ACOUSTIC SIGNAL PROCESSING FOR DISTANT-TALKING SPEECH RECOGNITION: NONLINEAR ECHO CANCELLATION IN A GENERIC MULTICHANNEL INTERFACE

Fabian Küch, Walter Kellermann, Wolfgang Herbordt, and Herbert Buchner *
Chair for Multimedia Communications and Signal Processing, University Erlangen-Nuremberg
{kuech,wk,herbordt,buchner}@LNT.de

ABSTRACT

In this contribution we describe full-duplex communication systems in a general framework using linear matrix formulations. Concentrating on the signal acquisition part echo cancellation, noise suppression and signal separation are identified as prevalent challenges. This framework is then extended to compensate for nonlinear acoustic echo paths and an efficient DFT-domain adaptive second-order Volterra filter based on partitioned block techniques is presented.

1 INTRODUCTION

In speech dialogue systems, reliable automatic speech recognition (ASR) for distant talkers in noisy and echoic environments is one of the main challenges on the way to seamless and natural human/machine interaction. Two classes of approaches are common and often combined for real systems: One class aims at enhancement of the recorded acoustic signal, whereas the other class concentrates on removing detrimental effects on the feature level and exploiting higher-level a priori knowledge to achieve robust ASR performance.

We consider here acoustic signal enhancement for a full-duplex human-machine interface after Fig. 1 where we allow for multichannel reproduction using loudspeaker arrays and for multichannel recording using microphone arrays, respectively. While most other components can be well modeled as linear systems, loudspeaker systems often involve nonnegligible nonlinearities, e.g. caused by overloaded amplifiers or nonlinearities in the electroacoustic transduction as common with low-cost loudspeakers driven at high volume.

In this contribution, we first describe the fundamental problems for acoustic signal acquisition in a general multichannel setup. Then, we extend this framework to include nonlinearities of loudspeaker systems. This leads to an efficient DFT-domain algorithm for acoustic echo cancellation for nonlinear echo path models, which can well be integrated with other state-of -the-art techniques for acoustic signal enhancement.

*This work was partly supported by grants from Intel Corp. (Beijing, China, Santa Clara, CA), from Grundig AG, Nuremberg, leading the German EMBASSI consortium, and from the European Commission as sponsor of the ANITA project.

2 MULTICHANNEL (ACOUSTIC) INTERFACES

From Fig.1 it is obvious that, ideally, the full-duplex communication system $\mathbf{G}$ processes the source signals $\mathbf{u}$ and the sensor signals $\mathbf{x}$ such that $\mathbf{w}$ corresponds to a desired sound impression $\mathbf{w}_\mathrm{d}$ at the listeners' ears and such that the output vector $\mathbf{z}$ consists of $P \leq M$ desired signals, respectively. Note that current speech recognition systems disregard all but one element of $\mathbf{z}$, which may change in some future applications.

Assuming for now that the matrices $\mathbf{H_{wv}}$, $\mathbf{H_{xv}}$, and $\mathbf{H_{xs}}$ can be modeled as linear discrete-time systems, $\mathbf{G}$ may also be linear and can be completely described by linear convolutions[1] as

$$\begin{pmatrix} \mathbf{v} \\ \mathbf{z} \end{pmatrix} = \mathbf{G} * \begin{pmatrix} \mathbf{u} \\ \mathbf{x} \end{pmatrix} = \begin{pmatrix} \mathbf{G_{vu}} & \mathbf{G_{vx}} \\ \mathbf{G_{zu}} & \mathbf{G_{zx}} \end{pmatrix} * \begin{pmatrix} \mathbf{u} \\ \mathbf{x} \end{pmatrix}, \quad (1)$$

where the submatrices $\mathbf{G_{vu}}$, $\mathbf{G_{vx}}$, $\mathbf{G_{zu}}$, and $\mathbf{G_{zx}}$ describe the signal processing between the respective signal vectors. The elements of the matrices $\mathbf{H}_*$ describe impulse responses of the acoustic environment with usually large numbers of filter taps (hundreds or thousands). We emphasize that, although we assume linear processing of the signals $\mathbf{x}, \mathbf{u}$ to produce $\mathbf{z}$, the actual signal processing for identifying the generally time-variant elements of $\mathbf{G}$ may of course be highly nonlinear.

