NONLINEAR ACOUSTIC ECHO CANCELLATION USING A HAMMERSTEIN MODEL

Lester S. H. Ngia and Jonas Sjöberg

Department of Applied Electronics
Chalmers University of Technology
S-41296 Göteborg, Sweden
ngia@ae.chalmers.se

ABSTRACT

In hands-free telephone or video conference application, there exists an acoustic feedback coupling between the loudspeaker and microphone, which creates the acoustic echo. Linear acoustic echo cancellers (AECs) are commonly used to remove this echo. However, they are unable to effectively cancel nonlinear distortions. This paper employs a Hammerstein model to describe the acoustic channel of a nonlinear system concatenated with a linear faded echo path. A feed-forward neural network is used to model the static nonlinearity and a Finite Impulse Response (FIR) structure is used to model the linear dynamic system. The formed nonlinear model is applied to real data collected in an anechoic chamber and it performs slightly better than linear models. Although the improvement is small, the results show some interesting insights on the characteristic of a loudspeaker's nonlinearities and their effect on the performance of an AEC.

1. INTRODUCTION

In a telephone connection between one or more hands-free telephones, a feedback coupling path is set up between the loudspeaker and microphone at each end. This acoustic coupling is due to the reflection of the loudspeaker's sound from walls, floor and other objects back to the microphone and the direct path from the loudspeaker to the microphone. The effects of the echo depend on the time delay between the incident and the reflected waves, the strength of the reflected waves and the number of paths through which the waves are reflected. Echoes arriving a few tens of milliseconds or more after the direct sound are very distracting and they decrease the quality of the speech. Such long delays can be caused by propagation time over long distances, the coding of the transmitted signals or the end acoustic echo path itself.

The data in the experiments is contributed by Ericsson Mobile Communication AB, Sweden.

The cancellation of this acoustic echo is crucial in hands-free telephone or video conference application, depicted in Fig. 1. The signal $u(t)$ is the far-end signal, $y(t)$ is the echo of $u(t)$, $x(t)$ is the near-end signal and $d(t)$ is the signal from the microphone. The AEC estimates $y(t)$ and subtracts it from $d(t)$. The remaining signal is indicated by $e(t)$. If the AEC is perfect, then $e(t) = x(t)$.

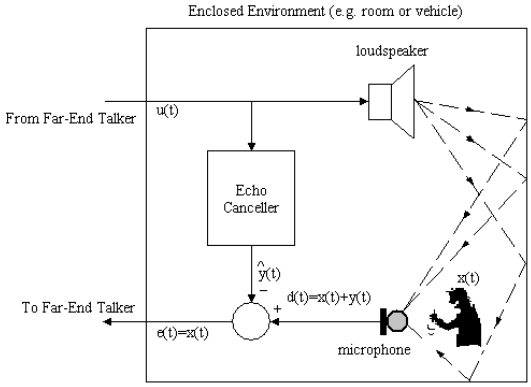

Figure 1: Acoustic echo canceller structure.

The echo path is not well approximated by the linear filters because it has a mixture of linear and nonlinear characteristics. The reflection of acoustic signals inside an enclosed environment is almost linearly distorted, but a loudspeaker introduces nonlinearities. The main causes of the nonlinearities are believed to be the suspension nonlinearity, which affects distortion at low frequency and the flux density inhomogeneity, which influences distortion at large output signal levels. These nonlinearities limits the performance of any linear cancellation algorithm [5].

In this paper a Hammerstein model [4] is used. It consists of a static nonlinearity concatenated with a linear dynamic model. The static nonlinearity is described by a neural net with one hidden layer and the linear dynamic part by a Finite Impulse Response filter (FIR). This kind of model structure is motivated by a physical reasoning, which will

be briefly discussed in the next section. The static nonlinearity is a function of $\Re^1 \rightarrow \Re^1$ followed by a tapped delay-line. This approach imposes a very specific structure on the model. However, it has the advantage of having reduced necessary parameters than if the tapped delay line is used first which gives a multi-input neural net as in [1][7]. To gain insights into the importance of the nonlinearity in different frequency and amplitude regions, separate models are estimated on the data from these different regions. The data used in the experiment is collected from a normal conversation using hands-free telephone in an anechoic chamber.

