New models for characterizing non-linear distortions in mobile terminal loudspeakers

Moctar I. Mossi, Christelle Yemdji
and Nicholas Evans
EURECOM
06560 Sophia-Antipols, France
{mossi, yemdji, evans}@eurecom.fr
www.eurecom.fr

Christian Herglotz, Christophe Beaugeant
and Philippe Degry
Infineon Technologies
06560 Sophia-Antipolis, France
{firstname.lastname}@infineon.com
www.infineon.com

***Abstract*—This paper presents two new approaches to model non-linear distortions which commonly arise in small loudspeakers used in mobile terminals. One is based upon a new, quantized frequency domain approach and another upon a parallelized polynomial filter approach. Both models are derived from practical studies of the input-output characteristics of real mobile terminal loudspeaker and artificial sinusoidal signals. The models are then used to predict the non-linear distortions in real speech signals. Comparisons to ground-truth distortions are performed to validate the models and confirm that both produce reliable predictions of non-linear distortions. In contrast to existing approaches, the new models are computationally efficient and are suitable for the real-time compensation of non-linear distortions.**

I. Introduction

Loudspeakers convert electrical signals into sound. With the miniaturization of mobile terminals the linearity of the loudspeaker is often adversely affected and, at sufficient levels, the associated non-linear distortion can become disturbing for the near-end listener. Linearity is also important for digital signal processing (DSP) algorithms which assume linear conditions. Therefore, without appropriate compensation, the performance of all downstream processes, will also be adversely affected, e.g. as in echo cancellation [1], [2].

One approach to mitigate such distortion involves loudspeaker linearization techniques which all rely on the non-linear modelling of the loudspeaker. Modelling typically involves an electro-acoustic and mechanical study of the loudspeaker to characterise its behaviour in non-linear conditions. These approaches, however, are generally too complex due to the high number of parameters which need to be estimated and the complex relationship between the electro-acoustic and mechanical properties [3]. The general conclusion of such studies show that loudspeakers are adequately characterised using a Volterra series for weak non-linearities and researchers have proposed many different loudspeaker models via such approaches [4], [5]. In general, however, all of these models rely on some restrictive assumptions such as slowly changes system characteristics, or limitations to second-order Volterra kernels for manageable complexity.

In this paper we present two new non-linear loudspeaker models which are both based on practical studies of input-output characteristics. The first model is based on frequency-domain, harmonic distortion modelling whereas the second approach is based on parallelized polynomial filters to model harmonic distortions. Both models are derived from the same set of empirical observations and are compared to real system outputs in order to demonstrate their effectiveness in predicting non-linear distortions in speech signals. The new models proposed in this paper aim to avoid the restrictive assumptions that are common to much of the existing work and are well-suited to the real-time compensation of non-linear distortions.

The remainder of this paper is organized as follows. Section II presents a system setup which is used to collect practical examples of non-linear loudspeaker distortions from real mobile terminals. This data is used to derive the two new non-linear loudspeaker models that are described in Section III. In Section IV we present an assessment of the two approaches by comparing loudspeaker outputs for real speech signals to those generated according to each of the two models. Finally we present our conclusions and perspectives in Section V.

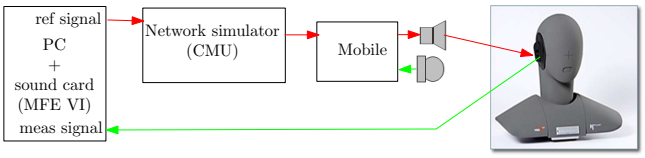

Fig. 1. An illustration of system setup. Reference signals are sent to a mobile terminal via a network simulator and are recorded with a high-quality microphone at the ear of a mannequin.

II. System characterization

Here we describe the experimental testbed that was used to acquire the empirical observations from which the two models are derived.

A. Global system setup

The system used for all of our experiments is illustrated in Figure 1. A PC is used to store and record all audio data that is sent to, or received from a mobile terminal via an MFE VI sound card [6] and a network simulator [7]. The loudspeaker output is recorded with an independent, high-quality microphone mounted in the ear of a mannequin [8]. The mobile terminal is placed in close proximity to the microphone, i.e. in handset mode rather than hands free mode, and all speech enhancement processes are deactivated. Since we aim to model loudspeaker distortions only we first verified the linearity of all other system, or channel elements. The sampling frequency of the input signals is $48kHz$. This is converted in the network simulator to $8kHz$ according to GSM specifications then recorded at $48kHz$ at the ear of the mannequin.