Using this matrix description, we can derive conditions for the elements of $\mathbf{G}$. With ASR in mind, we limit ourselves here to the recording part, i.e., to signal processing for obtaining the desired vector $\mathbf{z}$. Thus, we disregard the special problems of sound reproduction for providing a desired acoustic impression to the local listener(s) (see [1]) involving, e.g., wavefield synthesis, although this can

[1]The linear convolution $\mathbf{y} = \mathbf{A} * \mathbf{x}$ between a column vector $\mathbf{x}$ with elements $x_i(k)$ and a matrix $\mathbf{A}$ with time-invariant impulse responses $a_{ij}(k)$ is defined by $y_i(k) = \sum_{j=1}^{N} \sum_{n=-\infty}^{\infty} a_{ij}(k-n)x_j(n)$, where $y_i(k)$ is the i-th component of $\mathbf{y}$. The inverse $\mathbf{A}^{-1}$ is defined to fulfill $\mathbf{A}^{-1} * \mathbf{A} = \mathbf{I} \cdot \delta(k)$, where $\mathbf{I}$ is the identity matrix and where $\delta(k)$ is the discrete-time unit impulse. If $\mathbf{A}$ is not invertible, then, $\mathbf{A}^{-1}$ is the pseudoinverse of $\mathbf{A}$.

also be relevant in applications of ASR, e.g., in interactive gaming.

We note that the recorded signal vector capturing N microphone signals, $\mathbf{x}$, can be written as

$$\mathbf{x} = \mathbf{H_{xs}} * \mathbf{s} + \mathbf{H_{xv}} * \mathbf{v} + \mathbf{n_x}, \tag{2}$$

where $\mathbf{H_{xv}}$ and $\mathbf{H_{xs}}$ describe the multiple-input/multiple-output (MIMO) systems between the respective signal vectors, and $\mathbf{n_x}$ represents the components originating from undesired local interferers and noise sources.

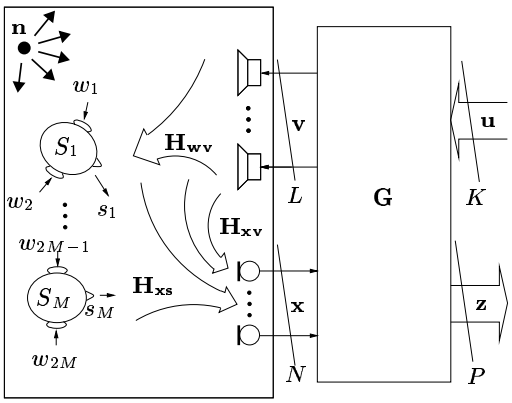


Figure 1: Full-duplex communication system.

Multichannel sound recording aims at the extraction of P desired signals $\mathbf{z}$ from $\mathbf{x}$, which can generally be described as output of a desired mixing matrix $\mathbf{D_{zs}}$ with the original $M \geq P$ source signals $\mathbf{s}$ as input. With (1), (2) we obtain

$$\begin{aligned} \mathbf{z} &= \mathbf{G_{zu}} * \mathbf{u} + \mathbf{G_{zx}} * \mathbf{x} \\ &= \mathbf{G_{zx}} * \mathbf{H_{xs}} * \mathbf{s} + (\mathbf{G_{zu}} + \mathbf{G_{zx}} * \mathbf{H_{xv}} * \mathbf{G_{vu}}) * \mathbf{u} \\ &\quad + \mathbf{G_{zx}} * \mathbf{n_x} \\ &\stackrel{!}{=} \mathbf{D_{zs}} * \mathbf{s}. \end{aligned} \tag{3}$$

This formulation defines three signal processing problems: cancellation of acoustic echoes resulting from $\mathbf{u}$, suppression of noise and interference $\mathbf{n_x}$, and source signal separation and deconvolution to obtain $\mathbf{D_{zs}} * \mathbf{s}$.

