TRANSFORM DOMAIN PREDICTION ERROR METHOD FOR IMPROVED ACOUSTIC ECHO AND FEEDBACK CANCELLATION

Jose M. Gil-Cacho, Toon van Waterschoot, Marc Moonen *

ESAT-SCD / IBBT-K.U.Leuven Future Health Department,
Katholieke Universiteit Leuven,
Leuven, Belgium.

Søren Holdt Jensen †

Aalborg University,
Dept. Electrical Systems,
Aalborg, Denmark.

ABSTRACT

The prediction error method (PEM) has been successfully applied in double-talk-robust acoustic echo cancellation (AEC) as well as in acoustic feedback cancellation (AFC). The main idea in both applications basically consists in decorrelating the adaptive filter input and error signals. This is done by whitening these signals with the inverse of a near-end signal model before the filter adaptation. The choice of the near-end model greatly affects the performance and complexity of the resulting AFC/AEC algorithms, oftentimes turning the algorithm impractical for real-world real-time applications. This paper proposes the use of discrete cosine transform (DCT), in conjunction with a simple near-end signal model, to boost the performance of PEM-based algorithms both in double-talk-robust AEC and AFC while only marginally increasing the computational complexity.

Index Terms— Prediction error method, acoustic echo cancellation, double-talk, acoustic feedback cancellation, transform domain.

1. INTRODUCTION

Acoustic feedback and acoustic echo are two well-known problems in speech communication applications, which are caused by the acoustic coupling between a loudspeaker and a microphone. On the one hand, acoustic feedback limits the maximum amplification that can be applied, e.g., in a hearing aid before howling, due to instability, appears [1],[2]. In many cases this maximum amplification is too small to compensate for the hearing loss, which makes acoustic feedback cancellation (AFC) algorithms an important component in hearing aids. On the other hand, acoustic echo cancellation (AEC) is widely used in mobile and hands-free telephony [3] where the existence of echoes degrades the intelligibility and listening comfort. The goal of AFC and AEC is essentially to identify a model for the feedback or echo path and to estimate the feedback or echo signal. The feedback or echo estimate is then subtracted from the microphone signal. These two applications in principle look the same and share many common characteristics, however they face different essential problems.

The main problem in AFC is the correlation, which is caused by the closed signal loop, that exists between the near-end signal and the loudspeaker signal. This correlation problem causes standard adaptive filtering algorithms to converge to a biased solution[1]. One of the solutions for this problem is therefore to reduce the correlation between the near-end signal and the loudspeaker signal. In AEC applications, on the other hand, the near-end signal is considered to be uncorrelated with the loudspeaker signal which is an approximation of reality. Except when the near-end signal is a white noise signal, the least-squares estimator is suboptimal which is typically the case in AEC. Moreover, practical AEC implementations rely on computationally simple stochastic gradient algorithms (e.g., NLMS). Therefore, it turns out that the presence of a near-end signal, in a so called double-talk (DT) scenario, will affect the adaptation in the AEC context by making the filter coefficients converge slowly and even diverge.

Reducing the bias in the feedback path model identification can be achieved by prefiltering the loudspeaker and microphone signals with the inverse near-end signal model before the adaptive filter [1],[2] using the prediction error method (PEM) [4]. The same concept has been successfully applied in [5] in order to achieve a DT-robust AEC by using knowledge of the near-end signal characteristics. In this way, the convergence properties of the echo path identification algorithm can be improved, even without the use of active DT detectors. For near-end speech signals, an auto-regressive (AR) model is commonly used [1] as it is indeed a very simple model. However, this single model fails to remove the speech periodicity, which causes the loudspeaker signal still to be correlated with the near-end signal during voiced speech. More complex models where different cascades of near-end signal models are used to remove the coloring and periodicity in voiced as well as unvoiced speech segments, e.g., the constraint pole zero linear prediction (CPZLP) [6] or the sinusoidal near-end model [2] have been proposed in the literature. However the overall AFC/AEC complexity increases dramatically.

