ROBUST AND LOW-COST CASCADED NON-LINEAR ACOUSTIC ECHO CANCELLATION

Moctar I. Mossi, Christelle Yemdji, Nicholas Evans

EURECOM
Sophia-Antipolis, France
{yemdji, mossi, evans}@eurecom.fr

Christophe Beaugeant, Philippe Degry

Infineon Technologies
Sophia-Antipolis, France
firstname.lastname@infineon.com

ABSTRACT

This paper addresses the problem of acoustic echo cancellation in non-linear environments. The first contribution relates to the use of a cascaded model which divides the loudspeaker enclosure microphone system into two main blocks; the first models the down-link transducers which are assumed to be the main source of non-linearity. The second block includes the acoustical channel and up-link transducers which are assumed to be linear and have a comparatively longer impulse response and higher time variability. The second contribution is a new non-linear adaptive echo canceler which is based on the cascaded model and has greater robustness to changes in the acoustic channel than an existing power filter approach.

Index Terms— Acoustic echo cancellation, non-linear echo, Volterra series

1. INTRODUCTION

The explosion of the mobile phone market and the need for low cost devices and miniaturization of components has led to the widespread use of lower quality, smaller loudspeakers that can introduce non-linearities in the acoustical coupling between the loudspeaker and microphone. This can result in non-linear echo artifacts in the up-link signal and thus echo cancellation algorithms are generally employed to improve speech quality.

Non-linearities generally degrade the performance of most echo cancellation algorithms that are based on the assumption of linearity and thus the problem of non-linear echo cancellation has emerged as an increasingly important problem [1].

There are two main approaches to tackle the problem of non-linearities in the acoustic path. The first approach is based on non-linear post filtering to suppress the residual non-linear echo [2]. In general the post-filter is preceded by a conventional linear adaptive filter. However, non-linearities have an adverse effect on linear filtering which impacts upon non-linear post processing and thus degrades global performance. The second, more popular approach is based on the use of a Volterra series and non-linear adaptive filtering [3]. Whilst there is less dependence on the performance of linear filtering the approach typically suffers from slow convergence.

In general the Volterra model takes a unified approach to estimate the overall Loudspeaker Enclosure Microphone (LEM) system. This involves the simultaneous tracking of non-linearities and changes in the acoustical channel, i.e. the path between the loudspeaker and microphone. This is potentially inefficient since the same acoustic path is estimated by each Volterra sub-filter. Since the sub-filter inputs are correlated converge is typically slow. This paper proposes a new method that can improve the convergence of the system using a cascaded LEM model. This approach uses a pre-processor which, aims to model the loudspeaker non-linearities, in series with a conventional linear adaptive filter to model the time varying acoustical channel. The linear adaptive filter is thus applied to a single input signal, which estimates the loudspeaker output, instead of being applied in parallel to the inputs of each sub-filter as in the Volterra model. Similar approaches to pre-processing based on clipping or polynomial models have already been proposed in [3, 4, 5]. In this paper we propose a new approach to pre-processing which is based upon the loudspeaker model proposed in [6]. The model in [6] uses parallel polynomial filters follow by a linear finite impulse response (FIR) filter to model the loudspeaker non-linearities and can be considered equivalent to power filters [7]. However, the proposed model is static, is thus dependent to the specific device and does not track slow variations which might occur over time. In this paper we propose an adaptive approach which explicitly models the loudspeaker characteristics and hence delivers superior echo cancellation performance.

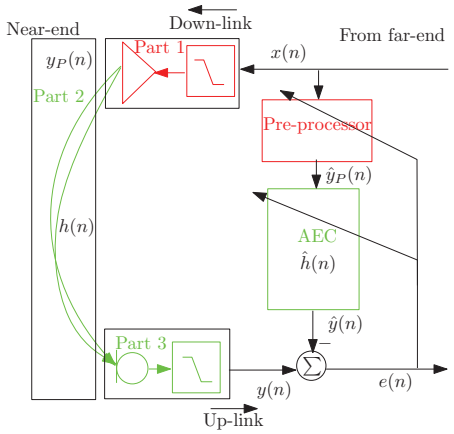

Fig. 1. The LEM is divided in two parts, the first one corresponds to non-linear model (red) and the second part is the linear model.