A major structural difference compared to a single-channel recording setup is given by the fact that sampling the acoustic wavefield by several microphones provides spatial information and allows spatially selective filtering. This can be immediately exploited for separating or suppressing the corresponding signals if a priori knowledge on the source signals is available.

2.1 Acoustic echo cancellation (AEC)

For compensating the feedback from the loudspeakers to the microphones independently of the signals $\mathbf{u}$, the MIMO system $\mathbf{G_{zu}}$ must ideally meet (see (3))

$$\mathbf{G_{zu}} \stackrel{!}{=} -\mathbf{G_{zx}} * \mathbf{H_{xv}} * \mathbf{G_{vu}}. \tag{4}$$

As the $\mathbf{G_{zx}}, \mathbf{G_{vu}}$ are observable, only $\mathbf{H_{xv}}$ must be identified. With monaural sound reproduction, AEC has to cope with long adaptive filters and the fact that the residual echo is not always observable as optimization error. Aside from the further increased computational complexity, the multichannel is even more difficult because of the correlation between the signals in $\mathbf{u}$ [2].

2.2 Noise and interference suppression

For suppression of local noise and interferers, the condition

$$\mathbf{G_{zx}} * \mathbf{n_x} \stackrel{!}{=} \mathbf{0} \tag{5}$$

must be fulfilled. Obviously, the signal-independent solution $\mathbf{G_{zx}} = \mathbf{0}$ prevents recording of the desired signals. Beamforming methods, however, lead to spatial filtering by $\mathbf{G_{zx}}$ so that diffuse noise and coherent interference $\mathbf{n_x}$ arriving from certain directions are suppressed, whereas desired signals from other directions remain undistorted. Moreover, adaptive beamforming can track changes of directions of arrival and signal statistics [3].

2.3 Source separation and dereverberation

To extract the signals of interest $\mathbf{s}$ in the desired form, the sensor signals need to be processed such that

$$\mathbf{G_{zx}} * \mathbf{H_{xs}} * \mathbf{s} \stackrel{!}{=} \mathbf{D_{zs}} * \mathbf{s}, \tag{6}$$

which requires

$$\mathbf{G_{zx}} = \mathbf{D_{zs}} * \mathbf{H_{xs}^{-1}} \tag{7}$$

for signal-independent solutions. Asking for undistorted source signals in $\mathbf{z}$, $\mathbf{D_{zs}}$ can be written as a diagonal matrix with P delayed unit impulses along the main diagonal. Then, (7) corresponds to a multichannel blind inversion problem ('dereverberation') for the elements on the main diagonal and an interference suppression problem for the elements on the off-diagonals of $\mathbf{G_{zx}} * \mathbf{H_{xs}}$, respectively.

Relaxing the requirements for $\mathbf{D_{zs}}$, blind source separation (BSS) is able to extract coherent signals $\mathbf{z} = \widehat{\mathbf{s}}$ by exploiting statistical independence or uncorrelatedness of the sources in conjunction with non-stationarity and non-whiteness assumptions [4]. Thereby, the off-diagonal elements of $\mathbf{G_{zx}} * \mathbf{H_{xs}}$ are minimized, while the elements on the main diagonal are usually not just delays, so that dereverberation is at best approximated.

3 NONLINEAR ACOUSTIC ECHO CANCELLATION

With nonlinearities in loudspeaker systems, the performance of linear AEC degrades, and high-level echo bursts are distorting the ASR input during high-volume announcements [5]. Therefore, it is desirable to extend the above concept to include nonlinear AEC. For clarity and notational convenience, we consider (4) for the case of monaural reproduction and acquisition ($L = N = 1$), and disregard sound rendering and the signal enhancement ($\mathbf{G_{vu}} = \mathbf{G_{zx}} = 1$). For a linear echo path (4) then

reads $g_{zu} = g_{xu} = g_{xv} = -h_{xv}$, and AEC has to form an estimate $\widehat{x}(k) = g_{xu} * u$ to compensate $h_{xv} * u$.

