A MAXIMUM LIKELIHOOD APPROACH TO BLIND AUDIO DE-REVERBERATION

Massimiliano Tonelli

ANWIDA Soft
Strada Montefeltro 81, 61100 Pesaro, Italy
tonelli@anwida.com

Nikolaos Mitianoudis

Electronic Engineering Dept., Imperial College
Exhibition Road, London SW7 2AZ UK
n.mitianoudis@imperial.ac.uk

Mike Davies

DSP Group, QMUL
Mile End Road, London E1 4NS UK
mike.davies@elec.qmul.ac.uk

ABSTRACT

Blind audio de-reverberation is the problem of removing reverb from an audio signal without having explicit data regarding the system and/or the input signal. Blind audio de-reverberation is a more difficult signal-processing task than ordinary de-reverberation based on deconvolution. In this paper different blind de-reverberation algorithms derived from kurtosis maximization and a maximum likelihood approach are analyzed and implemented.

1. INTRODUCTION

In non ideal acoustic spaces, reverb may deteriorate the listening experience. This is common for rooms that have not been explicitly designed for such functionalities. The recommended approach in this case is to improve the acoustic of the space by modifying its physical properties. However, this might not be possible for constraints of several natures (e.g. historical buildings, churches, cockpit of a car).

In similar conditions, speech intelligibility, or more generally the quality of the listening experience, is badly affected and often the "duty" of enhancing it is left entirely to the public addressing (PA) system. Therefore, it is a common trend for today's music equipment industry to build "intelligent" PA systems able to address such problems.

In other cases, it might be of interest to enhance audio streams obtained in environments with bad acoustic characteristics (e.g. recording restoration, pre-processing for speech recognition software).

If we assume the acoustic paths as linear, the acoustic system can be modelled using a linear transfer function. Under this hypothesis, two families of identification methods exist:

- "**input-output" system identification methods** (supervised learning) If both the input and the output of the system are known, a supervised learning strategy is feasible. If there is a necessity to characterize the system by the measurement of a "steady" transfer function, swept sinus or maximum length sequences (MLS), identification techniques offer excellent results. Otherwise, to estimate a time varying transfer function, adaptive filtering strategies or input-output block based cross-correlation methods can be used.

These kinds of approaches can be applied when a signal, almost unaffected by the surrounding reverb, is available (e.g. electro-acoustic speech/music set, where every single source is picked up by a microphone placed in the very proximity of the source or by a line signal).

- **"blind" identification methods** (unsupervised learning) If only the output of the system is available (e.g. a speaker talking in a room with no microphone, or acoustic instruments playing in a theatre), an input-output identification strategy is not possible.

This paper will discuss this second approach.

2. IMPULSE RESPONSE AND ROOM TRANSFER FUNCTION

The impulse response (IR) of a generic system is defined as its output, when a Dirac δ function is applied to its input. If the system is linear time invariant (LTI) the IR specifies the system completely.

In fact, if:

1) $x(n)$ is the signal obtained sampling the speaker
2) $y(n)$ is the reverberated signal that is perceived by the listener
3) $h(n)$ is the IR of the source-listener acoustic path

then $y(n)$ can be expressed as

$$y(n) = h(n) \otimes x(n) = \sum_{k=-\infty}^{+\infty} x(k) \cdot h(n-k) \quad , \qquad (1)$$

that is: the convolution between x and h.

An IR of finite length can be represented as a finite impulse response filter (FIR) having as coefficients the samples of $h(n)$.

The transfer function that characterizes the source-listener path in a acoustic space can be approximated by a linear, time-invariant system. Such transfer function is called room transfer function (RTF).

2.1. Impulse response, echogram and early reflections

The number of coefficients necessary to model an RTF using an FIR model is proportional to the sampling frequency (Fs) and to the time necessary for the energy present in the room to vanish, once the excitation source has been interrupted.

