A COMPARISON OF DIFFERENT LOUDSPEAKER MODELS TO EMPIRICALLY ESTIMATED NON-LINEARITIES

Leela K. Gudupudi[1], Christophe Beaugeant[2], Nicholas Evans[1], Moctar Mossi[2] and Ludovick Lepauloux[2]

[1]EURECOM, Sophia-Antipolis, France
lastname@eurecom.fr
[2]INTEL Mobile Communications, Sophia-Antipolis, France
firstname.lastname@intel.com

ABSTRACT

This paper investigates and questions the suitability of modelling non-linear loudspeaker distortion with scalar diagonal (SD) Volterra series. This approach, popular in studies of non-linear acoustic echo cancellation (NAEC), is compared to an alternative non-scalar diagonal (NSD) model. The new model is estimated empirically but based on the theoretical underpinnings of non-linear convolution. Using common, real-speech test signals, the loudspeaker outputs synthesised by each model are evaluated objectively through their comparison to real loudspeaker outputs measured in controlled conditions. Results show that non-linear distortion estimated with the NSD model better reflects that measured empirically. We also show that NAEC experiments conducted with SD loudspeaker models have the potential to over-exaggerate performance, whereas those conducted with an NSD model better reflect practical performance.

Index Terms— Nonlinear acoustic echo cancellation, Volterra series, non-linear convolution

1. INTRODUCTION

The mobile device market has continued to grow unabated in recent years. Coupled with rising demand, the drive towards miniaturisation and convergence has led to the use of ever-smaller transducers. Unfortunately, smaller transducers can introduce non-linearities which typically degrade the performance of speech enhancement processes which depend on linearity. Echo cancellation is one such process which can be severely affected [1, 2, 3]. Accordingly, non-linear acoustic echo cancellation (NAEC) is today an active research area.

Most research in NAEC assumes that the loudspeaker is the principal source of non-linearities [1, 3, 4, 5, 6, 7, 8] and that they dominate those introduced by the microphone [9]. As illustrated in Fig. 1, the traditional approach to NAEC models the loudspeaker enclosure microphone system (LEMS) as a Hammerstein model with cascaded non-linear and linear blocks related to the loudspeaker and acoustic path

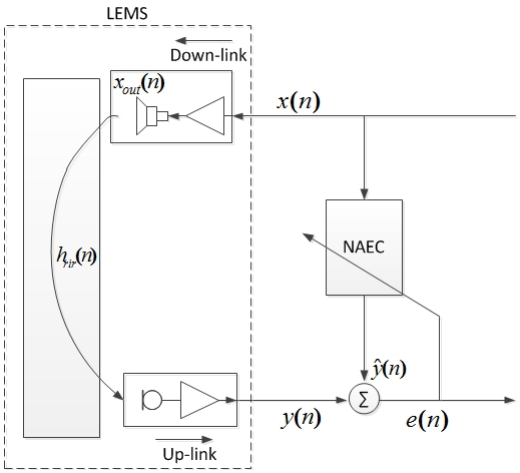

Fig. 1. System model illustrating a general approach to non-linear acoustic echo cancellation (NAEC).

respectively. Any research in NAEC is thus dependent on the accuracy of the non-linear loudspeaker model, be it used for NAEC itself, or to artificially synthesise non-linear test signals.

The most complex loudspeaker models typically involve a high number of parameters [10]. Lower complexity, scalar diagonal (SD) models based on Volterra series approximations are today the most popular. While they typically deliver efficient NAEC performance in well-controlled simulations, even slight model inaccuracies tend to degrade performance in real conditions. Alternative models have thus been investigated. Farina [11] proposed a new approach to measure loudspeaker non-linearities more accurately. This work was extended in [12] which introduced a new Volterra-based loudspeaker model based on non-linear convolution.

In search of improved NAEC performance, we have applied Farina's idea to the problem of NAEC using a non-scalar diagonal (NSD) loudspeaker model. This paper reports

the first evaluation and comparison of SD and NSD loudspeaker models through a two-fold assessment. First, under controlled experimental conditions with real speech test signals, we compare the output of each model to that recorded from a real loudspeaker. Second, we report a comparison of each model when used for NAEC assessment.