A common approach to modelling the nonlinear behaviour of loudspeakers is given by finite-length second-order Volterra filters [5, 6]. Volterra filters can be considered as multiple-input/single-output (MISO) systems, where the input signals result from products of samples of $u(k)$ taken at different time instants. Abbreviating $g := g_{zu}$, the output of the second-order Volterra filter can be written as

$$\widehat{x}(k) = \mathbf{g} * \mathbf{u}'. \tag{8}$$

where

$$\begin{aligned} \mathbf{g} \quad = \quad & \Big[g^{(1)}(k),\, g^{(2)}(k,k),\, g^{(2)}(k,k+1), \ldots, \\ & g^{(2)}(k,k+N^{(2)}-1)\Big] \end{aligned} \tag{9}$$

contains the linear kernel $g^{(1)}(k)$ with $0 \le k < N^{(1)}$ and the quadratic kernel $g^{(2)}(k,k')$ with $0 \le k,k' < N^{(2)}$. The augmented input vector $\mathbf{u}'$ reads

$$\begin{aligned} \mathbf{u}' \quad = \quad & [u(k),\, u(k)u(k),\, u(k)u(k-1), \ldots, \\ & u(k)u(k-N^{(2)}+1)\Big]^T. \end{aligned} \tag{10}$$

The compact form of (8) shows the relation to the general linear system description (1): Including nonlinear echo cancellation corresponds to an extension of the input vector $\mathbf{u}$, with the additional components being only used for AEC. (The corresponding elements of the sound rendering matrix $\mathbf{G_{vu}}$ must be set to zero.)

For deriving a new efficient realization of such a second -order Volterra filter, we rewrite (8) as

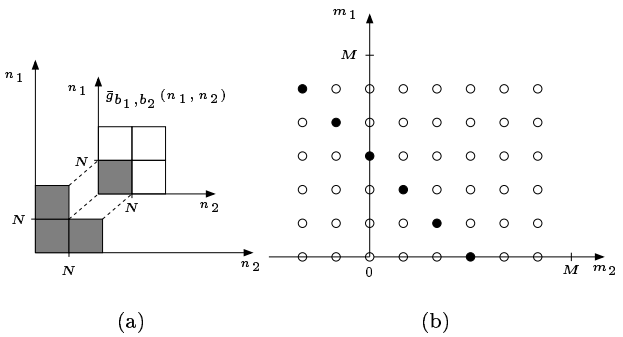
$$\begin{aligned} \widehat{x}(k) = & \sum_{n=0}^{N^{(1)}-1} g^{(1)}(n)u(k-n) \\ & + \sum_{n_1=0}^{N^{(2)}-1} \sum_{n_2=0}^{N^{(2)}-1} \tilde{g}^{(2)}(n_1,n_2)u(k-n_1)u(k-n_2), \end{aligned} \tag{11}$$

where $\tilde{g}^{(2)}(n_1,n_2) = g^{(2)}(n_1,n_2)$, for $n_1 = n_2$, and $\tilde{g}^{(2)}(n_1,n_2) = \frac{1}{2}g^{(2)}(n_1,n_2)$, otherwise.

3.1 Partitioned Block Frequency Domain Adaptive Volterra filter (PBFDAVF)

Efficient DFT-domain adaptive filtering algorithms assuring fast convergence to the optimum solution and sufficient robustness to cope with real-world scenarios are known for multichannel linear acoustic echo cancellation, adaptive beamforming, and blind source separation, respectively. For efficient combination, partitioned-block versions with scalable overlap were developed so that different FIR filter lengths are supported simultaneously for the different algorithms [7, 8, 9]. Correspondingly, a DFT-domain realization of adaptive second-order Volterra filters is desirable for its advantages regarding computational complexity and convergence behaviour.

