1)N-1$, we obtain for a computationally efficient realization $m = \lfloor (C+1)/N \rfloor - 1$ for the partioning parameter.

Measurements have shown that for the usual changes in a LRMS (caused by moving people, displaced chairs etc.), the segment energies $\underline{h}_l^T[k]\underline{h}_l[k]$ (where $\underline{h}_l[k]$ is defined equivalently to $\hat{\underline{h}}_l[k]$) remain nearly constant provided that loudspeaker and microphone are placed at fixed positions [GH95]. This allows the threshold parameters r_l to be selected optimally, if an estimate $\hat{\underline{h}}_l[0]$ is initially available. Such an estimate might be obtained by injecting white noise as the far-end speaker signal $x(n)$. Based on such an estimate, our choice for the threshold parameters is to select them proportionally to their appropriate segment energies, i.e.,

$$r_l = \frac{c_4}{\hat{\underline{h}}_l^T[0]\hat{\underline{h}}_l[0]} , \quad 0 \leqslant l < L . \tag{16}$$

Finally, the optimum length of the echo canceler, M_C, has to be determined. Based on the fact that the segment energies decay approximately exponentially with increasing l, the appropriate switching factors (11) are more and more likely to freeze the adaptation of $\hat{\underline{H}}_l[k]$. Thus, we conclude that there exists a reduced echo canceler length M_C with a corresponding number of segments L (i.e., $M_C = LmN$). Further increases in M_C do not improve the tracking behavior of the echo canceler substantially.

The proposed hands-free phone system has been implemented using one processor (ADSP-21020) for each phone and a sample rate of 8kHz. Linear 16bit converters have been used for both the loudspeaker and the microphone signal. Assuming it is permitted to insert a total delay of 16ms into the signal path for each phone, we selected a block size of $N = 64$ samples allowing 8ms for collecting a new data block and 8ms for processing the data. To perform a "sufficient" decorrelation of the input signal in order to achieve the optimum tracking behavior, we selected the DFT length as $C = 256$ points.

4 RESULTS

To demonstrate the obtained tracking behavior, 16 seconds of a recorded speech conversation has been used on a LRMS as depicted in Fig. 2. Its far-end speaker signal $x(n)$ and its local-speaker signal $w(n)$ are depicted in Fig. 4a and Fig. 4b, respectively. To discuss the tracking behavior, we use the LRMS mismatch defined by

$$J[k] = \sum_{i=0}^{LmN-1} \left(\hat{h}_i(kN) - h_i(kN)\right)^2 + \sum_{i=LmN}^{\infty} h_i^2(kN)\,, \qquad (17)$$

where, for practical reasons, "∞" in the last sum has to be replaced by some reasonable finite value.

For the simulation, the echo canceler has been initialized with $\underline{\hat{h}}_l[0]$, $0 \leqslant l < L$. After 11 seconds (far-end speaker present), we changed the LRMS by displacing a chair in the room. The obtained mismatch $J[k]$, as depicted in Fig. 4c, demonstrates the proper tracking behavior for different L. We emphasize that the choice $L = 6$ segments seems to be the optimum selection for the here considered LRMS. A further increase seems to not yield further improvements of the tracking behavior.

To evaluate the critical gain of the overall system (two connected phones), we replaced the AGC in the simulation with the constant gain factor g. Assuming a far-end speaker signal $x(n)$ and a change of the local LRMS as discussed above (constant far-end LRMS), an increase of the critical gain of 11.7dB has been obtained. Without any change in the LRMS, a 14dB increase has been observed.