As an example, using a sampling frequency of 22.05kHz, the FIR model of an RTF of a lecture room with a duration of 0.73 seconds, that requires about 16000 taps, is shown in Figure 1.

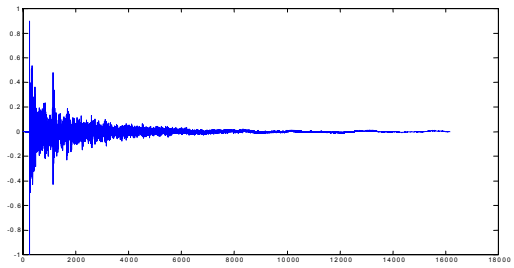

Figure 1: *IR of a lecture room*

An echogram, as shown below, is a useful tool to analyse the IR of an acoustic space. The echogram is calculated by the logarithm of the absolute value of the IR.

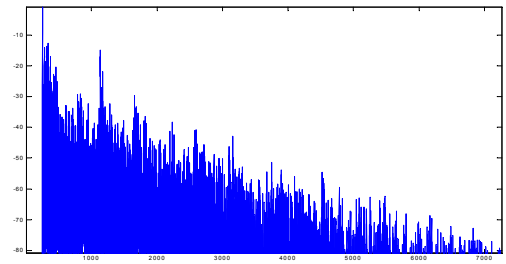

Figure 2: *Echogram of the previous IR.*

In the echogram it is easier to detect the most energetic, discrete reflections that characterize the IR. As reported in [1] the number of these reflections is small in comparison to the length of the whole IR. For example, in the reported IR there are only 61 coefficients that have intensity above –30 db (Figure 3). Usually these reflections are contained in the first 100ms of the IR (early reflections).

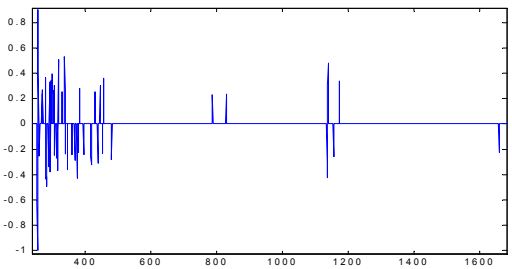

Figure 3: *Taps with an intensity greater than –30dB.*

3. INVERSE FILTERING

From "Iterative cepstrum-based approach for speech de-reverberation" by Radlovic and Kennedy, 1999 [2]:

"Let h(n) be a causal and stable, non-minimum phase impulse response between an acoustic source and a microphone placed some distance apart in a reverberant room.

If the requirement for signal processing is that the waveform of the input (source) signal remains unchanged after passing through the equalized room transmission system, we may think of the equalization problem as that of finding an inverse impulse response function p(k) such that

$$h(n) \otimes p(n) = \delta(n - N_d) \tag{2}$$

where n is a nonnegative time index, ⊗ signifies the discrete linear convolution, δ(k) is the unit sample sequence (δ(k)=1 for k=0 and δ(k)=0 for any other k) and N_d a delay."

3.1. Inverting and equalizing a single echo

Let us consider a single echo RTF of the kind:

$$h(n) = \delta(0) + \alpha \cdot \delta(k) \tag{3}$$

where k is the delay expressed in samples and α the reflection gain.
Its Z transform is

$$H(z) = 1 + \alpha \cdot z^{-k}; \tag{4}$$

this filter is also known as an FIR comb.
Its inverse transfer function is

$$G(z) = \frac{1}{1 + \alpha \cdot z^{-k}}; \tag{5}$$

this filter is also known as IIR comb.
The inverse Z transform of the IIR comb is

$$\alpha^{-n} \cdot \delta(n - m) \quad \text{with} \quad m = 0,\ k,\ 2k,\ 3k,\ \ldots \tag{6}$$

In the following example the IR of a single echo (k=50 and α=–0.95) and its truncated inverse filter are shown. Their convolution gives as a result a almost perfect approximation of a Dirac δ.

