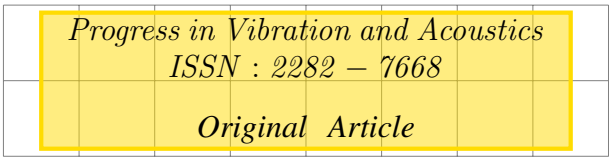
Progress in Vibration and Acoustics, September 2014, Volume 2, Issue 1, 1-37 doi: 10.12866/J.PIVAA.2014.09.001

Blind Speech Dereverberation



Massimiliano Tonelli[1]

[1] *Strada Tresole 18, 61122 Pesaro, e–mail: tonelli.acustica@gmail.com*

Abstract

Reverberation, a component of any sound generated in a natural environment, can degrade speech intelligibility or more generally the quality of a signal produced within a room. In a typical setup for teleconferencing, for instance, where the microphones receive both the speech and the reverberation of the surrounding space, it is of interest to have the latter removed from the signal that will be broadcast. A similar need arises for automatic speech recognition systems, where the reverberation decreases the recognition rate. More ambitious applications have addressed the improvement of the acoustics of theatres or even the creation of virtual acoustic environments. In all these cases dereverberation is critical. The process of recovering the source signal by removing the unwanted reverberation is called dereverberation. Usually only a reverberated instance of the signal is available. As a consequence only a blind approach, that is a more difficult task, is possible. In more precise terms, unsupervised or blind audio de-reverberation is the problem of removing reverberation from an audio signal without having explicit data regarding the system and the input signal. Different approaches have been proposed for blind dereverberation. A possible discrimination into two classes can be accomplished by considering whether or not the inverse acoustic system needs to be estimated. The aim of this paper is to show the leading research directions in blind speech dereverberation, and in particular to discuss the methods based on the explicit estimate of the inverse acoustic system, known as *reverberation cancellation techniques*. A novel dereverberation structure that improves the speech and reverberation model decoupling is also proposed. Experimental results are provided to confirm the capability of this algorithm to successfully dereverberate speech signals. [DOI:10.12866/J.PIVAA.2014.09.001] [1]

Keywords: blind dereverberation, blind multi–channel identification, deconvolution, speech enhancement

1 Introduction

In the recent years, an increasing interest in blind dereverberation techniques is observable. The aim of blind dereverberation is to estimate the original source signal by removing the reverberation components from the received signal(s), without knowledge of the surrounding acoustic

[1]Contributed by Technical Committee for publication in the Progress in Vibration and Acoustics. Manuscript received 8 August, 2014; final manuscript revised 22 August 2014; published online 2 September, 2014.

environment. A similar need arises for automatic speech recognition systems, where the reverberation decreases the recognition rate [Ferras, 2005] or for hand free communication, where the microphones receive both the speech and the reverberation of the surrounding space. Signal degradation due to reverberation is also a bottleneck for the performance and applicability of algorithms to practical problems (i.e. the cocktail party problem). A review article on blind speech dereverberation techniques was published in the 2005 by Naylor and Gaubitch in [Naylor and Gaubitch, 2005]. A review on the experimental validation of blind multimicrophone speech dereverberation was published in the 2007 by Eneman and Moonen in [Eneman and Moonen, 2007]. Books related to the blind dereverberation problem have been recently published [Benesty et al., 2005], [Huang et al., 2006a], [Benesty et al., 2008], [Naylor and Eds., 2008].

1.1 Blind dereverberation techniques, a possible classification

Even though several approaches have been proposed, a possible discrimination into two classes can be accomplished by considering whether or not the inverse IR needs to be estimated. In fact, all dereverberation algorithms attempt to obtain dereverberation by attenuating the IR effects or by undoing it. In a simplistic view, one approach tries to alleviate the *symptoms* of the signal degradation, while the other attempts to address its *cause*.

Due to the spatial diversity and temporal instability that characterize the IRs [Mourjopoulos, 1985], the first class of algorithms can offer, at the current state, more effective results in practical conditions [Thomas et al., 2007] [Naylor and Eds., 2008]. However, the algorithms belonging to the second class can potentially lead to ideal performances [Miyoshi and Kaneda, 1988]. It must be noted that practical dereverberation is still largely an unsolved problem.

To be consistent with the useful definitions reported in [Habets, 2007], the first class of algorithms will be addressed as *reverberation suppression* and the latter as *blind reverberation cancellation* methods.

Reverberation suppression methods are based on diverse set of techniques such as: beamforming [Veen and Buckley, 1988][Thomas et al., 2007], spectral subtraction [Naylor and Eds., 2008], temporal envelope [Unoki et al., 2006], LPC enhancement [Allen, 1974] [Yegnanarayana and Murthy, 2000].