B. System linearity

In addition to the non-linear distortion introduced by the loudspeaker, various other non-linear signal processing algorithms, such as the speech codec (here the Enhanced Full-Rate codec) and CMU simulator, may also contribute distortions and thus corrupt the model of distortions introduced specifically by the loudspeaker. Therefore, it is necessary to determine amplitude and frequency ranges where the other system elements can be considered to behave linearly. Any distortions under these conditions can thus be reliably attributed to the loudspeaker only. To determine the linear range we conducted some non-intrusive tests where artificial, pure sinusoidal signals were sent to the mobile terminal but were recorded in digital form immediately before the loudspeaker. Signals with different amplitudes and frequencies were considered. By comparing the single sinusoidal input to the output we can easily observe any non-linear behaviour and thus determine amplitude and frequency ranges for which the other system elements can be assumed to be largely linear. Of course this is not a comprehensive test for linearity. Nevertheless our experimental results show that, at least for sinusoidal signals, the system is effectively linear for the full amplitude range between the frequencies of 200Hz and 3700Hz. The following analysis is based

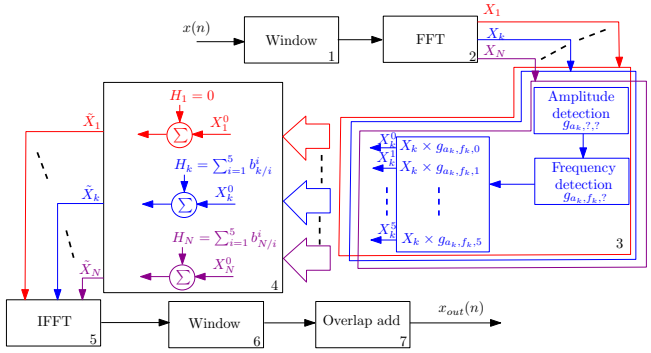

Fig. 2. The frequency domain model. The input signal is windowed and transformed into the frequency domain where the harmonic distortions are introduced according to the amplitude-dependent matrices described in Section II.

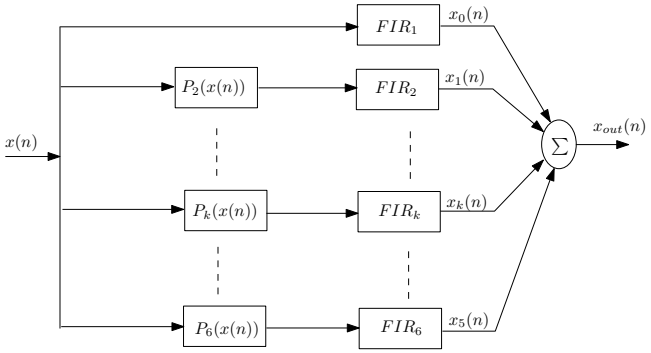

Fig. 3. I/O system of the polynomial model, the signal is processed in each stage to obtain a harmonic order distortion

on the response to sinusoids signals, we are sure that in this case that observed non-linearity may be reliably attributed to the loudspeaker and not to the systems elements.

C. Loudspeaker characterization

The non-linear behaviour of the loudspeaker is thus observed by repeating the same experiment described above but where signals are recorded after the loudspeaker. Here we consider single sinusoidal test signals with one of 10 different amplitudes in the range of 0dB (full-scale) to -27dB with a step size of -3dB and one of 80 different frequencies within the range of 50Hz to 4000Hz with a regular step size of 50Hz. Each of these signals may be denoted by $A_{i,ref}e^{2j\pi f_{i,ref}}$ where $A_{i,ref}$ is the amplitude and $f_{i,ref}$ is the frequency. This amounts to a total of 800 test signals. In order to observe the resulting harmonics the output signals are transformed into the frequency domain. Then, according to the same quantized frequency scale, the amplitudes at the output are set into a matrix, one for each input amplitude. Each matrix element thus gives the amplitudes at the output for each of the 80 fundamental reference frequencies and their generated harmonics. Here we suppose that intermodulation effects are negligible (see details in Section III-C). These matrices characterise the non-linear behaviour of the loudspeaker and are the basis of the new models that are described next.