In this paper the use of the discrete cosine transform (DCT) is proposed to boost the performance of the PEM adaptive filtering algorithms using row operations (PEM-AFROW) both in AFC and AEC while using a low–order AR near-end signal model. The idea of using a unitary orthogonal transform, like the DCT, of the adaptive filter signal is not new. Originally it was proposed to increase convergence rates in stochastic gradient algorithms such

*This research work was carried out at the ESAT Laboratory of KULeuven, in the frame of KULeuven Research Council CoE EF/05/006 'Optimization in Engineering' (OPTEC) and PFV/10/002 (OPTEC), Concerted Research Action GOA-MaNet, the Belgian Programme on Interuniversity Attraction Poles initiated by the Belgian Federal Science Policy Office IUAP P6/04 'Dynamical systems, control and optimization' (DYSCO) 2007-2011, Research Project FWO nr. G.0600.08 'Signal processing and network design for wireless acoustic sensor networks', EC-FP6 project 'Core Signal Processing Training Program' (SIGNAL) and was supported by a Postdoctoral Fellowship of the Research Foundation Flanders (FWO-Vlaanderen, T. van Waterschoot). The scientific responsibility is assumed by its authors

†Aalborg University, Dept. Electrical Systems,

as the least mean squares (LMS) algorithm [3], [7]. In this paper, however, the intention is to decorrelate the adaptive filter signal to achieve better double-talk robustness in AEC and achieve greater amplification rates, i.e., maximum stable gain (MSG), in AFC. The latter case was implicitly mentioned in [8]. The intention there was to have an efficient implementation of PEM-AFROW using the frequency domain adaptive filtering (FDAF). Some comments were given on complexity reduction but very little was said on performance increase. In [8], a discrete Fourier transform (DFT)-based FDAF was employed but, according to [7], the DFT is not the optimum transform for speech applications. The transformation that is closer to the optimal Karhunen-Loeve Transform (KLT) for *low-pass* signals, like speech signals, is the DCT [7]. Therefore the contribution of the paper is to use DCT-based transform domain (TD) PEM-AFROW (TD-PEM-AFROW) in speech applications to improve DT robustness in AEC and increase MSG in AFC.

The paper is organized as follows: Section 2 explains the signal model, algorithm and transformation, and it is shown how these are applied in AEC and AFC. In Section 3, simulation results are given and finally Section 4 concludes the paper.

2. TD-PEM-AFROW FOR AFC AND AEC

The acoustic feedback and echo cancellation concepts are shown in Fig. 1. The microphone signal is given as,

$$y(t) = x(t) + v(t) \tag{1}$$

with

$$x(t) = F(q,t)u(t) \tag{2}$$

$$v(t) = H(q,t)w(t) = \frac{1}{A(q,t)}w(t) \tag{3}$$

where q denotes the time shift operator and t is the discrete time variable, $v(t)$ is the near-end signal, $x(t)$ is the feedback or echo signal. $H(q,t)$ is the near-end signal model and $F(q,t)$ is the feedback or echo path between the loudspeaker and the microphone of order n_F. The feedback or echo canceler produces an estimate of the feedback or echo signal $x(t)$ which is then subtracted from the microphone signal $y(t)$. In the case of AFC the forward path $G(q,t)$ maps the microphone signal to the loudspeaker signal $u(t)$. In the case of AEC the echo-free error signal $e(t)$ is sent to the far-end and the loudspeaker signal $u(t)$ arrives from the far-end. In most applications the microphone signal is also corrupted by background noise $n(t)$ such that $y(t) = x(t) + v(t) + n(t)$.

The near-end signal can be modeled as an auto-regressive (AR) process with coefficients $A(q,t)$ of order n_A excited with a white noise signal $w(t)$ of time-dependent variance. These coefficients are calculated by means of linear prediction techniques and stored to form a filter (e.g., $L(q,t)$). Fig. 2 represents the concept of prefiltering the microphone and loudspeaker signal with the inverse model of the near-end speech signal. The signal model with (2) and (3) often fails to make the AFC/AEC completely remove the acoustic feedback or echo component in the microphone signal as will be shown in the simulations part. There are basically two reasons for this, one is the presence of noise and the second is that the model order n_A may be too low. This means that the adaptive filter does not only predict and cancel the feedback