Known DFT-domain approaches [10, 11] suffer from the inherent restriction that the memory length of the Volterra kernels can not be chosen differently for different orders. However, for modeling the nonlinear behavior of loudspeakers using Volterra filters it has been shown that the required memory length for the linear kernel is much larger than that of the quadratic kernel [5]. Therefore, we propose a new DFT-domain algorithm based on partitioned-block modeling which allows a more flexible choice with respect to the memory length of kernels. This algorithm can be viewed as a generalization of both, the frequency-domain adaptive Volterra filter [11] and the partitioned block frequency domain adaptive filter (PBFDAF,GMDF) for linear systems [12].



(a) (b)

Figure 2: Illustration of the zero-padding in $\bar{g}^{(2)}_{b_1,b_2}(n_1,n_2)$ (a) and the computation of $\widehat{X}^{(2)}_{b_1,b_2,\nu}(m)$ for $M = 6$ and $m = 3$ (b).

3.1.1 Partitioned Block Frequency-Domain Volterra Filter

We assume in the following that $N^{(1)}$ and $N^{(2)}$ are integer multiples of the so-called partition length N, i.e. $N^{(1)} = B^{(1)}N$, $N^{(2)} = B^{(2)}N$, where $N, B^{(1)}, B^{(2)} \in \mathbb{N}$. First, we define overlapping input signal data blocks of length M according to

$$u_{i,\nu}(l) = u(\nu\frac{N}{\alpha} + l - (i+1)N),\ 0 \le l < M, \tag{12}$$

where α is an overlap factor. Aiming at an efficient overlap/save method [13] to compute the output of the Volterra filter according to (11), we introduce zero-padded versions of partitions of the linear and quadratic kernels:

$$\bar{g}^{(1)}_b(n) \quad = \quad \begin{cases} g^{(1)}(n+bN), & 0 \le n < N, \\ 0, & N \le n < M \end{cases} \tag{13}$$

$$\bar{g}^{(2)}_{b_1,b_2}(n_1,n_2) = \begin{cases} \tilde{g}^{(2)}(n_1+b_1N, n_2+b_2N), & 0 \le n_1,n_2 < N, \\ 0, & N \le n_1,n_2 < M. \end{cases} \tag{14}$$

The definition of the zero-padded partitions $\bar{g}^{(2)}_{b_1,b_2}(n_1,n_2)$ according to (14) is illustrated in Fig. 2a for $B^{(2)} = 2$ and $b_1 = b_2 = 1$. Note that only the shaded areas contain nonzero coefficients. Introducing the signal blocks and the kernel partitions (12), (13) and (14) into (11),

respectively, yields $\widehat{x}_\nu(l) = \widehat{x}_\nu^{(1)}(l) + \widehat{x}_\nu^{(2)}(l)$ with

$$\widehat{x}_\nu^{(1)}(l) = \sum_{b=0}^{B^{(1)}-1} \sum_{n=0}^{M-1} \bar{g}_b^{(1)}(n) u_{b,\nu}(l-n), \quad (15)$$

$$\widehat{x}_\nu^{(2)}(l) = \sum_{b_1=0}^{B^{(2)}-1} \sum_{b_2=0}^{B^{(2)}-1} \widehat{x}_{b_1,b_2,\nu}^{(2)}(l), \quad (16)$$

$$\widehat{x}_{b_1,b_2,\nu}^{(2)}(l) = \sum_{n_1=0}^{M-1} \sum_{n_2=0}^{M-1} \bar{g}_{b_1,b_2}^{(2)}(n_1,n_2) \cdot u_{b_1,\nu}(l-n_1) u_{b_2,\nu}(l-n_2). \quad (17)$$