2. LOUDSPEAKER MODELS

Here we present two different, non-linear loudspeaker models, both based on a generalised Volterra series. As illustrated in Fig. 1, the far-end or downlink signal is referred to as $x(n)$. The loudspeaker signal $x_{out}(n)$ is modelled according to a generalised Volterra series with memory length M [5]:

$$\begin{aligned} x_{out}(n) = & \sum_{i_1=0}^{M-1} x(n-i_1).h_1(i_1) + \\ & \sum_{i_1=0}^{M-1}\sum_{i_2=0}^{M-1} x(n-i_1).x(n-i_2).h_2(i_1,i_2) + \\ & \sum_{i_1=0}^{M-1}\sum_{i_2=0}^{M-1}\sum_{i_3=0}^{M-1} x(n-i_1).x(n-i_2).x(n-i_3).h_3(i_1,i_2,i_3) + \cdots \end{aligned} \quad (1)$$

where h_p, $p = 1, 2, ..$ is the p-dimensional matrix which represents the p^{th} order non-linearity.

In many studies of NAEC, non-linearities are represented as a memoryless power series. Eq. (1) is then simplified to:

$$x_{out}(n) = \alpha_1 x(n) + \alpha_2 x^2(n) + \alpha_3 x^3(n) + \cdots \quad (2)$$

where α_1 replaces h_1 in Eq. (1) and is effectively the loudspeaker gain. For $p > 1$, h_p in Eq. (1) is replaced by α_p, the gain of the p^{th} order non-linear component. Since the diagonal terms in h_p are assumed to be scalar and since the non-diagonal terms are neglected, the model is referred to here as a scalar diagonal (SD) model.

This article investigates an alternative to the SD model based on the work in [12]. While the new model also assumes memoryless non-linearities and diagonal h_p matrices, the scalar constraints are relaxed. Eq. (2) is then rewritten as:

$$\begin{aligned} x_{out}(n) = & \sum_{i=0}^{M-1} x(n-i).h_1(i) + \sum_{i=0}^{M-1} x^2(n-i).h_2(i) + \\ & \sum_{i=0}^{M-1} x^3(n-i).h_3(i) + \cdots \end{aligned} \quad (3)$$

where, to simplify notation, we set $h_p(i, i, ..., i) = h_p(i)$. The model in Eq. (3) is thus referred to as a non-scalar diagonal (NSD) model. With fewer approximations than the SD model, and with non-scalar values of $h_p(i)$ in place of scalar values α_p, the NSD model has the potential to characterise non-linear loudspeaker behaviour more accurately than the SD model. Of course, estimation of $h_p(i)$ is comparatively more complex. It is described in the next section.

Whichever loudspeaker model is used, for complete echo modelling we assume that the microphone signal $y(n)$ is obtained from the convolution of $x_{out}(n)$ with a linear room impulse response (RIR), $h_{rir}(n)$:

$$y(n) = \sum_{i=0}^{L-1} x_{out}(n-i).h_{rir}(i) \quad (4)$$

where L is the length of $h_{rir}(n)$. We consider here the simple case where the microphone captures echo only (no local speech or noise).

3. NSD MODEL ESTIMATION

Loudspeaker characterisation is performed using the so-called exponential sine-sweep approach described in [11]. The experimental setup is illustrated in Fig. 2. A mobile terminal is placed before a head and torso mannequin at a distance of 30cm. The device operates in hands-free mode at maximum volume. Sine-sweep test signals $x(n)$ are played by the mobile terminal loudspeaker. They cover the frequency range between $f_1 = 40$Hz and $f_2 = 4$kHz, a duration of $T = 10$s and are sampled at $F_s = 8$kHz. They are generated according to [11, 12]:

$$x(n) = sin\left[2\pi f_1 . \frac{T}{ln\left(\frac{f_2}{f_1}\right)} . \left(e^{\frac{n}{T}.ln\left(\frac{f_2}{f_1}\right)} - 1\right)\right] \quad (5)$$

Test signals are then recorded with a high-quality microphone mounted within the mannequin ear before being deconvolved with the inverse sine-sweep signal $x_{inv}(n)$:

$$x_{inv}(n) = x(N-n-1) * exp(-\frac{n}{ln(\frac{f_2}{f_1})}.T) \quad (6)$$

where $N = F_s * T$ is the number of samples. Application of Eq. (6) obtains a series of vectors g_p; $p \in [1, N-1]$ usually referred to as measurement impulse responses (IRs). The linear IR is given by g_1 and the higher-order, non-linear or harmonic IRs are given by $g_p, p > 1$.