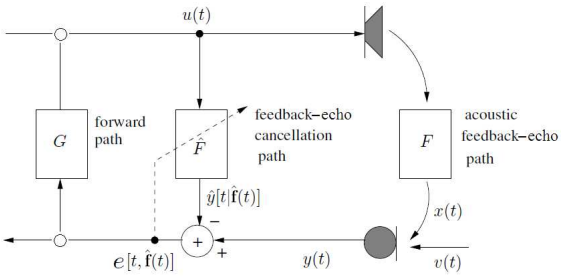


Fig. 1. AFC or AEC general set-up

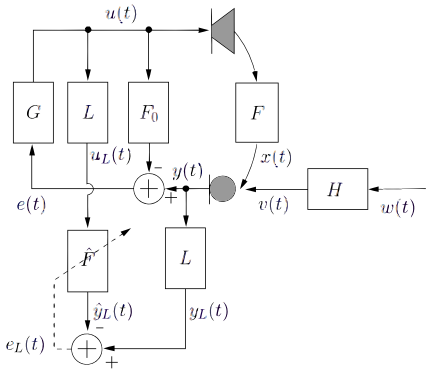


Fig. 2. AFC set-up with prefiltering of the loudspeaker and microphone signal

component in the microphone signal, but also part of the near-end signal, which results in a distorted feedback or echo compensated signal, smaller MSG in AFC and poor DT robustness in AEC. To further solve the problem of decorrelation using a minimal computational complexity increase, a DCT-based orthogonal transformation is proposed.

2.1. Transform Domain

The chosen orthonormal transformation is the discrete cosine transform (DCT) as it approaches the optimal KLT for speech signals [7]. The $n_F \times n_F$ DCT matrix coefficients $\mathbf{T}[kl]$ are given as

$$\mathbf{T}[k,l] = \begin{cases} \dfrac{1}{\sqrt{n_F}} & k = 1 \quad \text{and} \quad l = 1, ..., n_F \\ \left(\dfrac{2}{n_F}\right)^{1/2} \cos \dfrac{\pi(2l+1)k}{2n_F} & \\ k = 2, ..., n_F \quad \text{and} \quad l = 1, ..., n_F \end{cases} \tag{4}$$

The complete algorithm description using TD-PEM-AFROW for AEC is given in Algorithm 1. An equivalent AFC algorithm would be readily obtained by mapping the microphone signal back to the loudspeaker signal instead of transmitting it to the far–end as in the AEC case.

Algorithm 1: TD-PEM-AFROW for AEC

for $t = 1, 2, ...$ **do**
  j = mod(t,P);
  if j=0 **then**
    $\mathbf{f}(t-1) = \mathbf{T}^{-1}\hat{\mathbf{f}}(t-1)$;
    $\bar{\mathbf{y}}(t) = \bar{\mathbf{U}}(t)\mathbf{f}(t-1)$
    $\mathbf{d}(t) = \mathbf{y}(t) - \bar{\mathbf{y}}(t)$;
    $[\mathbf{a}, \delta^2] = \text{Levinson-Durbin}(\mathbf{d}, n_A)$;
    $\mathbf{u}_L(t) = \mathbf{U}(t)\mathbf{a}$;
  else
    $\mathbf{u}_L[2 : n_F](t) \leftarrow \mathbf{u}_L[1 : n_F - 1](t)$;
    $\mathbf{u}_L[1](t) = \mathbf{u}_{n_A}^T(t)\mathbf{a}$;
  end if
  $\mathbf{s}(t) = \mathbf{T}\mathbf{u}_L(t)$;
  $y_L(t) = \mathbf{y}_{n_A}^T(t)\mathbf{a}$;
  $\bar{y}_L(t) = \mathbf{s}^T(t)\hat{\mathbf{f}}(t-1)$;
  $e_L(t) = y_L(t) - \bar{y}_L(t)$;
  $\hat{\mathbf{f}}(t) = \hat{\mathbf{f}}(t-1) + \mu \dfrac{\mathbf{s}(t)}{\boldsymbol{\sigma}^2 + \delta^2 + \lambda} e_L(t)$;
  $e(t) = y(t) - \mathbf{s}^T(t)\hat{\mathbf{f}}(t)$;
end for