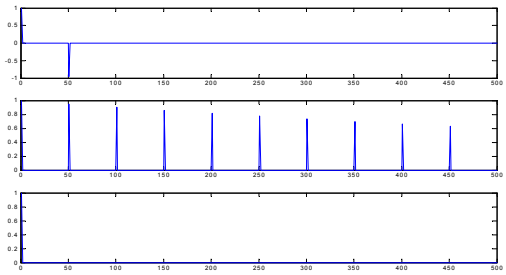

Figure 4: *Single echo, truncated inverse filter and their convolution.*

This simple example clarifies two facts:

1) By finding the inverse filter $P(n)$, perfect de-reverberation can be achieved.

2) The problem of equalizing a complex RTF can be seen, by linearity, as equalizing multiple single reflections. But even a single strong reflection may require a very long (theoretically infinitely long) and very resonant inverse filter. However, since reverberation is essentially due to energy that is decaying in a finite amount of time, a FIR structure is still a reasonable approximation.

4. BLIND RTF EQUALIZATION

Many techniques based on supervised identification approaches to measure an IR exist [3][4]. Once the IR has been measured, the inverse filter can be estimated and an equalized version of the signal can be obtained. Even if robust inverse filter design approaches have been proposed [5], several problems are still open in the application to practical audio systems. If the system input is unknown, the previous techniques cannot be applied and the equalization problem becomes even more complex.

A large class of blind identification methods are based on higher order statistics [6]. A non-Gaussian statistic is needed since second order statistic is "phase blind". This means that second order statistic cannot distinguish between different minimum and non-minimum phase system representations.

Furthermore, for signals having complex spectral structure (e.g. speech, music) the problem of blind RTF identification is complex since the observed power spectrum is a combination of the signal and reverb characteristics. Therefore de-reverberation cannot be obtained by direct spectral "whitening". A method of separating the two models must be provided.

Blind de-reverberation is still largely an unsolved problem.

4.1. Kurtosis and LP residual

LPC analysis has been proposed, as a way to decouple the spectral structure of the input and the reverb [7][8].

Kurtosis has been proposed as a suitable higher order statistic for blind system identification [8].

$$\text{kurtosis} = E\left\{\left(\frac{x - E\{x\}}{\sigma_x}\right)^4\right\} - 3 \qquad (7)$$

Kurtosis is a measure of the "peakedness" of the probability density function of a real-valued random variable [6].

A signal with sparse peaks and wide low level areas is characterized by a high positive kurtosis values.

In particular it has been observed [8] that, for clean speech signals, the kurtosis of LP residual can serve as a reverberation metric.

This observation can be explained roughly by the Central Limit Theorem (CLT): the sum of N arbitrarily (but identically) distributed RVs, S_N, converges to a Gaussian distribution.

By its nature, reverberation is the process of summing a large number of filtered and delayed copies of the same signal. Thus, very crudely, by the CLT, the reverberated signal has a more Gaussian distribution respect to the original one.

The LPC residual of speech is mainly constituted by the glottal pulses, so it is sparse and characterized by high, positive kurtosis. Reverberation causes this signal to assume a more Gaussian distribution with decreased kurtosis.

As a consequence, by building a filter that maximizes the kurtosis of the residual it may be possible to identify the inverse function of the RTF and therefore to equalize the system. This approach is blind since it requires only the evaluation of the kurtosis of the residual of the system output.

In [8] the following time domain single channel structure was proposed.