The remainder of this paper is organized as follow. In Section 2 we present the new cascaded model of the Loudspeaker Enclosure Microphone (LEM). In Section 3 we present the proposed non-linear acoustic echo canceler. Experimental work is presented in Section 4. Finally we present conclusions in Section 5.

2. LEM SYSTEM MODEL

In this section we present a general model of the LEM system. We also review the power filter presented in [7] and the new approach

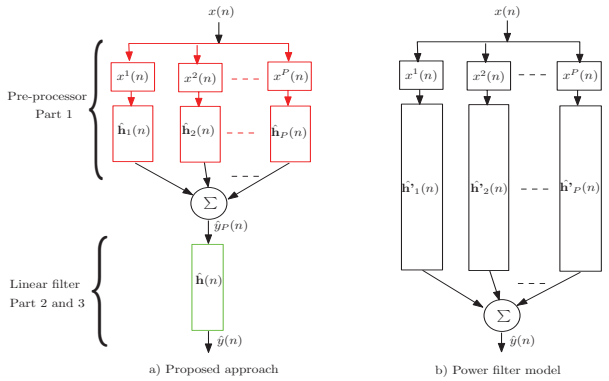

Fig. 2. Proposed model with P lower sub-filter ($h_p(n)$) in the pre-processor and the power filter model with P longer subfilter ($h'_p(n)$).

proposed in this paper.

The general LEM system illustrated in Figure 1 can be divided into three different parts. The first involves the down-link components and includes the amplifier and loudspeaker. This part of the system is well-known to have the greatest contribution to non-linearities [3, 6, 7]. Non-linearities stem from the the use of smaller loudspeakers and from higher signal levels in hands-free mode. With smaller components shorter impulse responses and lower variability is safely assumed [6]. The second part of the system is the acoustical channel which, in the absence of significant non-linearities, can be well-modelled by a linear filter [8]. The acoustical channel has a significantly longer impulse response and also a higher degree of time variability and thus filtering approaches are generally adaptive in nature [8]. The third part involves the up-link components and includes the microphone and amplifier. This part introduces less distortion and is generally assumed to be linear [3, 5, 7].

In view of their different characteristics and in contrast to the majority of current approaches, the idea here is to represent each part of the system with a separate, distinct model. The first part is distinctly non-linear whereas the second and third parts are predominantly linear. It is therefore desirable to use just two filters: one to represent the down-link path, which is assumed to have a short impulse response and be the principle source of non-linearities, and a second filter to represent both the acoustical channel and the up-link path. The linear part is dominated by the characteristics of the acoustical channel: a longer impulse response and higher variability. This strategy leads to a cascaded model of the acoustic echo path as illustrated in Figure 1 which includes a separate pre-processor and linear adaptive filter for acoustic echo cancellation (AEC).

With such an approach conventional linear adaptive filters are well suited to the second part. Being non-linear the down-link path is more troublesome but polynomial models [6] and power filters [7] are appropriate. A polynomial loudspeaker model as in [6] is used here, so that its combination with a linear filter (Figure 2 (a)) is comparable to the power filter model for non-linear AEC (Figure 2 (b)). Here the sub-filters of the power filter model are a combination of the pre-processor sub-filters $\hat{\mathbf{h}}_p(n)$ and the linear filter $\hat{\mathbf{h}}(n)$ and leads to the equality $\hat{h'}_p(n) = \hat{h}(n) * \hat{h}_p(n)$. For each sub-filter $\hat{h'}_p(n)$ we need at least the same number of taps as $\hat{h}(n)$ to model the LEM system with power filters. With more taps and high variability in the acoustical channel it becomes difficult to track the LEM system in this way which thus explains why the Volterra model is difficult to use in practice. An orthogonalization procedure was introduced in [7] to improve the performance when the length of $\hat{h'}_p$ is too large. The orthogonalization effect is explained in the following section which includes a detailed description of our approach.

3. CASCADED APPROACH TO NON-LINEAR AEC

In this section we present our new approach to non-linear acoustic echo cancellation with emphasis on the estimation of the loudspeaker model. We derive the Wiener solution to show the effect of input correlation, which led to the idea of orthogonalization of the inputs for power filters in [7], before deriving the adaptive solution.