2. LOUDSPEAKER SYSTEM MODELING

The equivalent mechanical and electrical circuit of a loudspeaker are shown in Fig. 2. The details on this brief physical modeling can be found in [6][8]. The linear differential equation for the mechanical circuit is:

$$m\frac{d^2x}{dt^2} + r_M\frac{dx}{dt} + \frac{x}{C_M} = Bli \tag{1}$$

and the electrical circuit is:

$$e = ir + L\frac{di}{dt} + Bl\frac{dx}{dt} \tag{2}$$

where m is the total mass of the coil, x is the cone displacement, r_M is the total mechanical resistance due to dissipation in the air load and the suspension system, C_M is the compliance of the suspension, B is the air gap's flux magnetic flux density, l is the length of the voice coil conductor, i is the current in the voice coil, e is the internal voltage of the generator, r is the total resistance of the generator and L is the inductance of the voice coil.

The principal causes of nonlinearities in a loudspeaker are nonlinear suspension and non-uniform flux density. The force deviation property of a loudspeaker's cone suspension system is usually approximated by a third-order polynomial, and (1) can be rewritten as:

$$m\frac{d^2x}{dt^2} + r_M\frac{dx}{dt} + \alpha x + \beta x^2 + \gamma x^3 = Bli \tag{3}$$

where α, β, γ are the polynomial's coefficients.

At low frequency, the derivatives in (3) are small compared to the effect of the nonlinearities introduced by the polynomial. Thus, nonlinear distortion can be expected to be more serious at low frequencies.

The second source of nonlinear distortion is non-uniform flux density, which is usually less than 1% if the amplitude of cone movement is small. However, at high amplitude operation, a loudspeaker has severe distortion because B in (1) and (2) is not a constant. It is a function of the movement x, which is approximated by a second-order polynomial:

$$B(x) = B_0 + B_1 x + B_2 x^2 \tag{4}$$

where B_0, B_1, B_2 are the polynomial's coefficients.

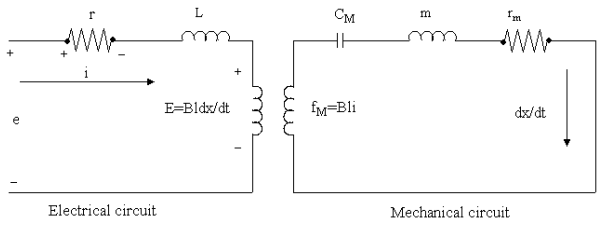

Figure 2: Equivalent mechanical and electrical circuits of a loudspeaker.

3. HAMMERSTEIN ECHO CANCELLER

The polynomial description in (3) and (4) can be seen as a volterra expansion and in [8] this model is applied to linearize a loudspeaker. However, high order polynomials often give poor modeling results and neural nets are tried instead [1][7]. The approach applied here is very close to the one in [1]. The main differences are that simulated data is used and an additional signal after the nonlinear loudspeaker is measured in [1]. The additional signal enables the possibility to estimate the parameters of a nonlinear part separately from those of a linear part. In a real application, however, this signal is unavailable. The results in this paper are based on real data and all the models' parameters (linear and nonlinear parts) are estimated simultaneously.

A Hammerstein model [4], as shown in Fig. 3, is used to model the loudspeaker and acoustics channel. The nonlinear static part consists of a feed-forward neural net with one hidden layer and the linear dynamic part is modeled by a FIR model:

$$y(t) = B(q)u(t) + e(t) \tag{5}$$

with

$$B(q) = q^{-n_k}(b_1 q^{-1} + \cdots + b_{n_b} q^{-n_b}) \tag{6}$$

where $[y(t)\ u(t)]_{t=1}^{N}$ are the input-output data and n_k is the delay. The noise term $e(t)$ is equals to $x(t)$ if the AEC is perfect.

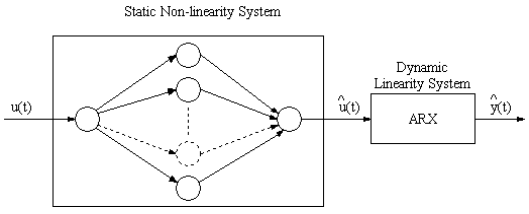

Figure 3: Hammerstein echo canceller model.