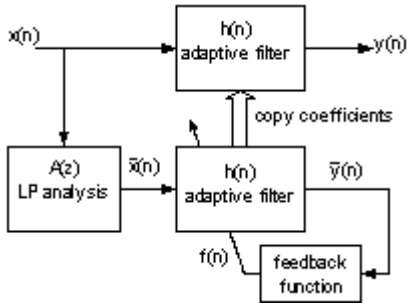

Figure 5: *A single channel time-domain LMS algorithm for maximizing kurtosis of the LP residual.*

During tests performed with this structure using batch processing, we have observed stability problems and unexpected results. In some apparently unpredictable conditions, the algorithm captures a harmonic component and creates a resonant spike in the residual. Therefore even if kurtosis is smoothly maximized, the inverse filter might converge to a highly resonant state.

By creating isolated strong peaks in the residual (thus making it sparser) kurtosis increases, but the result is irrelevant for de-reverberation purposes. This issue can be associated to the extreme sensitivity of kurtosis to "outlying" values.

Examining the simplified expression of kurtosis for a zero mean, unitary variance RV y

$$kurtosis = E\left\{y^4\right\} - 3 \qquad (9)$$

it can be noticed how its value is greatly affected by the values of y greater than 1. Similar criticisms of kurtosis have been raised in the context of Independent Component Analysis [9].

Let the RV y be generated by filtering a RV x:

$$y(n) = h^H(n) \cdot x(n) \qquad (10)$$

the derivative of kurtosis expression is

$$\frac{\partial kurtosis}{\partial h} = E\left\{4y^3 \frac{\partial y}{\partial h}\right\} = 4E\left\{y^3 \frac{\partial h \cdot x}{\partial h}\right\} = 4E\left\{y^3 \cdot x\right\}; \qquad (11)$$

therefore it is not bounded and can theoretically diverge.

4.2. Maximum likelihood approach

In order to minimize the sensitivity of the algorithm to "outlying" values we propose a maximum likelihood (ML) approach.

Assuming an FIR filter so that

$$y(n) = h^H(n) \cdot x(n)\,, \qquad (12)$$

the idea is to build h in order to achieve any desired probability density function of the output y:

$$\max_h E\{\log(P(y))\} = \max_h E\left\{\log(P(h^H x))\right\}. \qquad (13)$$

Defining the cost function as

$$J = E\{\log(P(y))\} \tag{14}$$

its gradient is

$$\frac{\partial J}{\partial h} = E\left\{\frac{\partial J}{\partial h}\log(P(y))\right\} = E\left\{\varphi(y)\frac{\partial h^H x}{\partial h}\right\} = E\{\varphi(y)\cdot x\} \tag{15}$$

where

$$\varphi(y) = \frac{1}{P(y)}\frac{\partial P(y)}{\partial y}, \tag{16}$$

therefore the update equation to maximize J is

$$h(n+1) = h(n) + \mu\cdot\nabla J(h(n)) = h(n) + \mu\cdot E\{\varphi(y)\cdot x\}. \tag{17}$$

The probability density function of y is chosen in order to have high kurtosis and bounded derivative.

A probability density function with these properties is

$$P(y) = \frac{1}{csh(y)} \tag{18}$$

giving

$$\varphi(y) = \frac{1}{P(y)}\frac{\partial P(y)}{\partial y} = csh(y)\frac{\partial}{\partial y}\frac{1}{csh(y)} = -tgh(y) \tag{19}$$

The update equation now becomes

$$h(n+1) = h(n) - \mu\cdot E\{tgh(y)\cdot x\} \tag{20}$$

Thus the hyperbolic tangent function replaces the kurtosis term. While the latter is unbounded and dominated by a cubic term, the former is bounded and insensitive to outliers.

5. RESULTS

In order to evaluate the proposed algorithm:

- an anechoic speech signal has been reverberated convolving it with a 16000 sample IR measured from a real room
- a 2000 tap FIR filter (corresponding to a time window of 90.7ms) has been used to equalize the real room IR.

To achieve a perfect equalization, a filter with a number of taps much greater than the length of the IR should be used. However, a shorter filter is expected to be less sensitive to noise during the adaptation and to be able to reduce the intensity of the early reflections in the considered time window (e.g. 2000 taps at a sampling frequency of 22050≈90.7ms). It is important to remember that the most energetic portion of an RTF is usually concentrated in the first 100ms.