All experiments were conducted in a non-anechoic acoustic booth with low reverberation. Since measurements reflect both loudspeaker behaviour and room acoustics, the measured responses g_p were equalised in order to suppress the influence of the latter. The assumed linear IR of the acoustic booth was measured using a similar procedure to that described above (Fig. 2). Here though, the mobile device is replaced with a

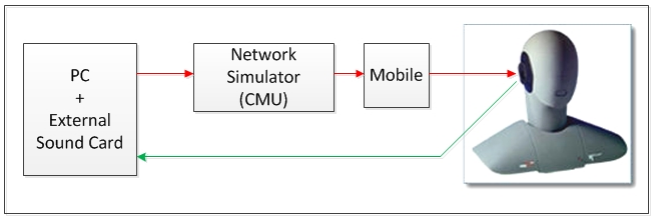

Fig. 2. An illustration of the experimental setup.

high quality loudspeaker with a flat frequency response in the region of interest (below 4kHz). An equalisation filter h_{eq}, which inverts the IR of the acoustic booth, was designed according to the approach described in [13]. The loudspeaker IR, denoted by $g_{eq,p}$ is then obtained by convolving the measured IRs g_p with h_{eq}:

$$g_{eq,p}(n) = h_{eq}(n) * g_p(n);\ p \geq 1 \tag{7}$$

Other experiments (not reported here) with different mobile devices showed that non-linearities above 5th order are negligible; they are dominated by lower order non-linearities. Accordingly, we considered only $g_{eq,p}$ for $p \leq 5$ in all experimental work reported here. Non-linear components h_p for $p \leq 5$ are computed through the linear combination of the equalised multiple IRs $g_{eq,p}$ using the method described in [12]. Examples are illustrated in Fig. 3.

4. SD MODEL ESTIMATION

For the SD model , we set $\alpha_1 = 1$. Weighting components α_p for $p > 1$ are chosen such that the mean linear-echo-to-total-non-linear-echo ratio ($LNLR_{tot}$) and the mean linear-echo-to-p^{th}-order-non-linear-echo ratio ($LNLR_p$) are the same as those of the NSD model. The recorded signal $y(n)$ is split into consecutive frames of 32ms duration. We consider also the linear and non-linear components $x_{out,k}(n)$ obtained from the NSD model as $x_{out,k}(n) = \sum_{i=0}^{M-1} x^k(n-i).h_k(i)$. The $LNLR_{tot}$ is computed according to:

$$LNLR_{tot} = \frac{1}{J}\sum_{j=1}^{J} LNLR_{seg}(j) \tag{8}$$

where J is the number of frames in the recorded signal x_{out} and where $LNLR_{seg}(j)$ is given by:

$$LNLR_{seg}(j) = 10log10\frac{\sum_{n=0}^{L-1} x^2_{out,1,j}(n)}{\sum_{n=0}^{L-1} x^2_{out,nl,j}(n)} \tag{9}$$

where $x_{out,1,j}(n)$ and $x_{out,nl}(n) = x_{out,2,j}(n) + x_{out,3,j}(n) + \cdots + x_{out,5,j}(n)$ are the linear and non-linear loudspeaker components respectively for frame j.
The $LNLR_p$ is computed as:

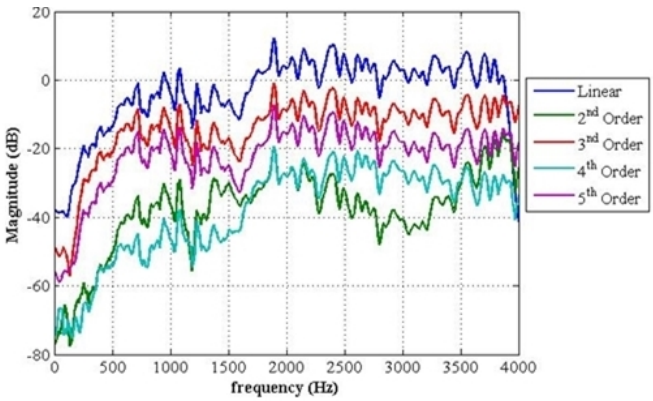

Fig. 3. Frequency responses of the linear component and up to 5th order non-linearities.

$$LNLR_p = \frac{1}{J}\sum_{j=1}^{J} LNLR_{seg,p}(j) \tag{10}$$

where the segmental $LNLR_{seg,p}(j)$ is given by:

$$LNLR_{seg,p}(j) = 10log10\frac{\sum_{n=0}^{L-1} x^2_{out,1,j}(n)}{\sum_{n=0}^{L-1} x^2_{out,p,j}(n)} \tag{11}$$

5. EVALUATION

This section presents a comparison of the two models. We compare loudspeaker output signals $x_{out}(n)$ synthesised from clean speech input signals $x(n)$ accordingly to Eqs. (2) and (3) for the SD and NSD model respectively to real empirically measured signals $x_{meas}(n)$. These are obtained from the same clean speech signals played by the mobile device loudspeaker and subsequently recorded at the mannequin ear using the experimental set-up described in Section 3. This signal is similarly equalised according to h_{eq} to remove room-effects.