III. Harmonics Distortion Modelling

Two models are described here: one is based upon a frequency domain approach and the other is based upon a polynomial approach.

A. Frequency domain model

The matrix model is based on the assumption that speech signals may be represented as a sum of sinusoids and thus that the non-linear effect of the loudspeaker may be modeled as the sum of the distortions on individual sinusoids. The decomposition into sinusoids is performed with the discrete Fourier transform (DFT) and the entire model is constructed in the frequency domain.

An overview of the system is illustrated in Figure 2. The input signal is first windowed into successive overlapping frames of length 1920 (40ms) with a sample rate of 48kHz, corresponding to a frame overlap of 75%. Each frame is transformed into the frequency domain where each component is denoted by $X_i = A_i e^{2j\pi f_i}$ and where i is the DFT bin, A_i is the amplitude and f_i is the frequency. Then, for each frequency f_i, we determine the nearest quantised sinusoidal reference frequency $f_{i,ref}$, in addition to the nearest reference amplitude $A_{i,ref}$, as described in Section II-C - i.e. we identify the '*closest*' or most applicable reference matrix. As explained in Section II-B, each reference sinusoid at the input leads, at the output, to (i) a fundamental sinusoid at frequency $f_{i,ref}$ and amplitude $A_{i,ref}(0)$ and (ii) 5 harmonics at frequencies $(k+1)\cdot f_{i,ref}$ with corresponding amplitudes $A_{i,ref}(k)$, for $k = 1...5$. For reasons of computational efficiency 5 harmonics are considered here. Experiments with higher numbers of harmonics showed only minor differences. $A_{i,ref}(0)$ and $A_{i,ref}(k)$ are obtained directly from the matrices described in Section II-C. We assume that if $A_i \approx A_{i,ref}$ and $f_i \approx f_{i,ref}$ then $\frac{A_i(k)}{A_i} \approx \frac{A_{i,ref}(k)}{A_{i,ref}}$, and hence we obtain the fundamental and harmonics generated by $A_i e^{2j\pi f_i}$ with the multiplication of A_i and the corresponding gain:

$$A_i(k) = g_{a_i,f_i,k} \times A_i, \tag{1}$$

where $g_{a_i,f_i,k} = \frac{A_{i,ref}(k)}{A_{i,ref}}$ is the gain applied to the k-th harmonic for an input signal of amplitude a_i and frequency f_i. This process corresponds to the 3^{rd} block in Figure 2. By combining all of the harmonics generated by each of the reference signals (block 4 in Figure 2) we obtain an approximation of the non-linear distortion in the frequency domain . Finally, a time domain signal is then resynthesized by applying an inverse DFT with overlap-and-add.

B. Polynomial model

Our so-called polynomial model is based upon a combination of polynomial and FIR filters. In contrast to the frequency domain model the idea here is to generate the different harmonics in the time domain according to different polynomial filters. The system is illustrated in Figure 3 where the polynomial filters are given by $P_k(x(n))$. Six parallelized branches aim to compute the linear response, $x_0(n)$, and the non-linear harmonics, $x_k(n)$. All signals are summed together with the original input signal to give the output $x_{out}(n)$.

The polynomial filter coefficients are determined according to the relationship between a cosine function at a multiple (harmonic) of the reference frequency and powers of the reference cosine:

$$cos(2\pi n \times f) = \sum_{i=0}^{n} \alpha_i cos^i(2\pi f). \tag{2}$$

Using trigonometric properties we determine the value of α_i for $n = 1, ..., 6$ (one fundamental frequency and five harmonics). These values correspond to the different coefficients in the polynomial model as given below:

$$\begin{aligned} P_1(x) &= x \\ P_2(x) &= 2x^2 - 1 \\ P_3(x) &= 4x^3 - 3x \\ P_4(x) &= 8x^4 - 8x^2 + 1 \\ P_5(x) &= 16x^5 - 20x^3 + 5x \\ P_6(x) &= 32x^6 - 48x^4 + 18x^2 - 1. \end{aligned} \tag{3}$$

Without added filtering the amplitude of the generated harmonics is independent of the input frequency and so an additional bank of FIR filters is used to adjust their amplitudes. If, for example, a particular range of input frequencies does not lead to any significant energy at the k-th harmonic, then a high-pass FIR filter, FIR_k, with high attenuation is applied to the output of the polynomial filter $P_k(x(n))$. For $k = 1$ the FIR filter is the impulse response which characterizes the coupling between the loudspeaker and the microphone in the ear of the mannequin.