From the echograms it can be noted that both the algorithms attenuate the most prominent early reflection (at sample position 1138). However, the kurtosis approach provides worst result and it introduces noise in the considered time window, making the processed IR noisier than the original one.

The batch ML algorithm delivers improved de-reverberation. Within the IR equalized using the ML algorithm, the strongest early reflection is attenuated by 13.5dB:

Original IR, strongest ER intensity = –14.7dB
Equalized IR, strongest ER intensity = –28.2dB
The average attenuation of all the IR intensity is of 4.93dB.
The average attenuation of the first 2000 taps is of 10.25dB.

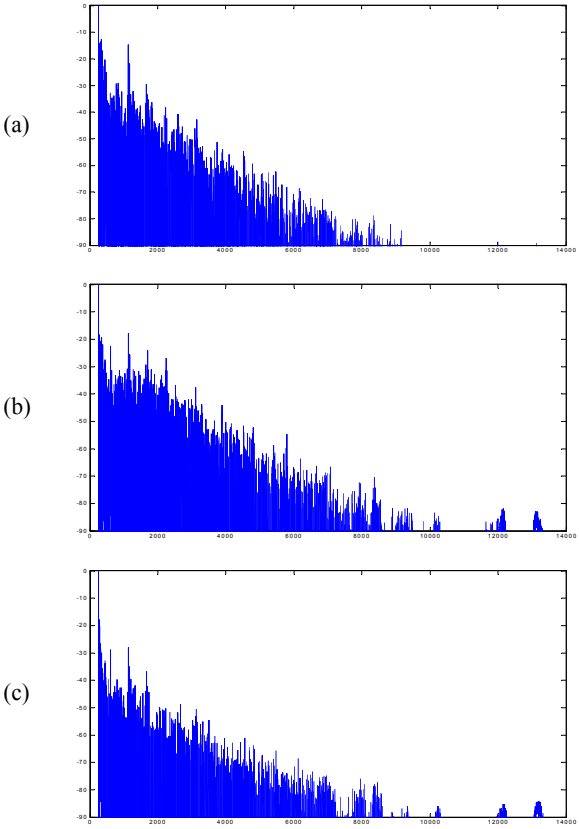

Figure 6: *(a) original IR, (b) IR equalized using the kurtosis based method (c) IR equalized using the ML based method*

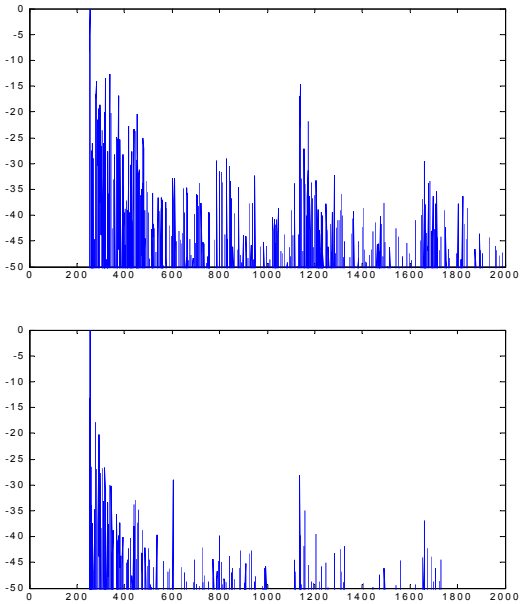

Figure 7: *Zoom of the echograms of the original IR and of equalized by ML algorithm.*

5.1. Multi-channel algorithm

Equalization with multiple microphones is potentially easier than using a single source [10]. As suggested by Gillespie et al [8], a multi-channel implementation can be directly extended from a single-channel system (see Figure 8). There a four channel system based on the improved-kurtosis, online algorithm was developed.
As stated before, the objective is to maximise the kurtosis of the LP residual of $y(n)$, where $y(n)$ is given by

$$y(n) = \sum_{c=1}^{C} h_c^T(n) \cdot x_c(n) \ , \qquad (21)$$

where C is the number of channels.