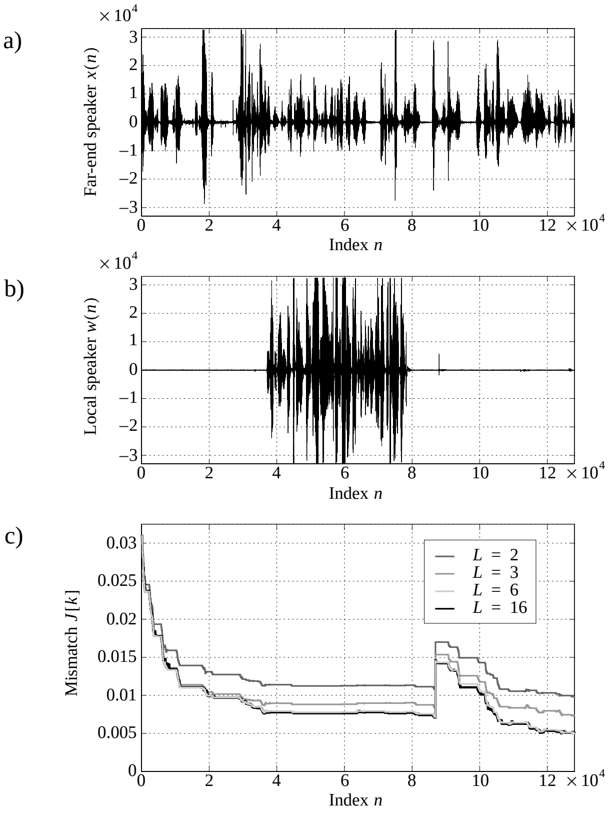

Figure 4: *Tracking behavior of the echo canceler: a) far-end speaker signal $x(n)$, b) local speaker signal $w(n)$, c) achieved mismatch $J[k]$ for different echo canceler lengths $M_C = LmN$.*

5 CONCLUSIONS

A new echo canceler consisting of a PFLMS algorithm with an incorporated adaptive step-size control has been proposed and implemented in a hands-free phone, together with an adaptive gain control. Due to the block processing of the PFLMS algorithm, our realization is computationally very efficient. With a sufficient length C of the DFT, the obtained tracking behavior approaches the one of the optimum RLS algorithm. The incorporated partitioning of the modeled LRMS allows for separate step-size controls in adapting each of the segments of the impulse response of the LRMS. This results in a good tracking behavior for actual non-stationary far-end/local speaker activities. In our implementation the additional input-output delay of the local speaker's signal is 16ms. It is noted that this delay could be reduced to one single sample if the first two blocks ($2N$ samples) were realized in the time-domain, whereas the actual adaptation were still performed in the frequency-domain. While exhibiting the same tracking behavior, this variant of the PFLMS algorithm would require increased computational complexity, which however, is still lower than the one of its comparable time-domain LMS algorithm.

6 ACKNOWLEDGMENTS

The authors are grateful to Prof. Dr. E. Hänsler for many fruitful discussions on the subject. Also, the authors would like to acknowledge the work of the students S. Marcoli and Th. von Hoff [MvH96], as well as A. Ganz and D. Huber [GH95]. The proposed realization of a hand-free phone system is partly based on results of their diploma theses.

References

[EK95] Pius Estermann and August Kaelin. On the comparison of optimum least-squares and computationally efficient DFT-based adaptive block filters. In *Proc. IEEE ISCAS, Seattle, USA*, volume 3, pages 1612–1615, 1995.

[Est96] Pius Estermann. *Adaptive Filter im Frequenzbereich: Analyse und Entwurfsstrategie*. PhD thesis, Swiss Federal Institute of Technology, Zürich, Switzerland, in preparation.

[GH95] Adrian Ganz and Dieter Huber. Kompensation akustischer Rückkopplung in Freihandtelefonen III. Diploma thesis, Signal and Information Processing Laboratory, ETH Zürich, 1995.

[Hei95] Peter Heitkämper. *Freisprechen mit Verstärkungssteuerung und Echokompensation*, volume 380 of *Fortschrittberichte VDI, Reihe 10: Informatik/Kommunikationstechnik*. VDI Verlag, 1995.

[Mar96] Jochen Marx. *Akustische Aspekte der Echokompensation in Freisprecheinrichtungen*, volume 400 of *Fortschrittberichte VDI, Reihe 10: Informatik/Kommunikationstechnik*. VDI Verlag, 1996.

[MvH96] Sandro Marcoli and Thomas von Hoff. Kompensation akustischer Rückkopplung in Freihandtelefonen IV. Diploma thesis, Signal and Information Processing Laboratory, ETH Zürich, 1996.

[Shy92] John J. Shynk. Frequency-domain and multirate adaptive filtering. *IEEE SP Magazine*, pages 14–37, January 1992.

[Som92] P.C.W. Sommen. *Adaptive Filtering Methods*. PhD thesis, Technical University Eindhoven, 1992.