To estimate the FIR filter coefficients we use reference signals to compute the gains, in a similar manner to that described in Section II. Filter gains are computed per harmonic using frame-by-frame FFTs of the input ($A_{i,ref}e^{2j\pi f_{i,ref}}$) and each individual output harmonic ($A_{i,ref}(k)e^{2j\pi k f_{i,ref}}$). Filter gains are then determined according to their average ratio:

$$G_k((k+1)f_i) = \frac{\overline{|A_{i,ref}(k)e^{j2\pi(k+1)f_i}|^2}}{\overline{|A_{i,ref}e^{j2\pi f_i}|^2}}, \quad (4)$$

The FIR filter is then the minimum phase filter which reflects the determined gain profile. After the estimation of all filter coefficients the system output is easy to compute. The input signal is passed through each combined polynomial and FIR filtering stage and the sum of the resulting signals gives the system output.

C. Constraints and limitations

Before we assess each of the two models we describe the limitations of each approach and their potential accuracy. The limits are defined by the complexity of the model, i.e. the size of the matrix described in Section II. For the frequency domain model this translates directly to the number of harmonics considered, which has a direct impact upon system accuracy. The bigger the matrix the better the accuracy, but the more complex the model. For the polynomial model accuracy depends on the number of stages (pair of polynomial filter and FIR) and the length of the FIR filters. Increases in the number of parameters will increase the complexity but less so than for the frequency domain model.

Finally, in the two approaches described above, intermodulation distortions are not considered. In the frequency domain model they are completely ignored. Some intermodulation distortions are generated with the polynomial model (though they were not considered directly in the design and polynomial parameter estimation). The only effect in this case is that they cannot be controlled independently from the harmonics. Whilst future work could consider the intermodulation effects, they were demeed to be of secondary importance in comparison to the more dominant harmonic distortions which are thus the sole focus in this paper.

IV. Experimental work

To compare the two models we assess each of them with real speech signals that are played at the loudspeaker of a mobile terminal and recorded at the ear of the mannequin as described in Section II. The signals measured at the ear are compared to the results obtained according to the two models as described in Section III. Three different metrics are used to assess models accuracy. First, signals are assessed in the time domain in terms of the segmental signal-to-estimate ratio (SER) given by:

$$SER(m) = 10 \times log_{10}\left(\frac{\sum_{i=m\times N}^{(m+1)\times N} x^2_{real}(i)}{\sum_{i=m\times N}^{(m+1)\times N} x^2_{model}(i)}\right) \quad (5)$$

where x_{real} is the speech signal recorded at the ear of the mannequin and x_{model} is the distorted speech predicted according to the model. Performance is also assessed in the frequency and cepstral domains

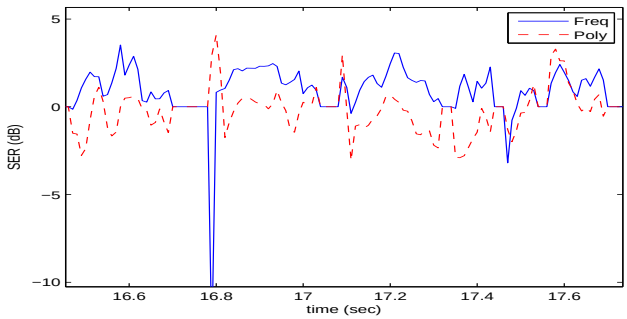

Fig. 4. Signal-to-estimate ratio (SER) against time for the two loudspeaker models. The frequency domain distortion model underestimates the real output whereas the polynomial model overestimates the real output.