The estimated linear models are used to initialize the nonlinear models before the iterative estimation algorithm is

			Tailored Model					Generic Model		
Class	**Dynamic Range**	**Frequency Range**	**Linear**	**Hammer-stein**	**Gain (dB)**	**Orders (n_b, n_k)**	**Neuron**	**Linear**	**Hammer-stein**	**Gain (dB)**
1	+/- 3200	>300Hz	100.11	99.73	0.03	(139,12)	1	107.14	106.81	0.03
2	+/- 6000	>300Hz	272.42	264.32	0.26	(187,21)	1	292.17	284.12	0.24
3	+/- 6000	>300Hz	141.14	136.58	0.29	(227,15)	3			
4	+/-1600	<300Hz	94.01	93.09	0.09	(112,14)	4			
5	+/-1600	>300Hz	59.92	59.93	0	(147,11)	4			

Table 1: Mean square fit performance of the models on the different validation data sets.

applied. This initialization procedure of a nonlinear model guarantees that a nonlinear model gets a better fit on the estimation data than a linear model [2]. This is more superior than a random initialization of the parameters which is often used in neural nets. The criterion of fit is the sum of squared errors

$$\hat{\theta}_N = \arg\min_{\theta} \frac{1}{N} \sum_{t=1}^{N} \big(d(t) - \hat{y}(t) \big)^2 \qquad (7)$$

In the linear FIR model (7) is solved using the least-squares (LS) algorithm but in the Hammerstein model an iterative search has to be used. The presented results use Levenberg-Marquardt algorithm [3].

The model structure does not contain any noise model. The reason is that the near-end signal $x(t)$ should not be eliminated by the AEC. Instead, it is the entity that needs to be extracted.

4. REAL DATA EXPERIMENTS

The data are collected in an anechoic chamber with 5 second of input speech signal transmitted from the loudspeaker of a hands-free phone and the echo is picked up by the microphone. The signals are sampled at 8kHz with a 16bit ADC.

The data is segmented into five classes according to its dynamic and frequency ranges as shown in Table 1. Fig. 4 illustrates the division of data for the estimation data set. This method of data segmentation is also used for the validation data set so that five different validation sets are obtained. Class 1 and 2 contain data with a low and normal dynamic range respectively. Both classes are subset of Class 3 which has data with a mixed dynamic range. The data in these classes has spectral frequency more than 300Hz. Class 4 and 5 have data with frequency lesser and greater than 300Hz respectively, but with a same low dynamic range.

The performance of the nonlinear models are compared with the linear FIR models using validation data sets. The results are depicted in the column *Tailored Model* of Table 1. Each nonlinear models is slightly better than its linear counterpart, except Class 5. The *tailored models* of the mixed dynamic range class, i.e. Class 3, are regarded as *generic models* since they are tuned on a larger variety of data. They are used to validate the data on Class 2 and 3 and the results are shown in the column *Generic Model* of Table 1. The linear and Hammerstein *generic models* perform worse on both classes than even the linear *tailored models*. The improvements in dB, the optimum FIR's orders and number of hidden neurons for each classes are also shown in Table 1. Fig. 5 depicts the static nonlinearity characteristics across the input data range and the impulse response of the dynamic linear FIR system of the generic Hammerstein model. As expected, the nonlinearity curve is very close to a straight line and a deviation is evident only at high amplitude input signal.

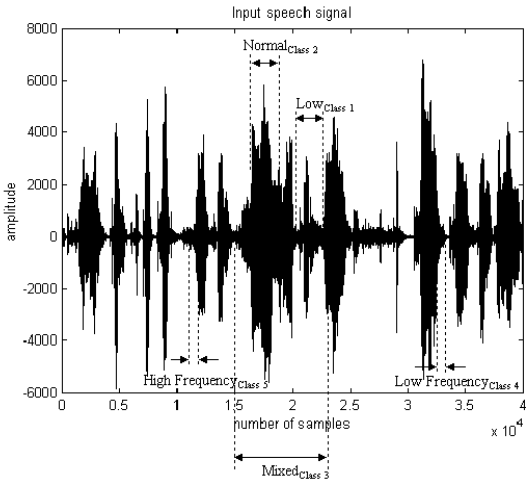

Figure 4: The division of speech signal into different segments depending on its amplitude and frequency ranges.