The vectors in Algorithm 1 are defined as

$$\mathbf{u}(t) = [u(t), ..., u(t - n_F + 1)]^T, \quad (5)$$

$$\mathbf{y}(t) = [y(t), ..., y(t - M + 1)]^T, \quad (6)$$

$$\mathbf{u}_{n_A}(t) = [u(t), ..., u(t - n_A + 1)]^T, \quad (7)$$

$$\mathbf{y}_{n_A}(t) = [y(t), ..., y(t - n_A + 1)]^T \quad (8)$$

The adaptive filter output (i.e., the feedback or echo estimate) may be expressed in vector notation as $\hat{y}_L(t) = \mathbf{u}_L^T(t)\hat{\mathbf{f}}(t)$, where the $n_F \times 1$ vector $\hat{\mathbf{f}}(t)$ contains the adaptive filter coefficients at time t and $\mathbf{u}_L(t) = [u_L(t), ..., u_L(t - n_F + 1)]^T$ is the input signal to the adaptive filter. The orthogonal matrix $\mathbf{T}$ transforms the adaptive filter input signal to the DCT domain as

$$\mathbf{s}(t) = \mathbf{T}\mathbf{u}_L(t) \quad (9)$$

The matrices are defined as

$$\mathbf{U}(t) = \begin{bmatrix} u(t) & \dots & u(t - n_A + 1) \\ \vdots & \ddots & \vdots \\ u(t - n_F + 1) & \dots & u(t - n_F - n_A + 2) \end{bmatrix}_{(n_F \times n_A)} \quad (10)$$

and

$$\bar{\mathbf{U}}(t) = \begin{bmatrix} u(t) & \dots & u(t - n_F + 1) \\ \vdots & \ddots & \vdots \\ u(t - M + 1) & \dots & u(t - M + 2 - n_F) \end{bmatrix}_{(M \times n_F)} \quad (11)$$

In the PEM-AFROW algorithm, the AR coefficients $\mathbf{a}$ and the variance δ^2 are calculated using the Levinson−Durbin recursion. P represents the frequency, in number of samples, at which this calculation is performed and M is the linear prediction window length. In Algorithm 1, $\boldsymbol{\sigma}^2$ is an $n_F \times n_F$ diagonal matrix whose elements are the power estimates of the elements in $\mathbf{s}(t)$ (i.e., $s[k](t)$ for $k = 1, ..., n_F$) such that

$$\sigma^2[k](t) = (1 - \alpha)\sigma^2[k](t - 1) + \alpha s^2[k](t), \quad (12)$$

α is a small factor chosen in the range $0 < \alpha \leq 0.1$, λ is also a small constant to avoid division by zero and δ^2 accounts for the energy variations in the near-end excitation signal.

The elements of the transformed input vector, $\mathbf{s}(t)$, appear to be approximately decorrelated with one another [3] [7]. Moreover, an appropriate power normalization (i.e., with $\boldsymbol{\sigma}^2$) can convert the input autocorrelation matrix to a normalized matrix whose eigenvalue spread will be much smaller than that of the original input signal, thereby improving the convergence behavior of stochastic gradient algorithms (e.g., LMS) in the transform domain. Although improving the convergence was the first idea of TD adaptive filtering, it turns out that the implicit decorrelation of the transformed input vector can be exploited in PEM-based AFC and AEC. The DCT is performed at each sample whereas FDAF typically works on a frame-by-frame basis [3] and so a better convergence is expected.

It is finally noted that there may be several other orthogonal transforms suitable for adaptive filtering algorithms. The DCT is one of the most popular orthogonal transforms and closest to the optimal KLT in speech applications.