through the log-spectral and cepstral distances as given in Equations 6 and 8 respectively. The cepstral distance is intended to give a more perceptually-related assessment, or at least one which is better correlated to subjective assessment than the spectral distance. The spectral distance (SD) is given by:

$$SD(m) = \sqrt{E\{(L_{x_{real}}(m) - L_{x_{model}}(m))^2\}} \quad (6)$$

where $L_{x_s}(m)$ is an N column vector of the m^{th} frame given by:

$$L_{x_s}(m) = 20 \cdot log_{10}(DFT[x_s(mN-1), ..., x_s((m+1)N)]) \quad (7)$$

The cepstral distance (CD) is given by:

$$CD(m) = \sqrt{\sum_N [C_{x_{real}}(m) - C_{x_{model}}(m)]^2} \quad (8)$$

where $C_{x_s}(m)$ is an N column vector of the m^{th} frame given by:

$$C_{x_s}(m) = IDFT\{ln|DFT[x_s(mN-1)...x_s((m+1)N)]|\} \quad (9)$$

In Equations 7 and 9 the index s is either $real$ or $model$. In all cases measurements come from consecutive frames of 20ms ($N = 960$) in length. For all experiments reported here performanceis evaluated using a dataset of 3 speech signals with a total length of 1 minutes.

A. Time domain assessment

The SER provides an impression of global system performance and, when plotted against time as in Figure 4, profiles illustrate variation in the error against time between modeled and ground-truth distortions. Figure 4 shows a profile for an example speech signal which typifies performance across the whole speech dataset. The solid blue profile illustrates performance for the frequency domain model and the dashed red profile illustrates performance for the polynomial model. On average the two systems give similarly accurate distortion estimates: despite some deviations the SER for both models is generally within a margin of $+/-2dB$. Figure 4 also shows that the polynomial model generally overestimates the distortion (SER< 0) whereas the frequency domain model generally underestimates the distortion (SER> 0). This can be explained by the complete absence of intermodulation harmonic estimation in the frequency domain model, leading to lower energies in x_{model} than in x_{real}. In contrast, the polynomial model leads to an overestimation of intermodulation, and consequently more energy in x_{model} than in x_{real}.

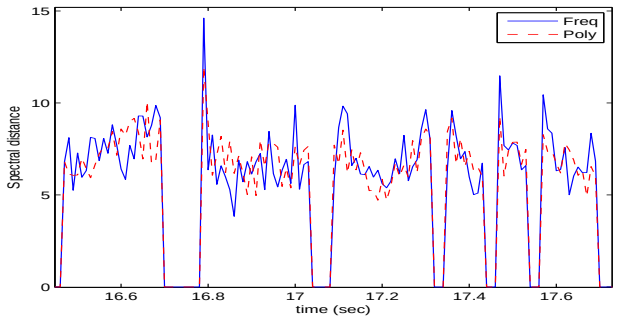

Fig. 5. Frequency domain assessment with the log-spectral distance

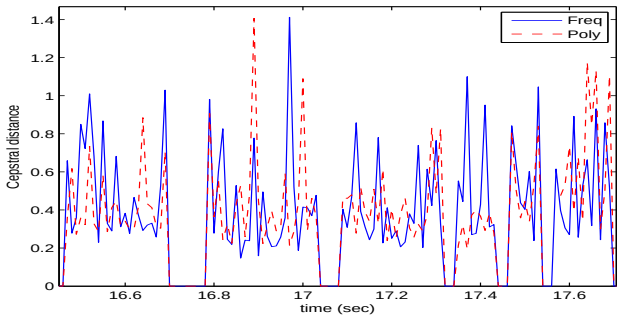

Fig. 6. Frequency domain assessment with the cepstral distance

Overall, the two models lead to approximately the same amount of error with a mean absolute SER of 1.33dB and 1.28dB for the frequency domain and polynomial models respectively, for the complete speech dataset. The profiles in Figure 4 contain some significant breakdowns, especially for the frequency domain model (around 16.8s for instance). Listening tests reveal that these peaks occur typically only during speech/non-speech transitions, i.e. at 16.8 and 17.5 seconds in Figure 4. This can be explained by the fact that the frequency domain model generates harmonics from the speech signal at either side of the transition. Considering a silence/speech transition this leads to a form of pre-echo as the harmonics are generated for the entire frame being processed. This is the classical pre-echo effect inherent in frequency domain processing. These transitions are generally less perturbing with the polynomial model, despite important differences that can still be noticed in the SER measurement during such periods, in addition to informal listening tests.