5.1. Model Accuracy

Differences between the measured signal $x_{meas}(n)$ and those synthesised with each of the two models are assessed in terms of the Cepstral Distance (CD):

$$CD(m) = \sqrt{\sum_{N} \left[C_{x_{meas}}(m) - C_{x_{out}}(m)\right]^2} \tag{12}$$

where N is the number of samples in each frame and where $C_{x_{meas}}(m)$ and $C_{x_{out}}(m)$ are vectors of cepstral coefficients

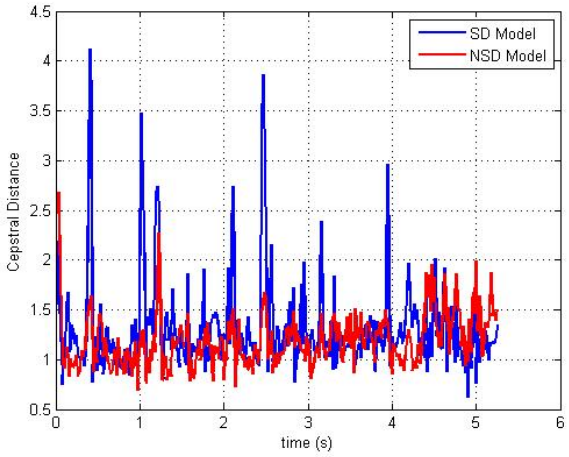

Fig. 4. An illustration of the cepstral distance between measured loudspeaker signals and those synthesised with SD and NSD models.

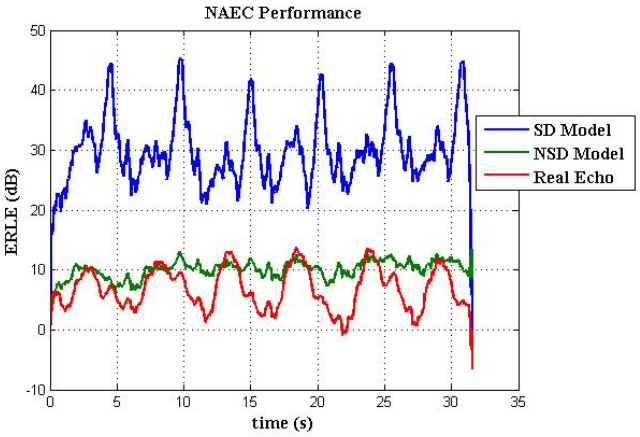

Fig. 5. NAEC performance in terms ERLE with either real, measured echo signals or those synthesised with with SD or NSD models.

as defined in [14]. The CD provides a more perceptually correlated assessment than alternative approaches based on energy or power differences. The CD profiles illustrated in Fig. 4 show that the difference between the measured signal and that synthesised with the NSD model is consistently lower than that between the measured signal and the signal synthesised with the SD model. The NSD model thus better reflects real recordings. This result was confirmed with extensive informal listening tests which showed that signals synthesised with the NSD model sound less artificial and are perceptually closer to the measured signal than those synthesised with the SD model.

5.2. Influence of the Models on NAEC Evaluation

Here we report a comparison of the two models in the context of NAEC. We assess the performance of a typical NAEC algorithm in treating either real, measured echo signals or echo signals synthesised with either SD or NSD models.

These experiments were performed using an NAEC preprocessor based on an adaptive polynomial loudspeaker model followed by an adaptive filter based on a conventional normalised least mean square algorithm. Full details are provided in [14]. For all experiments reported here, test signals $x(n)$ are speech signals of 30s duration, which is sufficient to ensure NAEC convergence. Loudspeaker non-linearities are synthesised through NSD and SD models as described in Sections 3 and 4. The resulting loudspeaker signals $x_{out}(n)$ are convolved with an acoustic echo path h_{rir}, reflective of a typical reverberant room, in order to synthesise microphone signals $y(n)$. Performance is assessed in terms of the echo return loss enhancement (ERLE) [15].