3. SIMULATION RESULTS

Simulations were performed using speech signals. The sampling frequency in every simulation was 8 kHz. In the AEC simulations the far-end (FE) or loudspeaker signal was a female speech signal and the near-end (NE) signal a male speech signal; in the AFC simulations the near-end signal was the same female speech signal as in the AEC simulation. In the AEC simulations, the microphone signal consists of three concatenated segments of speech: the first 12 s segment consists of echo only, the second segment is the sum of echo + near-end signal generating a DT situation of 13 s, and the third segment is echo only again. The performance measures consist of *misadjustment* (MSD) for both AFC and AEC and the *maximum stable gain* (MSG) for AFC. The MSD between the estimated feedback path $\hat{\mathbf{f}}(t)$ and the true feedback path $\mathbf{f}$ represents the accuracy of the feedback path estimation and is defined as,

$$\text{MSD}(t) = 10 \log_{10} \frac{\left\| \hat{\mathbf{f}}(t) - \mathbf{f} \right\|_2^2}{\|\mathbf{f}\|_2^2} \quad (13)$$

The achievable amplification before instability occurs is measured by the MSG, which is derived from the Nyquist stability criterion [1] and defined as

$$\text{MSG}(t) = -20 \log_{10}[\max_{\omega \in \phi} |J(\omega, t)[F(\omega) - \hat{F}(\omega, t)]|] \quad (14)$$

where ϕ denotes the set of frequencies at which the loop phase is a multiple of 2π (i.e., the feedback signal $x(t)$ is in phase with the near-end signal $v(t)$), and $J(\omega, t)$ denotes the forward path processing before the amplifier, i.e., $G(\omega, t) = J(\omega, t)K$ with K the forward path gain.

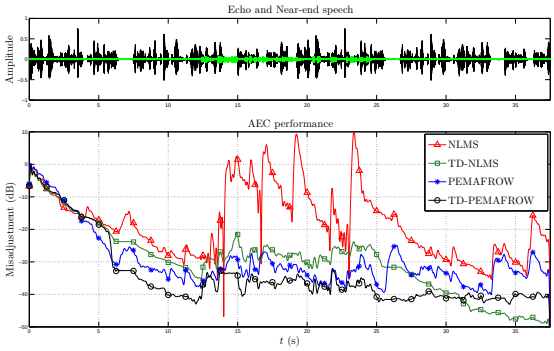
The near-end signal to echo ratio (SER) was set at two different levels: -25 and -15 dB which are typically found in hands-free mobile communications. The AR model order in AEC was chosen $n_A = 1$ following the indications given in [5]. A white (Gaussian) background noise at 35 dB SNR was added to the microphone signal. In AFC, the forward path gain K was set 3 dB below the MSG without feedback cancellation. Two different AR model orders were chosen as in [5]: $n_A = 12$ which is

common in speech coding for formant prediction and $n_A = 55$ being high enough to capture all near-end signal dynamics. The step sizes of the adaptation were tuned such that every algorithm had the same initial convergence rate. This choice aims to make a fair comparison of the resulting steady-state error of the solution. Two measured acoustic impulse responses were obtained from real devices, i.e., an 80-tap echo path from a mobile device for AEC and a 100-tap feedback path from a hearing aid for AFC simulations. In every case sufficient order was assumed. The linear prediction window length M was chosen to be 20 ms (160 samples), which corresponds to the frame in which speech is considered stationary. Four algorithms were compared in total, namely normalized least mean squares (NLMS), transform domain NLMS (TD-NLMS), PEM-AFROW and transform domain PEM-AFROW (TD-PEM-AFROW).