The multi-channel update equation thus becomes

$$h_c(n+1) = h_c(n) + \mu \cdot f(n) \cdot \tilde{x}_c(n) \qquad (22)$$

where the feedback function f(n) can take various forms (e.g. kurtosis or sigmoid). To jointly optimise the filters, each channel is independently adapted using the same feedback function.

While the extension to a multi-channel system is simple, the differences between a single channel approach and a multi-channel one are quite dramatic (indeed [8] provided no results for the single channel case). The reason is that it is often possible to represent the inverse filter using a set of FIR filters [10]. Recall that in the single channel case even a single echo can only theoretically be inverted using an IIR filter. Also, even when the full inverse is not attained the multi-channel setup is more similar to a beam-former and, since it mixes together filtered versions of the inputs, it can exploit constructive and destructive phase interference to de-reverberate. This greatly simplifies the filter complexity (see results).
This implies:

1) greater insensitivity to noise and better gradient estimation
2) less computational demand
3) less memory requirements

However the down side is it cannot offer "one channel to one channel" signal enhancement.

5.2. Synthetic example - equalization using a 4 channel system

Four long echoes (with delays of 2000, 2100, 2200 and 2300 samples and gains respectively of –0.5, 0.5, –0.5, 0.5) have been used to add reverb to the speech file. Figure 9 shows the impulse response of one of the four adaptive filters using the kurtosis based criterion.

Here, it can be seen how the filter models the time delay between the taps (100, 200, 300 samples), rather than the actual delays of 2100, 2200 and 2300.
The spectrograms of the original, reverberated and equalized files are shown in Figure 10. The algorithm manages to suppress almost completely the echoes present in the reverberated file.

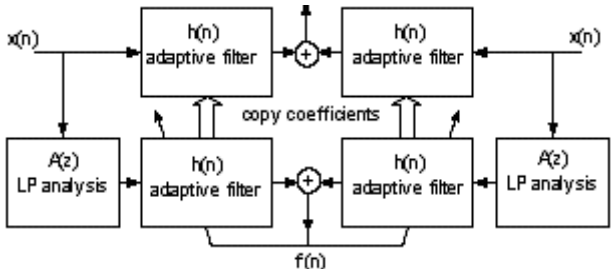

Figure 8: *Two channel system.*

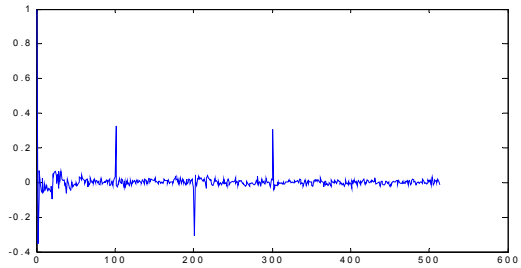

Figure 9: *Impulse response of 1 of the 4 filters of the multi-channel system.*

6. CONCLUSION

It has been demonstrated how the ML approach is effective in reducing the strength of the early reflections of an acoustic IR. However, the performances are poor in removing the late reverberant tail.

In this sense, two approaches should be investigated:

1) verify if a longer filter used with a larger quantity of input data can effectively remove the long reverberant tail (this approach is considered to be unlikely effective due to RTF's coherence instability).