Filter estimation is performed according to the Minimum Mean Square Error (MMSE) criterion. The Mean Square Error (MSE) is given by:

$$E\{e^2(n)\} = E\{(y(n) - \hat{y}(n))^2\}$$

where $y(n)$ is the echo signal and $\hat{y}(n)$ is the estimated signal given by:

$$\hat{y}(n) = \hat{\mathbf{h}}^T(n)\hat{\mathbf{y}}_P(n)$$

$\hat{\mathbf{h}}(n)$ is an N-column vector which represents the echo path and $\hat{\mathbf{y}}_P(n)$ is an N-column vector which contains the loudspeaker output given by:

$$\hat{y}_P(n) = \sum_{p=1}^{P} \hat{\mathbf{h}}_p^T(n)\mathbf{x}_p(n)$$

$\hat{\mathbf{h}}_p(n)$ is the estimated filter vector of length N_p and $\mathbf{x}_p(n) = [x^p(n), ..., x^p(n - N_p + 1)]^T$. The error can thus be written as:

$$e(n) = y(n) - \hat{\mathbf{h}}^T(n) \sum_{p=1}^{P} \hat{\mathbf{h}}_p^T(n)\mathbf{X}_p(n) \qquad (1)$$

where $\mathbf{X}_p(n) = [\mathbf{x}_p(n), ..., \mathbf{x}_p(n - N + 1)]$. As Equation 1 contains too many unknowns we need to assume that $\hat{y}_P(n) = y_P(n)$, i.e. that the estimate is equal to the true value. The MMSE solution of $\hat{\mathbf{h}}(n)$ is then given by:

$$\hat{\mathbf{h}} = \mathbf{R}_{y,y_P}\mathbf{R}_{y_P}^{-1}$$

where $\mathbf{R}_{y,y_P}$ is the cross-correlation between the microphone signal and the output of the loudspeaker and $\mathbf{R}_{y_P}$ is the auto-correlation of the loudspeaker output. This solution thus depends on knowledge of the loudspeaker output and will be discussed later in this section.

Here we derive the pre-processor sub-filters while assuming that only the filter $\hat{\mathbf{h}}_k$ is unknown whereas the others are known ($\hat{\mathbf{h}} = \mathbf{h}$ and $\hat{\mathbf{h}}_{p\neq k} = \mathbf{h}_{p\neq k}$). The MMSE solution is given by:

$$\frac{\delta E\{e(n)^2\}}{\delta \mathbf{h}_k} = \frac{\delta E\{(y(n) - \mathbf{h}^T(n) \sum_{p=1}^{P} \hat{\mathbf{h}}_p^T(n)\mathbf{X}_p(n))^2\}}{\delta \mathbf{h}_k}$$
$$= E\{\mathbf{X}_k(n)\hat{\mathbf{h}}^T(n)\big(y(n) - \mathbf{h}^T(n) \sum_{p=1}^{P} \hat{\mathbf{h}}_p^T(n)\mathbf{X}_p(n)\big)\}$$

If we suppose that $\mathbf{X}_k(n)\hat{\mathbf{h}}^T(n) = \tilde{\mathbf{y}}_k(n)$,$\tilde{\mathbf{y}}_k(n)$ has a length of N_k then:

$$\frac{\delta E\{e(n)^2\}}{\delta \mathbf{h}_k} = E\{\tilde{\mathbf{y}}_k(n)\big(y(n) - \hat{\mathbf{h}}^T(n) \sum_{p=1}^{P} \mathbf{h}_p^T(n)\mathbf{X}_p(n)\big)\}$$
$$= \mathbf{R}_{y,\tilde{y}_k} - \mathbf{R}_{Y_{p\neq k},\tilde{y}_k} - \mathbf{h}_k\mathbf{R}_{\tilde{y}_k}$$

where $\mathbf{R}_{y,\tilde{y}_k}$ is the cross-correlation between the echo signal and the corresponding output, $\mathbf{R}_{Y_{p\neq k},\tilde{y}_k}$ is the cross-correlation between

the other sub-filter outputs and the output of the sub-filter k. The estimate of the filter $\mathbf{h}_k$ in the MMSE sense is given by:

$$\hat{\mathbf{h}}_k = \left(\mathbf{R}_{y,y_k} - \mathbf{R}_{Y_{p\neq k},\tilde{y}_k}\right)\mathbf{R}_{y_k}^{-1} \tag{2}$$