The DFT-domain representation of (15) reads

$$\widehat{X}_\nu^{(1)}(m) = \sum_{b=0}^{B^{(1)}-1} \bar{G}_b^{(1)}(m) U_{b,\nu}(m), \quad (18)$$

where $U_{b,\nu}$ and $\bar{G}_b^{(1)}(m)$ denote the DFT of $u_{b,\nu}(l)$ and $\bar{g}_b^{(1)}(n)$, respectively. Following [14], the DFT-domain representation of (17) can be written as

$$\widehat{X}_{b_1,b_2,\nu}^{(2)}(m) = \frac{1}{M} \sum_{m_1=0}^{M-1} \bar{G}_{b_1,b_2}^{(2)}(m_1, m-m_1) \cdot U_{b_1,\nu}(m_1) U_{b_2,\nu}(m-m_1), \quad (19)$$

where $\bar{G}_{b_1,b_2}^{(2)}(m_1,m_2)$ is the two-dimensional DFT of $\bar{g}_{b_1,b_2}^{(2)}(n_1,n_2)$. The computation of $\widehat{X}_{b_1,b_2,\nu}^{(2)}(m)$ is illustrated in Fig. 2b for $M = 6$. The bin pairs (m_1, m_2) that have to be considered for summation in (19) are marked by $\bullet$ for $m = 3$, which shows that the summation in (19) is performed along the line $m_1 + m_2 = m$ in the (m_1, m_2)-plane. Due to the linearity of the DFT, $\widehat{x}_\nu^{(2)}(l)$ according to (17) reads in the DFT domain:

$$\widehat{X}_\nu^{(2)}(m) = \sum_{b_1=0}^{B^{(2)}-1} \sum_{b_2=0}^{B^{(2)}-1} \widehat{X}_{b_1,b_2,\nu}^{(2)}(m). \quad (20)$$

The IDFT of $\widehat{X}_\nu(m) = \widehat{X}_\nu^{(1)}(m) + \widehat{X}_\nu^{(2)}(m)$ is denoted by $\breve{x}_\nu(l)$. Finally, we obtain the output sequence $\widehat{x}(k)$ according to (11) by applying the overlap/save method, i.e., the first N elements of $\breve{x}_\nu(l)$ are discarded and we set

$$\widehat{x}(k) = \breve{x}_\nu(k - \nu N + N), \quad (21)$$

for $\nu N \leq k \leq (\nu-1)N + M - 1$, i.e., for the last $M - N$ elements of $\breve{x}_\nu(l)$.

3.1.2 *Adaptation of the PBFDAVF*

Assuming that local noise $\mathbf{n_x}$ and desired sources can be disregarded, AEC aims at minimizing the power of the residual echo $e(k) = x(k) - \widehat{x}(k)$, where $e(k) = z(k)$ if $n_x(k) = s(k) = 0$. The update equations for the DFT-domain kernel coefficients $\bar{G}_b^{(1)}(m)$ and $\bar{G}_{b_1,b_2}^{(2)}(m_1,m_2)$ that are presented in the following preserve the time-domain constraints on the kernel coefficients according to (13) and (14). Analogously to the linear constrained partitioned block FDAF [12] we define

$$\widetilde{z}_\nu(l) = \begin{cases} 0, & 0 \leq n < N, \\ x_\nu(l) - \widehat{x}_\nu(l), & N \leq n < M. \end{cases} \quad (22)$$

The DFT-domain representation of $\widetilde{z}_\nu(l)$ is denoted by $\widetilde{Z}_\nu(m)$. Following [12] we introduce the time-domain update term

$$\Delta g_b^{(1)}(n) = r^{(1)}(n)\, \mathcal{F}_{1\mathrm{D}}^{-1} \left\{ \widetilde{Z}_\nu(m) U_{b,\nu}^*(m) \right\}, \quad (23)$$

where $\mathcal{F}_{1\mathrm{D}}^{-1}\{\cdot\}$ denotes the one-dimensional (1D) inverse DFT. Furthermore, we have introduced the 1D window function

$$r^{(1)}(n) = \begin{cases} 1, & 0 \leq n < N, \\ 0, & N \leq n < M. \end{cases} \quad (24)$$