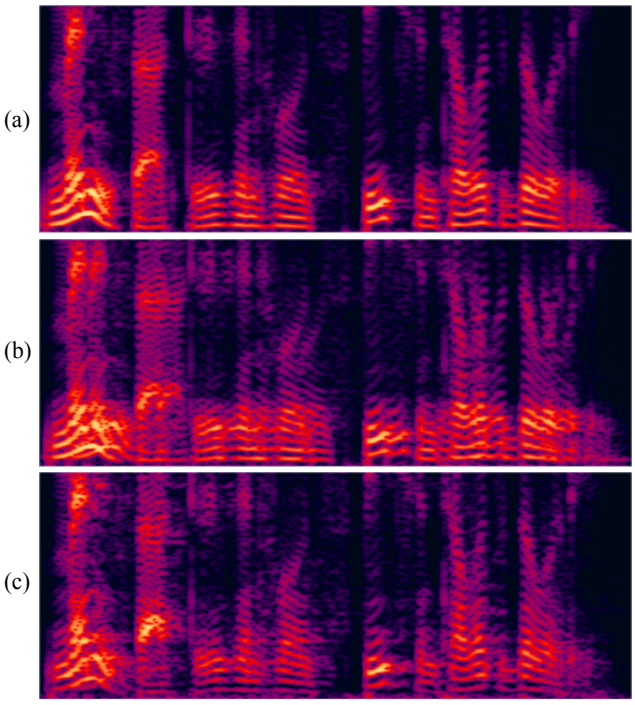

Figure 10: *(a) original speech signal, (b) reverberated, (c) equalized by the multi-channel system.*

2) use a two step process cascading the proposed algorithm with a denoiser specifically designed to attenuate the long term reverb

The LP residual approach has been proposed as a method of decoupling the harmonic structure of speech and reverb. However, it can be criticized since both the LPC and the de-reverb FIR filter are convolutional operators. Thus there is a worrying ambiguity in the identifiability of the de-reverb filter. Essentially such algorithms are making an implicit assumption that these filters are in some sense orthogonal. Alternative methods based on the transients present within the input signal may provide a better criterion. This new approach could be extended to a more general class of musical signals.

Since the generation of sound by musical instruments is often associated with impulsive like phenomena, it should be investigated if and how the metric based on the kurtosis of LP residual can be generalized.

The ML algorithm has not been optimised for real-time purposes, even if its modification in this sense seems to be straightforward. It would be interesting to develop a multi-channel versions of the algorithm with an approach similar to the one proposed in [8].

Finally other techniques of blind identification are reported in the literature. It would be of interest to apply them to the dereverberation problem and compare their results.

7. REFERENCES

[1] W. G. Gardner, "The virtual acoustic room," Master's thesis, MIT Media Lab, 1992.

[2] B. D. Radlovic, R. A. Kennedy, "Iterative cepstrum-based approach for speech de-reverberation," *Proc. of ISSPAA'99,* vol. 1, pp. 55-58.

[3] A. Farina, "Simultaneous measurements of impulse response and distortion with a swept-sine technique," *108th AES Convention*, Paris, France, 2000, Reprint 5093.

[4] S. Haykin, *Adaptive Filter Theory.* New Jersey: Prentice-Hall, 2002.

[5] O. Kirkeby and P. A. Nelson, "Digital Filter Design for Inversion Problems in Sound Reproduction," *J. Audio Eng. Soc.,* vol. 47, no. 7/8, pp. 583–595, 1999.

[6] C. L. Nikias, J. M. Mendel, "Signal processing with higher order spectra," *IEEE Signal Processing Magazine,* vol. 10, no. 3, pp. 10–37, 1993.

[7] K. Kokkinakis, V. Zarzozo, A. K. Nandi, "Blind separation of acoustic mixtures based on linear prediction analysis," *Proc. of ICA,* 2003, pp. 343–348.

[8] B. W. Gillespie, D. A. F. Florencio, H. S. Malvar, "Speech de-reverberation via maximum-kurtosis subband adaptive filtering," *Proc. of ICASSP'01*, Salt Lake City, pp. 3701–3704.

[9] A. Hyvarinen, J. Karhunen and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[10] M. Miyoshi, Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. on Acoustics, Speech and Signal Processing,* vol. 36, no. 2, pp. 145–152, 1988.