Equation 2 shows that the estimation of the pre-processor sub-filters are dependent due to their inter-correlation. As the inputs of the pre-processor sub-filters are the powers of the same signal this may lead to a degradation of the estimation as a direct consequence of the inter-correlation. To overcome this limitation an orthogonalization procedure is introduced in [7] and shows that better performance is achieved when the sub-filter inputs are orthogonal. Orthogonalization leads to $\mathbf{R}_{y,\tilde{y}_k} = \mathbf{R}_{y_k,\tilde{y}_k}$ and $\mathbf{R}_{Y_{p\neq k},\tilde{y}_k} = \mathbf{0}$ so that the filter parameters become independent. In the proposed model we did not use orthogonalization since, with fewer taps in the pre-processor filters, it does not improve performance. As shown in [7] a further bias correction would be needed to improve performance and would lead to an overly complex solution in our case.

As we are in a short-term stationary environment adaptive filters are a necessity. The Least Mean Square (LMS) adaptive filter can easily be derived using an approach similar to that described in [4, 5, 9]. The LMS algorithm for the sub-filter $\mathbf{h}_k(n)$ is given by:

$$\begin{aligned}\hat{\mathbf{h}}_k(n+1) &= \hat{\mathbf{h}}_k(n) + \frac{1}{2}\mu_k \frac{\delta e(n)^2}{\delta h_k} \\ &= \hat{\mathbf{h}}_k(n) + \mu_k \frac{\delta e(n)}{\delta h_k} e(n) \\ &= \hat{\mathbf{h}}_k(n) + \mu_k \mathbf{X}_k(n)\hat{\mathbf{h}}^T(n)e(n)\end{aligned} \tag{3}$$

whereas the linear filter is given by:

$$\hat{\mathbf{h}}(n+1) = \hat{\mathbf{h}}(n) + \mu\hat{\mathbf{Y}}_P(n)e(n) \tag{4}$$

Equations 3 and 4 show that the linear filter and the pre-processor filter estimates are dependent. The problem of the dependency between filters is discussed in [5] where the authors suggest that linear filter adaptation is done before adaptation of the pre-processor. Here we start with the linear filter $\hat{\mathbf{h}}(n)$ and the sub-filters $\hat{\mathbf{h}}_1(n)$ and $\hat{\mathbf{h}}_2(n)$, since their inputs are the least correlated.

4. EXPERIMENTAL WORK

In the following we report a performance comparison of the cascaded filtering approach to both linear and power filtering approaches. In all cases tests were conducted with two different acoustic environments and the focus is on the robustness of the approaches to changes in the acoustic channel.

4.1. Test set-up

The non-linear environment is simulated with $P = 5$ sub-filters, each with 100 taps, for the first part of the system illustrated in Figure 2(a). A single, longer filter with either 200 or 300 taps is used to model the second part (acoustic channel and up-link). Additive noise is introduced to obtain two separate noise conditions of 30 dB and 40 dB SNR which is the ratio of the echo and noise signal energies. Artificial changes in the acoustical channel are introduced every 10 seconds simply by appending a sequence of zeros to the beginning of the impulse response, i.e. the impulse response is shifted or delayed by 2.5 ms in each case. This is done to assess the dynamic re-convergence performance of each algorithm. Here a normalized

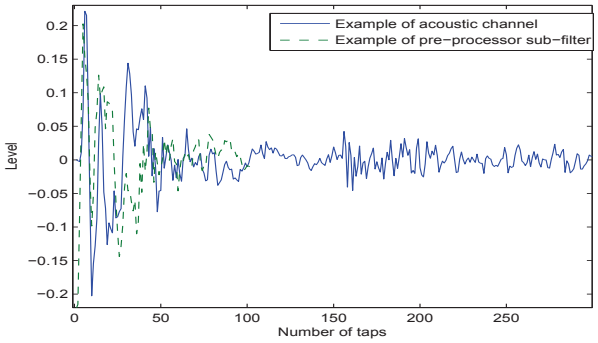

Fig. 3. Example impulse responses. The longer blue profile is that used for the acoustic echo path whereas the shorter green profile is that used for the pre-processor. The former is derived from real office environments and the latter is obtain by truncation of other acoustic channel impulse responses.