The DFT-domain update equation is then obtained by the 1D-DFT of $\Delta g_b^{(1)}(n)$, i.e.,

$$G_{b,\nu+1}^{(1)}(m) = G_{b,\nu}^{(1)}(m) + \mu\, \Delta G_{b,\nu}^{(1)}(m), \quad (25)$$

with the step-size parameter $\mu > 0$ and

$$\Delta G_{b,\nu}^{(1)}(m) = \mathcal{F}_{1\mathrm{D}} \left\{ \Delta g_b^{(1)}(n) \right\}. \quad (26)$$

Taking the results of [11] into account, we introduce a time-domain windowed update term of the quadratic kernel coefficients according to

$$\Delta g_{b_1,b_2\nu}^{(2)}(n_1,n_2) = r^{(2)}(n_1,n_2) \frac{1}{M} \mathcal{F}_{2\mathrm{D}}^{-1} \left\{ \widetilde{Z}_\nu(m) U_{b_1,\nu}^*(m_1) U_{b_2,\nu}^*(m_2) \right\}, \quad (27)$$

where $m = (m_1 + m_2)$ modulo M. $\mathcal{F}_{2\mathrm{D}}^{-1}\{\cdot\}$ denotes the 2D-IDFT and $r^{(2)}(n_1,n_2)$ represents the 2D window function

$$r^{(2)}(n_1,n_2) = \begin{cases} 1, & 0 \leq n_1, n_2 < N, \\ 0, & N \leq n_1, n_2 < M. \end{cases} \quad (28)$$

The desired constrained update equation for the 2D-DFT domain coefficients finally reads

$$G_{b_1,b_2,\nu+1}^{(2)}(m_1,m_2) = G_{b_1,b_2,\nu}^{(2)}(m_1,m_2) + \mu\, \Delta G_{b_1,b_2,\nu}^{(2)}(m_1,m_2), \quad (29)$$

where

$$\Delta G_{b_1,b_2,\nu}^{(2)}(m_1,m_2) = \mathcal{F}_{2\mathrm{D}} \left\{ \Delta g_{b_1,b_2\nu}^{(2)}(n_1,n_2) \right\}, \quad (30)$$

and where $\mathcal{F}_{2\mathrm{D}}\{\cdot\}$ represents the 2D-DFT. [2]

[2] Note that for the special case $B^{(1)} = B^{(2)} = 1$ (implying $N^{(1)} = N^{(2)}$) the adaptation of the kernel coefficients according to (25) and (30) simplifies to the approach [11].

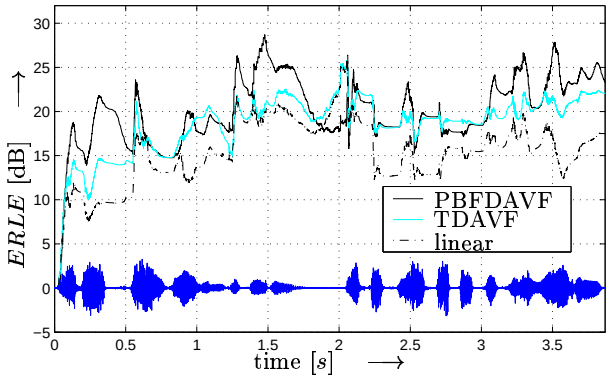


Figure 3: Comparison of different nonlinear approaches and a linear NLMS for a realistic AEC scenario.

3.1.3 Simulation Results

Our evaluation of the above PBFDAVF is based on recorded speech data from a low-cost loudspeaker placed in an enclosure with low reverberation. As performance measure we use the *Echo Return Loss Enhancement* ($ERLE$) which is defined by

$$ERLE = 10\log_{10}\frac{\mathrm{E}\left\{x^2(k)\right\}}{\mathrm{E}\left\{e^2(k)\right\}}[\mathrm{dB}]. \qquad (31)$$

The parameters of the PBFDAVF have been chosen to $N = 64$, $M = 2N$, $B^{(2)} = 1$ and $B^{(1)} = 6$, respectively, and an overlap factor $\alpha = 4$ has been used. The resulting $ERLE$ curve is compared to a time-domain adaptive Volterra filter (TDAVF) with memory lengths $N^{(1)} = B^{(1)}N$ and $N^{(2)} = B^{(2)}N$, applying a normalized least mean square (NLMS) algorithm [15], and to a linear time-domain echo canceler using NLMS adaptation in Fig. 3. We notice that extending the linear AEC to a second-order Volterra filter leads to an improved performance if the nonlinear distortion in the echo path is caused by loudspeakers and that the proposed PBFDAVF provides a faster convergence speed compared to a corresponding time-domain approach.