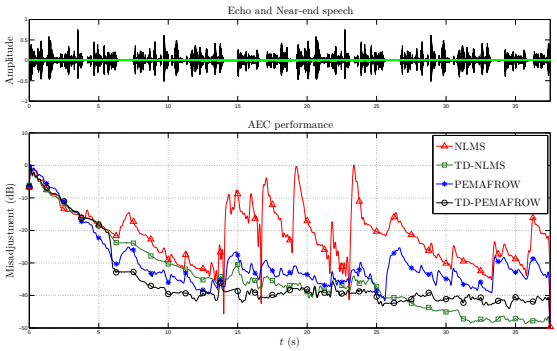
3.1. Discussion

AEC: Fig. 3.1 shows the AEC performance in terms of MSD at different SER. On the one hand it is observed that, as expected, NLMS performs very poorly during DT periods resulting in near-end speech distortion and no echo cancellation; therefore it will be excluded from the following discussion. On the other hand, it is observed that TD-PEM-AFROW outperforms the other algorithms during DT periods in both -15 dB and -25 dB SER. These two SER situations require a different analysis: In the case of SER -15 dB the TD-PEM-AFROW is consistently better than PEM-AFROW for around $5-6$ dB in average and around 10 dB better than TD-NLMS. The latter, however, offers reasonable robustness against DT. In the case of SER -25 dB the outstanding performance of TD-PEM-AFROW is demonstrated showing that the MSD remains around the same value as before the DT, and very importantly, with only small deviations compared to the other algorithms. This is of great importance since any deviation of the filter coefficients will lead to undesired echo (whose level is much higher) disturbing the error signal. This is exactly the weakness of PEM-AFROW in very low SER, since its MSD is around $8-9$ dB higher than for TD-PEM-AFROW and moreover its variance is also higher. In the -15 dB SER case, TD-PEM-AFROW still obtains an improvement in MSD of $3-4$ dB with respect to the -25 dB SER case, whereas PEM-AFROW barely gets 1 dB improvement. Surprisingly enough TD-NLMS obtains better MSD values than PEM-AFROW in this case.

AFC: Fig. 4 shows the AFC performance in terms of MSD and Fig. 5 in terms of MSG. Before continuing it is necessary to clarify that the solid line (i.e., instantaneous gain K) represents the limit at which the system is still stable; if the instantaneous gain K rises above the MSG, then the system becomes unstable and howling will appear. Between the solid line and the dotted one (i.e., the achievable MSG before feedback cancellation is applied) some "ringing" and therefore near-end distortion will appear (but not yet instability). If the MSG of an algorithm is above this threshold this means that some more amplification, represented by the MSG, could be applied in the forward gain of the system without instability. In both Fig. 5(a)-(b) it is shown that the NLMS is close to instability, meaning that some ringing distorting the near-end signal appears and no additional amplification would be possible without howling. Interestingly enough, TD-NLMS remains stable as shown in Fig. 5(a)-(b) and even performing better in terms of MSG than PEM-AFROW with an AR model order of 12 as shown in Fig. 5(a). Again TD-PEM-AFROW greatly outperforms the other algorithms: the MSD is consistently better for about 5 dB than that of PEM-AFROW and the MSG is shown to be much higher even with a low AR model order. It is worth noting that TD-PEM-AFROW also shows better performance both in terms of MSD and MSG than those shown in [2] using more complex near-end signal models.



(a) Misadjustment SER -15 dB

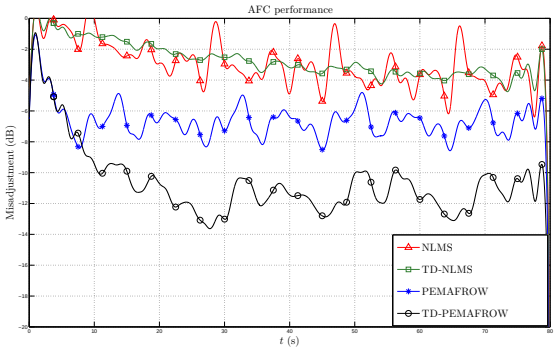


(b) Misadjustment SER -25 dB

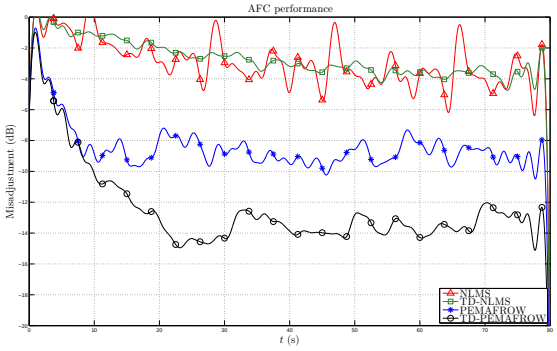
Fig. 3. AEC performance: Misadjustment with different SER