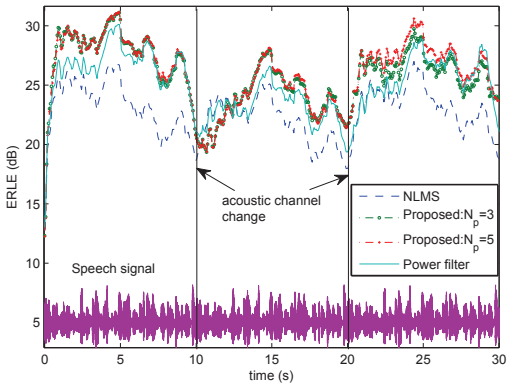

Fig. 4. ERLE against time for different algorithms and the far-end speech signal. Here 200 taps are used for the acoustical channel filter and the SNR is 40dB.

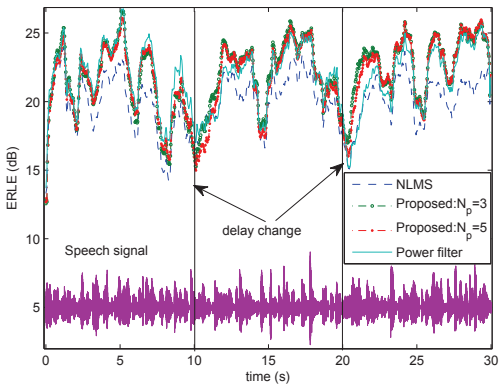

Fig. 5. As for Figure 4 except for an acoustical channel filter with 300 taps and an SNR of 30dB.

step size is used for the cascaded method which is given by:

$$\mu_k = \frac{\mu_{kn}}{\left\| \mathbf{X}_k(n)\hat{\mathbf{h}}^T \right\|^2 + \delta_k},$$

where δ is the regularization factor.

4.2. Common filter parameterizations

For the test whose results are illustrated in Figure 4 the number of taps for the acoustical channel filter is set at 200. The SNR is 40 dB. Illustrated are the ERLE profiles for four different algorithms. A set of common filter parameters were applied to all four algorithms and were chosen to maintain stability during changes in the acoustical path. The four filters and configurations are: the linear NLMS algorithm ($N = 200$, $\mu = .5$, $\delta = 1e-7$), a power filter without orthogonalization (see explanation in Section 3) ($N_{p=1,,5} = 200$, $\mu_1 = .5$, $\mu_{p=2,..,5} = .01$, $\delta_1 = 1e-7$, $\delta_{p=2,..,5} = 1e-4$) and the new cascaded algorithm ($N = 200$, $\mu = .5$, $\mu_{p=1,..,5} = .01$, $\delta_1 = 1e-7$, $\delta_{p=1,..,5} = 1e-4$) with two different lengths for the down-link model $N_{p=1,,5} = 5$ and $N_{p=1,,5} = 3$.

As shown in Figure 4 all algorithms start to converge at approximately the same rate but the NLMS algorithm reaches a lower maximum ERLE than other algorithms. This is due to the poor robustness of NLMS algorithms to noise and non-linearities [1]. Both the power filter and new cascaded methods achieve a higher level of ERLE. Upon initialization and each change in the acoustical channel we observe that the cascaded method converges more quickly than the power filter. This is due to the fact that changes in the acoustical channel have less impact on the pre-processor as it is designed to converge slowly. Robustness can be improved by decreasing the step-size or by increasing the pre-processor filter length but this has the effect of slower convergence at initialization.

4.3. Optimized filter parameterizations

Figure 5 illustrates performance where the acoustical channel filter has 300 taps and where the SNR is 30 dB. Changes in the acoustical channel are introduced every 10 seconds as before. For this experiment filter parameters were optimized independently for all four filters and the different acoustic conditions. The filters and configurations are: the linear NLMS algorithm ($N = 300$, $\mu = .7$, $\delta = 1e-7$), a power filter without orthogonalization ($N_{p=1,,5} = 300$, $\mu_1 = .7$, $\mu_{p=2,..,5} = .3$, $\delta_1 = 1e-7$, $\delta_{p=2,..,5} = 1e-3$) and the cascaded method ($N = 300$, $\mu = .7$, $\mu_{p=1,..,5} = .01$, $\delta_1 = 1e-7$, $\delta_{p=1,..,5} = 1e-4$) with two different lengths for the loudspeaker model $N_{p=1,,5} = 3$ and $N_{p=1,,5} = 5$.