Results are illustrated in Fig. 5. While NAEC performance in the case of loudspeaker signals synthesised with the SD model is similar to that obtained in previous work [14], poorer performance is observed in the case of real, measured echo signals. While NAEC performance in the case of signals synthesised with the NSD approach also differs from that with real, measured echo signals, the difference is significantly reduced.

These observations confirm the significant, favourable bias in results generated with the popular SD model and emphasise its potential influence on the evaluation of NAEC performance. Results generated with the NSD model better reflect practical measurements and thus the new model is an appealing alternative to be considered for future work. Results thus derived will exhibit less bias than those reported previously in the open literature, and provide a more realistic estimation of practical NAEC performance.

6. CONCLUSIONS

This paper reports our work to assess the suitability of Volterra series derivatives in modelling the non-linear distortion introduced by small loudspeakers. We compared the outputs of two loudspeaker models to empirically measured, real loudspeaker outputs. The work suggests that the new NSD model described in this paper approximates more reliably practical non-linear loudspeaker behaviour. Assessments with non-linear acoustic echo cancellation show that the SD model can lead to favourably-biased indications of performance. In contrast, the NSD model more closely reflects empirically measured results and is thus an appealing, alternative model for future evaluations of non-linear acoustic echo cancellation performance.

7. REFERENCES

[1] A. Stenger, L. Trautmann, and R. Rabenstein, "Nonlinear acoustic echo cancellation with 2nd order adaptive volterra filters," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 2, Mar 1999, pp. 877–880 vol.2.

[2] A. N. Birkett and R. A. Goubran, "Limitations of handsfree acoustic echo cancellers due to nonlinear loudspeaker distortion and enclosure vibration effects," in *Applications of Signal Processing to Audio and Acoustics, 1995., IEEE ASSP Workshop on*. IEEE, 1995, pp. 103–106.

[3] R. Niemistö and T. Mäkelä, "On performance of linear adaptive filtering algorithms in acoustic echo control in presence of distorting loudspeakers," in *Proceedings of the Eight International Workshop on Acoustic Echo and Noise Control, IWAENC*, 2003, pp. 79–82.

[4] F. Kuech, A. Mitnacht, and W. Kellermann, "Nonlinear acoustic echo cancellation using adaptive orthogonalized power filters," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, vol. 3, 2005, pp. iii/105–iii/108 Vol. 3.

[5] A. Guerin, G. Faucon, and R. Le Bouquin-jeannes, "Nonlinear acoustic echo cancellation based on volterra filters," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 6, pp. 672–683, 2003.

[6] A. Stenger and W. Kellermann, "Adaptation of a memoryless preprocessor for nonlinear acoustic echo cancelling," *Signal Processing*, vol. 80, no. 9, pp. 1747–1760, 2000.

[7] L. Ngia and J. Sjobert, "Nonlinear acoustic echo cancellation using a hammerstein model," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 2, 1998, pp. 1229–1232 vol.2.

[8] M. Mossi, C. Yemdji, N. Evans, C. Beaugeant, and P. Degry, "Robust and low-cost cascaded non-linear acoustic echo cancellation," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 2011, pp. 89–92.

[9] M. Mossi, "Non-linear acoustic echo cancellation with loudspeaker modelling and pre-processing," Ph.D. dissertation, Thesis, 10 2012.

[10] W. Klippel, "Loudspeaker nonlinearities - causes, parameters, symptoms," in *Audio Engineering Society Convention 119*, Oct 8-10 2005.

[11] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Audio Engineering Society Convention 108*, Feb 2000.

[12] A. Farina, A. Bellini, and E. Armelloni, "Non-linear convolution: A new approach for the auralization of distorting systems," in *Audio Engineering Society Convention 110*, May 2001.

[13] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *The Journal of the Acoustical Society of America*, vol. 66, p. 165, 1979.

[14] M. Mossi, C. Yemdji, N. Evans, C. Hergoltz, C. Beaugeant, and P. Degry, "New models for characterizing mobile terminal loudspeaker distortions," in *International Workshop on Acoustic Echo and Noise Control, (IWAENC)*, 2010.

[15] E. Hänsler and G. Schmidt, *Acoustic echo and noise control: a practical approach*. Wiley. com, 2005, vol. 40.