5. DISCUSSION OF RESULTS

The enhancements in the mean square fit value of the Hammerstein model as shown in Table 1 conform with the theory presented in Sec. 2. The nonlinearities are more dominant

when the input signal to the loudspeaker is large, i.e. in Class 2 and 3 data, and when the signal has low spectral frequency, i.e. in Class 4 data. However, the gain of about 0.3dB of Class 3 and 4 is more than the gain of 0.09dB of Class 4. Thus, it may appear that the loudspeaker may not has severe suspension nonlinearity.

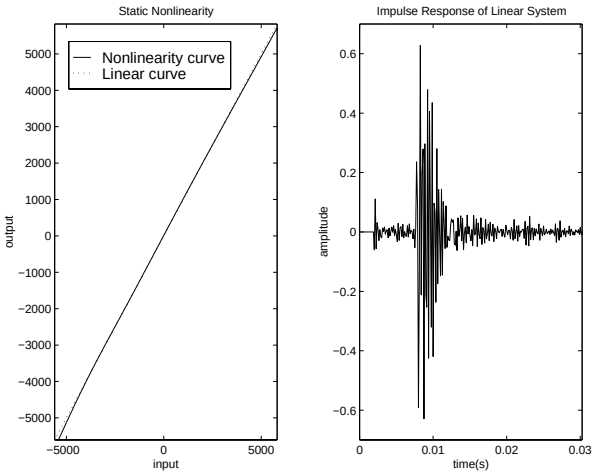

Figure 5: The static nonlinearity transfer function and the linear system's impulse response of the generic Hammerstein model.

The gain obtained for both Class 2 and 3 is not very large. One possible explanation is that the signal is not be large enough to cause distinguish flux density inhomogeneity. Experimentally, the nonlinearity distortions of a loudspeaker increases almost proportionally with the input signal level [9]. The improvement of a nonlinear echo canceller over a linear one is also more prominent when a loudspeaker is used at higher output power operating region [1]. To an extent, this plausible cause also explains the fact that one has almost the same gain for both the normal and mixed dynamic range classes. Nonetheless, the experiment is conducted under the dynamic range of a normal phone conversation. The small improvement of Class 3 model, i.e. the *generic model*, is also illustrated by the nearly linear curve of the static nonlinearity's input-output transfer function of the neural network as shown in Fig. 5.

Comparing the results between the *Tailored* and *Generic Model* of Class 1 and 2 data in Table 1, it is evident that the performance becomes better for the tailored models. This suggests that an AEC may perform better with models that are suited for different dynamic ranges than a single generic model. In addition, there is a gain of using nonlinear models within each dynamic range for both models. This shows that it is impossible to find a local linear model for each range.

6. CONCLUSION

Hammerstein models with a static nonlinearity of a neural net and a linear dynamic FIR have been estimated and compared to linear FIR models using acoustic echo data. The main points are:

- The nonlinear models give a small improvement over the linear FIR models.
- The nonlinear distortions are more pronounced when the input power to a loudspeaker is large and less obvious for low frequency signal which agree with the theory of a loudspeaker's nonlinearities.
- It is possible to acquire a more efficient AEC if different models are used in different dynamic ranges.

7. REFERENCES

[1] A. N. Birkett, R. A. Goubran. "Acoustic echo cancellation using NLMS-neural network structures". In *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, pages 3035–3038, 1995.

[2] J. Sjöberg. "On estimation of nonlinear black-box models: How to obtain a good initialization". In *IEEE Workshop in Neural Networks for Signal Processing*, pages 72–81, 1997.

[3] J.E. Dennis and R.B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Englewood Cliffs, NJ, 1983.

[4] L. Ljung. *System Identification: Theory for the User*. Prentice-Hall, Englewood Cliffs, NJ, 1987.

[5] M. E. Knappe, R. Goubran. "Steady-state performance limitations of full-band acoustic echo cancellers". In *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, pages 73–76, 1994.

[6] M. Rossi. *Acoustics and Electroacoustics*. Artech House, Inc., Norwood, MA, 1988.

[7] P. Chang, C. Lin, B. Yeh. "Inverse filtering of a loudspeaker and room acoustics using time-delay neural network". *Journal of the Acoustical Society of America*, 95(6):3400–3408, Jun 1994.

[8] X. Y. Gao, W. M. Snelgrove. "Adaptive linearization of a loudspeaker". In *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, pages 3589–3592, 1991.

[9] Y. Kaneda. "A study of non-linear effect on acoustic impulse response measurement". *Journal of the Acoustical Society of Japan*, 16(3):193–195, May 1995.