In contrast to the first test all methods converge after initialization to a similar level of ERLE but both the power filter and the cascaded filter ultimately give better performance than the NLMS algorithm. With a higher step-size the power filter initially outperforms the cascaded filter, albeit only marginally. Upon changes in the acoustical channel, however, the cascaded filter is more robust than the power filter which has slower convergence but eventually reaches similar levels of performance. Larger step-sizes for the pre-processor in the cascaded filter do not improve initial performance as the algorithm becomes unstable. We also note that upon each change in the acoustical channel the power filter exhibits similarly poor convergence and levels of ERLE as the NLMS algorithm. Even if it ultimately achieves higher levels of ERLE it generally remains lower than the ERLE obtained by the cascaded filter.

In contrast to the results in Figure 4 we observe here that the cascaded filter has slightly lower performance for $N_p = 5$ than for $N_p = 3$. This is due to the fact that, with more taps, more time is needed for convergence.

Robustness to changes in the acoustical channel is a challenging problem. We acknowledge that the artificial changes in the acoustical channel that are used here to assess each algorithm are more abrupt than typically encountered in practice. Results nonetheless demonstrate the importance and perhaps account for why some researchers prefer acoustic echo suppression to cancellation [10].

5. CONCLUSIONS

This paper presents a cascaded approach to non-linear acoustic echo cancellation with loudspeaker non-linearities and acoustic echo variation. The MMSE solution is first derived to explain the effect of sub-filter input correlation on adaptation. We then propose an adaptive approach which exploits the reduced effect of correlation when the sub-filters have fewer taps and introduce a procedure to increase system stability and robustness to changes in the acoustical channel. This is achieved by configuring the pre-processor to converge slowly by using a small step-size. At the expense of poorer initial convergence the approach is shown to be more stable and robust than a parallel model based on power filters.

An open issue with the current system relates to system complexity, which is directly related to the number of sub-filters, which in turn has a bearing on the estimation of loudspeaker non-linearities. Work thus far has studied up to orders of $P = 5$ which might lead to prohibitive complexity for real-time implementation in low cost mobile devices. Further work should therefore address the influence of P on reliable estimation of non-linearities.

6. REFERENCES

[1] M. I. Mossi, N. W. D. Evans, and C. Beaugeant, "An assessment of linear adaptive filter performance with nonlinear distortions," *ICASSP*, March 2010.

[2] O. Hoshuyama and A. Sugiyama, "An acoustic echo suppressor based on a frequency-domain model of highly nonlinear residual echo," *ICASSP*, vol. 5, May 2006.

[3] A. Guerin, G. Faucon, and R. Le Bouquin-Jeannes, "Nonlinear acoustic echo cancellation based on volterra filters," *IEEE Trans. on Speech and Audio Proc.*, vol. 11, pp. 672 – 683, Nov 2003.

[4] B. S. Nollett and D. L. Jones, "Adaptation of a memoryless preprocessor for nonlinear acoustic echo cancelling," *NISP*, Sept 1997.

[5] A. Stenger and W. Kellermann, "Adaptation of a memoryless preprocessor for nonlinear acoustic echo cancelling," *Signal Processing*, vol. 80, pp. 1747–1760, Feb 2000.

[6] M. I. Mossi, C. Yemdji, N. W. D. Evans, C. Herglotz, C. Beaugeant, and P. Degry, "New models for characterizing non-linear distortions in mobile terminal loudspeakers," *IWAENC*, Sept 2010.

[7] F. Kuech, A. Mitnacht, and W. Kellermann, "Nonlinear acoustic echo cancellation using adaptive orthogonalized power filters," *ICASSP*, vol. III, pp. 105–108, March 2005.

[8] C. Breining, P. Dreiseitel, E. Hänsler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, and J. Tilp, "Acoustic echo control, an application of very-high-order adaptive filter," *IEEE SP magazine*, pp. 42–69, July 1999.

[9] S. Haykin, *Adaptive Filter Theory* 4^{th} *Ed*, Prentice Hall, March 2001.

[10] O. Birkenes, "A phase robust spectral magnitude estimator for acoustic echo suppression," *IWAENC*, Aug 2010.