Blind reverberation cancellation methods can be distinguished into two sub–classes: the techniques that are based on the IR blind estimation followed by its inversion [Huang et al., 2006b] and the ones that attempt to directly estimate the inverse system [Gillespie et al., 2001] [Yoshioka et al., 2006b] [Nakatani et al., 2003]. While the first methods have the benefit of providing the access to the IR estimation, and this is of interest for the extraction of many acoustic parameters (i.e. T_{60}, EDT, C80 etc. [H.Kutruff, 2000]), the calculation of the inverse system is not trivial even in the non blind case [Mourjopoulos et al., 1982] [Kirkeby and Nelson, 1999] [Hikichi et al., 2006] and it might lead to inaccuracies [Mourjopoulos, 1985], [Radlovic et al., 2000], [Fielder, 2003]. Therefore, it is probably more consistent for the dereverberation purpose to achieve a direct estimation of the inverse system.

Blind reverberation cancellation and suppression methods can be combined to offer hybrid strategies [Wu and Wang, 2005][Furuya and Kataoka, 2007].

Another useful distinction within dereverberation algorithms is between single or multi–channel structures. Multi-channel approaches can take advantage of spatial diversity. While it might be seem that the step leading from a single channel structure to its multi-channel version is a simple generalization, the multi–channel framework can rely on strategies not applicable to the single-channel case [Miyoshi and Kaneda, 1988], [Liu and Dong, 1997].

This article will briefly describe the main blind dereverberation techniques and introduce a novel dereverberation algorithm based on the Natural Gradient.

2 Reverberation suppression methods

2.1 Beamforming

Beamforming is a spatial filtering technique that can discriminate between signals coming from different directions. Beamforming applications are several and diverse [Veen and Buckley, 1988] (i.e. radar, sonar, telecommunication, geophysical exploration, biomedicine, image processing, acoustics). In acoustics, beamforming is employed to create sensors with an electronically configurable directivity pattern. This can be used to separate a source in a noisy environment or to minimize the interference caused by reverberation. Enhanced speech obtained from a far field acquisition is a typical application. In a similar way, this technique can be used to obtain a higher intelligibility of the diffused signal by designing a loudspeaker array that can focus the acoustic energy in a confined spatial region, minimizing the reflections due to the surrounding walls and objects.

2.1.1 FIR filters and beamforming

The simplest form of a beamformer is a linear combination of the signals acquired by an N equi-spaced linear array of omni–directional microphones.

$$y(k) = \sum_{i=1}^{N} h_i x_i(k) \tag{1}$$

where h_i is the weight and x_i the signal at the i-th sensor and $y(k)$ the beamformer output. This beamformer will be called for simplicity *linear beamformer*. This equation has the same structure of an FIR filter.

The weight h_i can be a simple real number or a more complex filter. In this last form beamforming, as it will be shown in chapter 5.2, has a close similarity to the structures employed in multichannel blind reverberation cancellation methods.

When the linear beamformer operates at a single frequency ω (narrow band hypothesis), the analogy with an FIR filter is intuitive.

- In an FIR filter, the time interval between two consecutive samples in the filtered signal is determined by the sampling period T.
- In a linear beamformer, the time interval present between two sensor is determined once it is specified the Direction Of Arrival (DOA) θ of the impinging wave, its velocity, c and the distance between the sensors, d.

The delay present between two neighboring sensors is [1]

$$\tau(\theta) = \frac{d}{c}\sin(\theta). \tag{2}$$

[1] A far field hypothesis is assumed: the sources are far enough to consider the wave as plane. This relation is not valid in the close field.

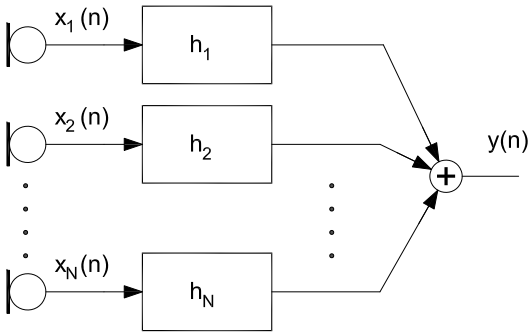


Figure 1: *Beamformer structure.*

While an FIR filter is based on the combination of uniformly spaced temporal samples, the linear beamformer is based on the combination of uniformly spaced spatial samples. As in an FIR filter, the beamformer weights h determine completely its response, and the FIR filter design techniques can be used, replacing the concept of frequency with the one of directivity.