B. Spectral and cepstral domain assessment

In order to give an assessment that is more reflective of human perception we also computed log-spectral (SD) and cepstral distances (CD) to assess model accuracy. Figures 5 and 6 show profiles for SD and CD respectively, for each of the two models. As for time domain measurements with the SER, the two models show similar performance. SD for both frequency domain and the polynomial models they are relatively close (averages of 7.11dB cf. 7.10dB across the entire speech datasets). As illustrated in Figure 6, the CD between modeled and ground-truth distortions is reasonably similar. The global mean of the CD for this typical example is about 0.52 for the frequency domain model and 0.50 for the polynomial model. We found similar averages between 0.5 and 0.7 for the whole speech dataset.

There are noticeable peaks in the SD profiles. These peaks correspond to the peaks in the SER profiles, i.e. during transitions. The CD profiles, however, show more erratic behaviour. Even if the CD remains relatively low, such erratic behaviour can be explained by the fact that the CD better reflects human perception and is hence more sensitive to perceptual distortion than the other distances considered. The peaks appear during different periods for the two models. Even if the mean distances are similar, the CD reflects the fact that the deviations between modeled and real signals sound different for both models: the kind of deviation introduced by the polynomial model does not appear for the same kind of speech signal as for the frequency domain model. Listening tests confirm this assessment. On one hand, the polynomial model interferes with the timbre of the signal, sometimes overly exaggerating certain frequencies compared to real recorded signals. On the other hand the deviations introduced by the frequency domain model are more noticeable during transitions, even within the speech signal, for instance during transitions between voiced and unvoiced speech. In any case the CD is relatively small and the variation over time is not that high. This indicates that the two models give a good approximation of system behaviour. This last point is also confirmed by listening tests during which the differences are audible, but the model outputs are comparable to the real recorded signals.

V. Conclusion

In this paper we present two new models of non-linear harmonic distortion in mobile terminal loudspeakers. Both models may be used to give relatively accurate predictions of loudspeaker behaviour, through a fixed set of coefficients determined empirically, and can be seen as a good first approximation of small loudspeakers. Nevertheless the models do not match perfectly with reality and thus there remains some potential for improvement. The lack of reliable intermodulation modeling seems to be the main drawback of both approaches. To increase accuracy one can envisage the introduction of at least first order intermodulation effects.

The frequency domain model is of higher complexity than the polynomial model. Considering that loudspeaker distortions are generally at low frequencies and for high level signals, such properties could be introduced into the models to reduce complexity. The complexity of the polynomial model could also be reduced by combining the FIR filters with the polynomial filters. Power control could also be added so that different FIR filters could be applied to successive frames, i.e. the frequency response of the FIR filters could vary from frame-to-frame according to the power of the input signal.

References

[1] Mossi, Moctar Idrissa, N. W. D. Evans, and C. Beaugeant, "An assessment of linear adaptive filter performance with nonlinear distortions," in *ICASSP 2010, 35th International Conference on Acoustics, Speech, and Signal Processing, March 14-19, 2010, Dallas, Texas, USA*, 03 2010.

[2] L. A. Azpicueta-Ruiz, M. Zeller, J. Arenas-Garcia, and W. Kellermann, "Novel schemes for nonlinear acoustic echo cancellation based on filter combinations," in *ICASSP '09*, 2009, pp. 193–196.

[3] W. Klippel, "Loudspeaker nonlinearities - causes, parameters, symptoms," *J Audio Eng Soc*, vol. 54, pp. 907–939, Oct 2006.

[4] W. A. Franck, "An efficient approximation to the quadratic volterra filter and its application in real-time loudspeaker linearization," *Signal Processing*, vol. 45, pp. 97–113, July 1995.

[5] X. Gao and W. Snelgrove, "Adaptive linearization schemes for weakly nonlinear systems using adaptive linear and nonlinear fir filters," *Circuits and systems, proc of the* 33^{rd} *Midwest Symposium*, vol. 1, pp. 9–12, Aug 1990.

[6] H. acoustic GmbH, "HQS-mobile rev.04," *Head acoustic standard documentation*, pp. 13–20, June 2008.

[7] RHODES&SCHWARTZ, "R&S CMU200 universal radio communication tester, specifications," *Data Sheet version 09.00*, Oct 2008.

[8] ITU, "Objective measuring apparatus, head and torso simulator for telephonometry," *ITU-T P.58 : TERMINALS AND SUBJECTIVE AND OBJECTIVE ASSESSMENT METHODS*, Aug 1996.