4. CONCLUSION

This paper has investigated the performance of a DCT-based TD-PEM-AFROW algorithm in terms of double-talk robustness in AEC and general improvement in AFC, with marginal complexity increase. Although the direct application of the DCT matrix requires $O(n_F^2)$ operations a fast DCT can be applied with $O(n_F \log n_F)$ operation only [9]. TD-PEM-AFROW is compared with standard NLMS, TD-NLMS and PEM-AFROW in different scenarios i.e., different SER in AEC and different AR model orders in AFC. It is shown that the combination of a prewhitening of the input and microphone signals together with transform-domain filter adaptation, successfully leads to an algorithm that solves the problem of decorrelation in a very efficient manner. The TD-PEM-AFROW algorithm is very robust in DT situations and boosts the performance of the simplest AFC (i.e., using only an AR model for the near-end signal). In the AFC context it actually outperforms state-of-the-art solutions that use more complex models for the near-end signal.

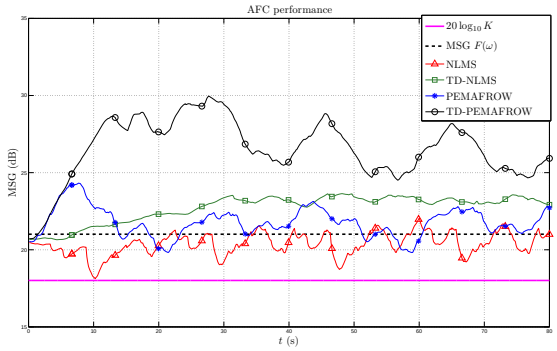


(a) Misadjustment AR model order 12

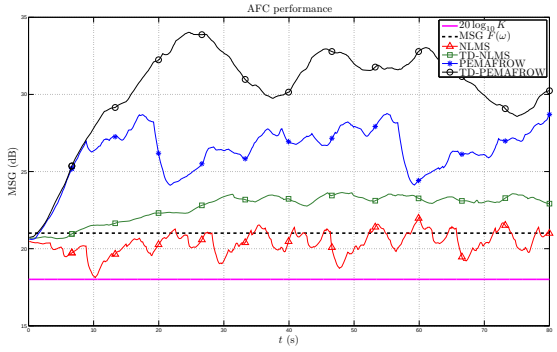


(b) Misadjustment AR model order 55

Fig. 4. AFC performance: Misadjustment, with forward gain 3 dB below the MSG before feedback cancellation is applied, at different AR model orders



(a) MSG AR model order 12



(b) MSG AR model order 55

Fig. 5. AFC performance: MSG, with forward gain 3 dB below the MSG before feedback cancellation is applied, at different AR model orders

5. REFERENCES

[1] 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. 2005.

[2] K. Ngo, T. van Waterschoot, M. G. Christensen, S. H. Jensen M. Moonen, and J. Wouters, "Prediction-error-method-based adaptive feedback cancellation in hearing aids using pitch estimation," in *In European Signal Processing Conference (EUSIPCO),*, Aalborg, Denmark, 2010.

[3] S. Haykin, *Adaptive Filter Theory (4th Edition)*, Prentice Hall Upper Saddle River, New Jersey, USA., 2002.

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

[5] T. van Waterschoot, G. Rombouts, P. Verhoeve, and M. Moonen, "Double-talk-robust prediction error identification algorithms for acoustic echo cancellation," *IEEE Trans. Signal Process.*, vol. 55, no. 3, pp. 846–858, March 2007.

[6] T. van Waterschoot and M. Moonen, "Adaptive feedback cancellation for audio applications," *Signal Processing*, vol. 89, no. 11, pp. 2185–2201, Nov 2009.

[7] B. Farhang-Boroujeny and S. Gazor, "Selection of orthonormal transforms for improving the performance of the transform domain normalised LMS algorithm," *Radar and Signal Processing, IEE Proceedings F*, vol. 139, no. 5, pp. 327 – 335, October 1992.

[8] G. Rombouts, T. van Waterschoot, and M. Moonen, "Robust and efficient implementation of the PEM-AFROW algorithm for acoustic feedback cancellation," *J. Audio Eng. Soc.*, vol. 55, no. 11, pp. 955–966, November 2007.

[9] K. R. Rao and P. Yip, *Discrete Cosine Transform: Algorithms, Advantages, Applications*, New York: Academic., 1990.