2.1.2 FIR filter frequency response and linear beamformer directivity pattern

The frequency response of an FIR filter of length N, with impulse response h at a sampling period T is given by

$$H(\omega) = \sum_{i=1}^{N} h_i e^{-j\omega T(i-1)} \tag{3}$$

that can be written vectorially as

$$H(\omega) = \mathbf{h}^T \mathbf{d}(\omega) \tag{4}$$

where

$$\mathbf{h}^T = [h(1), h(2), \ldots, h(N)] \tag{5}$$

and

$$\mathbf{d}(\omega) = [1, e^{j\omega T}, \ldots, e^{j\omega(N-1)T}]. \tag{6}$$

The squared absolute value of the frequency response is the filter power response. To obtain the spatial response of a linear beamformer, it is sufficient to replace the sampling frequency T with $\tau(\theta) = \frac{d}{c}\sin(\theta)$

$$H(\omega) = \mathbf{h}^T \mathbf{d}(\theta, \omega) \tag{7}$$

where

$$\mathbf{d}(\theta, \omega) = [1, e^{j\omega\tau(\theta)}, \ldots, e^{j\omega(N-1)\tau(\theta)}]. \tag{8}$$

The square of the absolute value of $H(\omega)$ is the beamformer *directivity pattern* or *beam–pattern*.

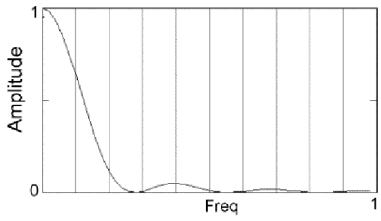


Figure 2: *Amplitude response of a simple low pass filter.*

In the temporal sampling, the highest frequency that can be represented without ambiguity is $f_{Nyquist} = f_s/2$, where f_s is the sampling frequency. If a signal is sampled with an insufficiently high f_s, the sampled signal, due to the ambiguity in the representation (aliasing), will contain frequency components that are not really present in the original signal. In a similar way, when a plane wave is spatially sampled with sensors placed at a uniform distance d, spatial aliasing can happen. It is possible to show that the highest frequency that can be reconstructed without ambiguity is

$$f_{max} = \frac{c}{2d}. \tag{9}$$

Spatial aliasing can cause undesired images in the beam–pattern.

2.1.3 A simple beamformer

A simple FIR low pass filter is given by the average of N neighboring samples

$$h(n) = [1/N, 1/N, \ldots, 1/N]\,. \tag{10}$$

The filter frequency cut is linked to the filter length N. The frequency response of this system can be calculated by 3 and its power spectrum is reported in Fig.2.

For a linear beamformer, the same $h(n)$ can be interpreted as a peak of sensitivity at 0°, as shown in Fig.3(a). The spatial resolution of the beamformer can be increased by augmenting the number of sensors, as shown in Fig. 3(b).

In the FIR filter design it is usual to weight the coefficients with a windowing function to obtain a smoother frequency response. This at the expense of frequency resolution. An optimal choice in this sense, that offers equiriple behavior in the stopband, is the Dolph–Chebyshev window [Mitra, 2002]. An example of a beamformer obtained by using the weights reported in 10 and smoothed by a Dolph-Chebyshev window is reported in Fig. 3(c).

2.1.4 Steering and 2-D beamformers

The previous example considered a beamformer with a maximum of sensitivity at 0°. If it is desired to steer the maximum to a different direction, it is sufficient to properly delay the signals acquired by the sensors. It can be shown [Veen and Buckley, 1988] that the coefficients $\mathbf{h}$ of a narrowband beamformer operating at frequency ω_0 steered to θ_0 are given by

$$\mathbf{h} = d(\theta_0, \omega_0). \tag{11}$$

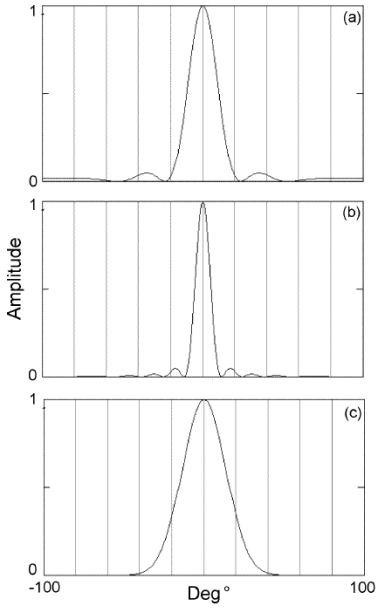


Figure 3: *Beampatterns of linear beamformers: (a)16 sensors, (b)32 sensors, (c)Dolph-Cebyshev 16 sensors.*

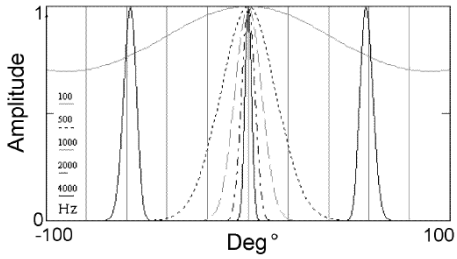


Figure 4: *Wideband behavior of a 16 sensor Dolph-Chebyshev beamformer.*

A linear beamformer can discriminate the direction of arrival from a 2–D space. A 3–D discrimination can be obtained by a 2-D matrix of equi–spaced sensors. The smoothing window coefficients are given by the factorization of the 1–D window. Therefore, the 2–D linear beamformer can be viewed as the composition of 1–D linear beamformers.