4 Conclusion

We discussed the general configuration of multichannel communication systems by using a compact linear matrix representation. For the case that the echo path to be modeled is nonlinear, we have extended the concept by including adaptive second-order Volterra filters. An efficient and fast converging DFT-domain algorithm for the adaptation of a second-order Volterra filter has been proposed. Simulation results have shown that the proposed approach leads to an increased echo attenuation compared to a linear AEC if the echo signal is corrupted by nonlinear distortion due to loudspeaker nonlinearities.

5 References

[1] W. Herbordt et al., "Full-duplex multichannel communication: Real-time implementations in a general framework," in *Proc. IEEE Int. Conf. on Multimedia and Expo (ICME)*, Baltimore, July 2003.

[2] J.Benesty, D.R. Morgan, and M.M. Sondhi, "A hybrid mono/stereo acoustic echo canceler," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 6, no. 5, pp. 468–475, Sep. 1998.

[3] W.Herbordt and W.Kellermann, "Adaptive beamforming for audio signal acquisition," in *Adaptive Signal Processing: Application to Real-World Problems*, J. Benesty and Y. Huang, Eds. Springer, Berlin, Jan. 2003.

[4] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, Inc., New York, 2002.

[5] A. Stenger et al., "Nonlinear acoustic echo cancellation with 2nd order adaptive Volterra filter," in *Proc. ICASSP*, Phoenix, March 1999.

[6] A. Fermo, A. Carini, and G. L. Sicuranza, "Simplified Volterra filters for acoustic echo cancellation in GSM receivers," in *European Signal Processing Conference (EUSIPCO)*, Tampere, Sept. 2000.

[7] H.Buchner, J. Benesty, and W.Kellermann, "Multichannel frequency-domain adaptive filtering with application to acoustic echo cancellation," in *Adaptive Signal Processing: Application to Real-World Problems*, J. Benesty and Y. Huang, Eds. Springer, Berlin, Jan. 2003.

[8] W. Herbordt and W. Kellermann, "Frequency-domain integration of acoustic echo cancellation and a generalized sidelobe canceller with improved robustness," *European Transactions on Telecommunications*, vol. 13, no. 2, pp. 123–132, Mar. 2002.

[9] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of a class of blind source separation algorithms for convolutive mixtures," in *Proc. ICA*, Nara, Japan, April 2003.

[10] D. Mansour and A. H. Gray, "Frequency domain nonlinear adaptive filtering," in *Proc. ICASSP*, Atlanta, March 1981.

[11] S. Im and E. J. Powers, "A block LMS algorithm for third-order frequency-domain Volterra filters," *IEEE Signal Processing Letters*, vol. 4, no. 3, pp. 75–78, April 1997.

[12] E. Moulines, O. A. Amrane, and Y. Grenier, "The generalized multidelay adaptive filter: Structure and convergence analysis," *IEEE Transactions on Signal Processing*, vol. 43, no. 1, pp. 14–28, Jan. 1995.

[13] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms and Applications*, Prentice Hall, New Jersey, 3rd edition, 1996.

[14] S. Im and E. J. Powers, "A fast method of discrete third-order Volterra filtering," *IEEE Transactions on Signal Processing*, vol. 44, no. 9, pp. 2195–2208, Sept. 1996.

[15] V. J. Mathews, "Adaptive polynomial filters," *IEEE Signal Processing Magazine*, pp. 10–26, July 1991.