2.1.5 Beamformer response to a wide band excitation

If a narrowband beamformer designed for a specific frequency is hit by a wideband plane wave, it will appear large for high frequencies, therefore the spatial resolution will be high, and small for low frequencies, and as a consequence the spatial resolution will decrease. At high frequencies spatial aliasing might happen, and unwanted maxima of sensitivity might appear in the beampattern. The behavior of a beamformer at different frequencies is shown in Fig.4.

Since the beampattern exhibits a non constant lobe width at different frequencies, the interfering signal will not be completely suppressed, but only low-pass filtered. Design techniques for beamformers with a constant beampattern have been proposed, among them the approach of Ward et al. [Ward et al., 1995].

2.1.6 Beamforming and dereverberation

A beamformer can be used to reduce reverberation. If a beamformer is oriented toward a source positioned within a reverberant room, the reverberation that does not fall within the beampattern is attenuated. This implies the knowledge of the direction of arrival of the source. A popular method in this sense is the *delay and sum* beamformer (DSB), where the observed microphone signals are delayed to compensate for different times of arrival and then weighted and summed [Elko, 1996]. This causes the constructive summation of the components due to the direct path and the attenuation of the incoherent components due to reverberation [Elko, 1996]. It can be shown that the DSB forms a beam in the direction of the desired source.

A beamformer is however not capable of reducing the components of reverberation that fall inside the beampattern. In other words, since reverberation comes from all possible directions in a room, it will always enter the path of the beam.

Gaubitch has shown that the expected improvement in direct–to–reverberant ratio [Gaubitch, 2006] that can be achieved with a DSB is

$$E\left\{\overline{DRR}\right\} = 10\log_{10}\left(\frac{D'^2\sum_{m=1}^{M}\sum_{l=1}^{M}\frac{1}{D_m D_l}}{\sum_{m=1}^{M}\sum_{l=1}^{M}\frac{\sin(k||l_m-l_l||)}{k||l_m-l_{m+1}||}cos(k(D_m-D_l))}\right) \tag{12}$$

where D_m is the distance between the source and the m-th microphone, l_m is the m–th microphone three dimensional coordinate vector and $D^{'} = min_m(D_m)$ is the distance from the source to the closest microphone. $k = 2\pi f/c$ is the wave number with f denoting frequency and c being the speed of sound in air.

The expected improvement that can be achieved with the DSB depends only on the distance between the source and the array and the separation of the microphones and is consequently independent of the reverberation time. The performance increases by augmenting the number of microphones and the distance from the source.

In summary, beamforming and in particular the delay–and–sum beamformer are simple approaches that can provide moderate improvement in dereverberation.

2.2 Spectral subtraction

2.2.1 Noise reduction based on spectral subtraction

Spectral subtraction is not a recent approach to noise compensation and was first proposed in 1979 [Boll, 1979]. There is however a vast amount of more recent work in the literature relating to different implementations and configurations of spectral subtraction.

Spectral subtraction will be described below as summarized in [Pacheco and Seara, 2006].

Spectral subtraction is usually applied to additive noise reduction. Its main advantage is the simplicity of implementation and the low computational requirements. Speech degraded by additive noise can be represented by

$$y(n) = x(n) + v(n) \tag{13}$$

where $x(n)$ is a speech signal corrupted by the additive noise $v(n)$. Assuming that speech and noise are uncorrelated, the enhanced speech, $\hat{x}(n)$, is obtained from

$$\left|\hat{X}(k)\right| = \begin{cases} |Y(k)| - |\hat{V}(k)|, & \text{if } |Y(k)| > |\hat{V}(k)| \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

where $X(k)$, $Y(k)$ and $V(k)$, denote the short-time magnitude spectra of $x(n)$, $y(n)$ and $v(n)$ of the kth frame. These short–time magnitude spectra are usually obtained from the discrete Fourier transform (DFT) of sliding frames, typically in the order of 20–40 ms. The noise spectrum is estimated during non-speech intervals. Noise reduction is thus achieved by suppressing the effect of noise from the magnitude spectra only. The subtraction process can be in true magnitude terms or in power terms. Phase terms are ignored.

In the particular case of magnitude spectral subtraction, the enhanced speech is reconstructed by using the phase information of the corrupted signal $y(n)$

$$\hat{x}(n) = IDFT\left[|\hat{X}(k)|e^{j\angle Y(k)}\right] \tag{15}$$

where IDFT represents the inverse discrete Fourier transform.

The main inconvenience with this approach is the generation in the processed signal of an annoying interference, termed musical noise. This noise is composed of tones at random frequencies [Virag, 1999] and is mainly due to the rectification effect caused by equation 14. The spectral subtraction process can also be described as a filtering operation in the frequency–domain as [Virag, 1999]

$$X(k) = G(k)Y(k), \text{ with } 0 \leq G(k) \leq 1. \tag{16}$$

Equations, that allow one to design the filter $G(k)$ to perform magnitude subtraction, power spectral subtraction and Wiener filtering, have been similarly proposed in [Pacheco and Seara, 2006], [Habets, 2004], [Virag, 1999]. Here, the formulation proposed in [Pacheco and Seara, 2006] is reported:

$$G(k) = \begin{cases} \left\{1 - \alpha\left[\frac{V(k)}{Y(k)}\right]^{\gamma_1}\right\}^{\gamma_2}, & \text{if } \left[\frac{V(k)}{Y(k)}\right]^{\gamma_1} < \frac{1}{\alpha+\beta} \\ \beta & \text{otherwise .} \end{cases} \tag{17}$$

The following control parameters adjust $G(k)$:

- α (oversubtraction factor). It controls the amount of denoising at the expense of raising distortion.
- β (spectral flooring). Instead of assigning negative values of $|\hat{X(k)}|$ to zero, a threshold β can be set in such a way that musical noise caused by the rectification effect can be reduced.
- Exponents γ_1 and γ_2. Determines the path between $G(k) = 1$ and $G(k) = 0$. Three classical methods are defined: magnitude subtraction, with $\gamma_1 = \gamma_2 = 1$, power spectral subtraction, defined by $\gamma_1 = 2$ and $\gamma_2 = 0.5$, and Wiener filtering, with $\gamma_1 = 2$ and $\gamma_2 = 1$.

Speech distortion and residual noise cannot be minimized simultaneously. Parameter adjustment is dependent on the application. As a general rule, human listeners can tolerate some distortion, but they are sensitive to fatigue caused by noise. Automatic speech recognizers usually are more susceptible to speech distortion [Pacheco and Seara, 2006].

In [Evans et al., 2005], the limitations of spectral subtraction have been analyzed. In the same paper, the exact relation between the clean speech spectrum $X(k)$ and the noise, $V(k)$, and distorted signal, $Y(k)$, spectra is given:

$$X(k) = \left[|Y(k)|^2 - |V(k)|^2 - X(k)\cdot V^*(k) - X^*(k)\cdot V(k)\right]^{1/2} e^{\theta_X(k)}. \tag{18}$$

This expression suggests that three sources of error in a practical implementation of spectral subtraction thus exist:

- phase errors, arising from the differences between the phase of the corrupted signal $\theta_Y(e^{jw})$ and the phase of the true signal $\theta_X(k)$
- cross–term errors, from neglecting $X(k) \cdot V^*(k)$ and $X^*(k) \cdot V(k)$
- magnitude errors, which refer to the differences between the true noise spectrum $|V(k)|$ and its estimate $|\hat{V}(k)|$

Except for the worst levels of SNR, errors in the magnitude make the greatest contribution. However, as noise levels in the order of 0 dB are approached phase and cross-term errors are not negligible and lead to degradations that are comparable to those caused by magnitude errors.

2.2.2 Dereverberation based on spectral subtraction

The use of spectral subtraction for speech dereverberation of noise–free speech was proposed by Lebart et al. in [Lebart et al., 2001].

Spectral subtraction dereverberation methods are based on the observation that reverberation creates correlation between the signal measured at time $t0$ and at time $t0 + \Delta t$. Therefore, reverberation can be reduced by considering as noise the contribution of the signal at time $t0$ to the signal at time $t0 + \Delta t$. The problem of reverberation suppression differs from classical de–noising in that the *reverberation noise* is non stationary.

The non–stationary reverberation–noise power spectrum is usually based on a statistical model of late reverberation, that assumes that the room IR can be modelled as a zero-mean random sequence modulated by a decaying exponential [Polack, 1988]

$$h(n) = v(n)e^{-\tau n}u(n) \tag{19}$$

where $v(n)$ represents a white zero–mean Gaussian noise, $u(n)$, the unit step function, and τ is a damping constant related to the reverberation time T_{60}

$$\tau = 3\ln(10)/T_{60}. \tag{20}$$

The non–stationary reverberation–noise power spectrum, due to late reverberation, can be modelled as [Pacheco and Seara, 2006]

$$V(n) = e^{-2\tau T_d}X(n - T_d) \tag{21}$$

where T_d is the number of samples that identify the threshold that separates the direct component from the late reverberant one (usually between 40 and 80 ms). Therefore, it is the number of samples where reverberation is not suppressed. $V(n)$ is an exponentially attenuated power spectrum of the acquired signal $x(n)$.

To achieve an effective dereverberation, an accurate and consistent estimation of the reverberation time is necessary. Since τ is related to the reverberation time, the T_{60} should be estimated blindly from the captured signal. Different approaches have been proposed to tackle this problem [Ratnam et al., 2004], [Zhang et al., 2006], [Wen et al., 2008]. The main difficulty is the requirement of silence regions between spoken words. Particularly in short utterances, this condition may not be fulfilled, with a resulting error in the estimate of T_{60}.

In contrast to deconvolution methods, the reverberation suppression method based on spectral subtraction is not sensitive to the fluctuation of impulse responses, therefore it is more robust in practical applications. On the other hand the nonlinear processing distortion (i.e. musical noise)

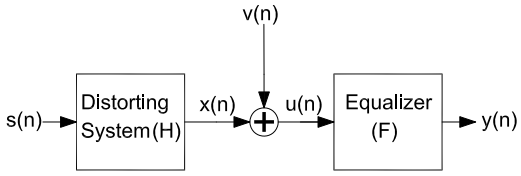


Figure 5: *Model of the blind SISO deconvolution problem.*

and the necessity of an accurate blind estimate of the reverberation time can degrade the quality of the processed reverberant speech.

Spectral subtraction is also used as a post–processing step in blind reverberation cancellation algorithms [Furuya and Kataoka, 2007], [Wu and Wang, 2005]. If this step is applied after the reverberation cancellation, as it is done in the referred papers, the exponential decaying model is not anymore valid, and a modified representation for the reverberation noise is required.

3 Reverberation cancellation methods

3.1 Blind deconvolution

Blind deconvolution techniques are an unsupervised learning approach that identifies the inverse of an unknown linear time–invariant, possibly non minimum phase, system, without having access to a training sequence (i.e a desired response). An overview of existing blind deconvolution techniques can be found in [Haykin, 1994], [Haykin, 2000], [Haykin, 2002], [Joho, 2000].

Blind deconvolution is often called, in the communication framework, blind equalization. A discrete-time model of the linear equalization problem is shown in Fig.5.

The following model assumptions are made

- The source signal $s(n)$ is a discrete–time, real, stationary stochastic process with zero mean and discrete–time power spectrum $P_s(\omega)$.
- The distorting system is linear and time invariant (LTI) with discrete–time transfer function $H(\omega)$.
- $v(n)$ is noise, statistically independent of $s(n)$, modelled as a discrete–time, real, stationary stochastic process with zero mean and power spectrum $P_v(\omega)$.
- $u(n) = x(n) + v(n)$ is the observed signal.
- The equalizer is an LTI system with discrete–time transfer function $W(\omega)$.

In a blind approach no information about $H(\omega)$ is available. Therefore, to achieve equalization, an estimate of the system transfer function or of its inverse must be found. In a noise free scenario, only second order statistics (SOS) of the received signal $x(n)$ (i.e. the system output) are needed to equalize the magnitude $|H(e^{j\omega}|$. However SOS are not sufficient to equalize the phase component. In fact, the phase response of an LTI system is not available in the output SOS. This can be observed

from the expression of the power spectrum (i.e. the Fourier transform of the autocorrelation) $P_x(e^{j\omega})$ of $x(n)$

$$P_x(e^{j\omega}) = \sum_{k=-\infty}^{\infty} r_x(k)e^{-jk\omega} \tag{22}$$

that is linked to the power spectrum of the system input $s(n)$ by the relation

$$P_x(\omega) = P_s(\omega)|H(\omega)|. \tag{23}$$

If a different LTI system $H'(\omega) = H(\omega) * A(\omega)$, where $A(\omega)$ is an allpass filter (i.e. unit magnitude and arbitrary phase response), the power spectrum $P_x(\omega)$ is unchanged. SOS cannot distinguish between $H'(\omega)$ and $H(\omega)$. This is why SOS is often described as phase-blind.

A unique relationship between the magnitude $|H(\omega)|$ and the phase $\angle H(\omega)$ of an LTI system exists only when H is either minimum phase or maximum phase (i.e. the transfer function of the system is stable and has all its zeros confined either to the interior or exterior of the unit circle in the z–plane) [Oppenheim and Schafer, 1989]. Therefore, a whitening filter, that equalizes only the magnitude response $|H(\omega)|$ of a system, is insufficient for blind equalization of mixed-phase systems.

4 Higher order statistics (HOS) methods

A possible approach to overcome the limitation of SOS and to recover the phase information is by using Higher Order Statistics (HOS).

Let $u(n), u(n+\tau_1), \ldots, u(n+\tau_{k-1})$ denote the random variables obtained by observing the process at times $n, n+\tau_1, \ldots, n+\tau_{k-1}$.

The second, third and fourth–order cumulants for a stationary random process are given by [Haykin, 2002]

$$c_2(\tau) = E\left\{u(n) \cdot u(n+\tau)\right\} \tag{24}$$

$$c_3(\tau_1, \tau_2) = E\left\{u(n) \cdot u(n+\tau_1) \cdot u(n+\tau_2)\right\} \tag{25}$$

$$\begin{aligned} c_4(\tau_1, \tau_2, \tau_3) = {} & E\left\{u(n) \cdot u(n+\tau_1) \cdot u(n+\tau_2) \cdot u(n+\tau_3)\right\} \\ & - E\left\{u(n) \cdot u(n+\tau_1)\right\} \cdot E\left\{u(n+\tau_2) \cdot u(n+\tau_3)\right\} \\ & - E\left\{u(n) \cdot u(n+\tau_2)\right\} \cdot E\left\{u(n+\tau_3) \cdot u(n+\tau_1)\right\} \\ & - E\left\{u(n) \cdot u(n+\tau_3)\right\} \cdot E\left\{u(n+\tau_1) \cdot u(n+\tau_2)\right\} \end{aligned} \tag{26}$$

The p^{th}–order moment M_p of a random variable A is

$$M^p(A) = E\left\{A^p\right\} \tag{27}$$

where $E\{\}$ is the statistical expectation.

As an example, $M^1(A) = E\{A\}$ is the mean, and $M^2(A) - (M^1(A))^2 = E\{A^2\} - (E\{A\})^2$ is the variance of A.

The generalization to a stationary random process $u(n)$ is the $k^{th} - order$ moment function R_u, defined as

$$R_u[\tau_1, \ldots, \tau_{k-1}] = E\{u(n) \cdot u(n+\tau_1) \ldots u(n+\tau_{k-1})\} \tag{28}$$

It can be observed that:

- the second–order cumulant $c_2(\tau)$ is the same as the autocorrelation function $r(\tau)$ (the second order moment function $E\{u(n) \cdot u(n+\tau]\}$);
- the third order cumulant $c_3(\tau)$ is the same as the third order moment function $E\{u(n) \cdot u(n+\tau_1) \cdot u(n+\tau_2)\}$;
- the fourth order cumulant differs from the fourth order moment function.

A $polypectrum$ of order k is defined as the k^{th}–dimensional Fourier transform of the k^{th}-order cumulant c_k [Haykin, 2002]

$$C_k(\omega_1, ..., \omega_{k-1}) = \sum_{\tau_1=-\infty}^{\infty} ... \sum_{\tau_{k-1}=-\infty}^{\infty} c_k[\tau_1, ..., \tau_{k-1}] \cdot e^{-j(\omega_1\tau_1 + ... + \omega_{k-1}\tau_{k-1})} \tag{29}$$

for $k = 2$, the ordinary power spectrum is obtained

$$P(\omega) = \sum_{k=-\infty}^{\infty} c_2(\tau) \cdot e^{-j\omega k} \tag{30}$$

for $k = 3$, we have the $bispectrum$

$$C_3(\omega_1, \omega_2) = \sum_{\tau_1=-\infty}^{\infty} \sum_{\tau_2=-\infty}^{\infty} c_3[\tau_1, \tau_2] \cdot e^{-j(\omega_1\tau_1 + \omega_2\tau_2)} \tag{31}$$

and the $trispectrum$ for $k = 4$

$$C_4(\omega_1, \omega_2, \omega_2) = \sum_{\tau_1=-\infty}^{\infty} \sum_{\tau_2=-\infty}^{\infty} \sum_{\tau_3=-\infty}^{\infty} c_4[\tau_1, \tau_2, \tau_3] \cdot e^{-j(\omega_1\tau_1 + \omega_2\tau_2 + \omega_3\tau_3)}. \tag{32}$$

Cumulants and polyspectra can be considered respectively as a generalization of the autocorrelation function and of power spectra.

When a real–valued stationary random process is considered, its power spectrum is real, therefore no phase information can be extracted from it,on the other hand, it can be shown that polyspectra preserve phase information. In particular, for a bispectrum the following relationship holds [Nikias and Raghuveer, 1987]

$$B_x(\omega_1, \omega_2) = B_s(\omega_1, \omega_2) H(\omega_1) H(\omega_2) H^*(\omega_1 + \omega_2) \tag{33}$$

where, the higher–order spectrum of the received signal $x(n)$ is linked to the the higher-order spectrum of the source signal $s(n)$ by a complex valued relation from which both magnitude and

phase of the transfer function $H(\omega)$ can be identified. In [Mendel, 1991] a closed form for an FIR model of the identified system impulse response is given

$$h(n) = \frac{R_y[P, n]}{R_y[-P, P]} \qquad n = 0, \ldots, P \tag{34}$$

where y is the system output and P the FIR model order.

Unfortunately, equation 34 has limited practical use as the model order P must be known, and the estimate of the moment function $R_y[p, n]$ comes with large variances, even without noise present. Nevertheless, it demonstrates the usability of HOS for blind system identification.

A review of blind identification methods based on the explicit use of higher order spectra can be find in Giannakis [Giannakis and Mendel, 1989], Mendel [Mendel, 1991] and in Nikias and Mendel [Nikias and Mendel, 1993]. See [Hatzinakos and Nikias, 1994] Hatzinakos and Nikias for the application to deconvolution.

A different approach started with the work of Wiggins [Wiggins, 1978], where it was shown that blind estimation of the equalizer $W(\omega)$ could be achieved by maximizing the non–Gaussianity of the received signal. A possible measure of the non-Gaussianity is Kurtosis, a measure of the *peakedness* of the probability distribution of a real-valued random variable, was proposed as a metric for the non-Gaussianity. Kurtosis is defined[2] as the fourth moment around the mean divided by the square of the variance (that is the second moment) of the probability distribution minus 3

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3 \,, \tag{35}$$

therefore it implies the computation of the fourth and second moments only. The previous expression is also know as kurtosis *excess* and is commonly used because γ_2 of a normal distribution is equal to zero, while it is positive for distributions with heavy tails and a peak at zero, and negative for flatter densities with lighter tails. Distributions of positive [negative] kurtosis are thus called super-Gaussian [sub-Gaussian]. A signal with sparse peaks and wide low level areas is characterized by a high positive kurtosis values.

Donoho [Donoho, 1981] provided a statistical foundation to Wiggins' method, by pointing out that, by the central limit theorem [Papoulis, 1984], a filtered version of a non-Gaussian i.i.d. process appears *more Gaussian* than the source itself. He also concluded that general HOS can be used to reflect the amount of Gaussianity of a random variable.

Later Shalvi and Weinstein [Shalvi and Weinstein, 1990] provided a theoretical foundation to the non Gaussianity maximization approach extending it to any non–Gaussian source signal and a necessary and sufficient condition for blind deconvolution of nonminimum phase linear time–invariant systems.

A different class of HOS–based blind deconvolution techniques indirectly exploits higher order statistics of the received signal, by employing a static non linear function.

This approach is due to Bellini [Bellini and Rocca, 1986], [Bellini and Rocca, 1988], [Bellini, 1994]. In these papers, a technique for blind deconvolution, which is efficient when the input sequence is i.i.d. and the channel distortion is small, is proposed. The algorithms are know as Bussgang algorithms because the deconvolved signal exhibits Bussgang statistics when the algorithm converges in the mean value.

Let us recollect the definition given before:

$s(n)$ - the source signal

[2]Other definitions of kurtosis exist (i.e. $\frac{\mu_4}{\mu_2^2}$ or $\mu_4 - 3\mu_2^2$ [Cadzow, 1996])

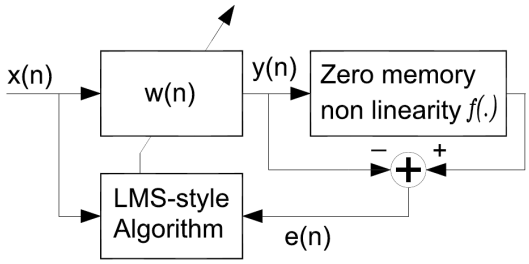


Figure 6: *Bussgang type equalizer structure.*

$x(n)$ - the received signal

$y(n)$ - the estimate of the source signal (the equalized signal)

$w(n)$ - the equalizer impulse response

The Bussgang SISO algorithm for a LTI system is

$$y(n) = \sum_{p=0}^{P} w(p)e(n-p) \tag{36}$$

and

$$w_{new}(p) = w(p) + \mu f(y(k))x(k-p) \tag{37}$$

where $e(n) = f(y(n)) - y(n)$ is the estimation error, P is the equalizer order, μ the adaptation parameter and $f(.)$ the Bussgang nonlinearity [3]

The standard Bussgang algorithm has a very slow convergence. To address this problem, Amari et al. proposed in [Amari et al., 1996] an on–line adaptive algorithm for blind deconvolution, called Natural Gradient (NG). The same algorithm was also discovered by Cardoso et al. and described in [Cardoso and Laheld, 1996] as the *relative gradient*.

The SISO Natural Gradient Algorithm (NGA) for a LTI system is [Amari et al., 1997]

$$y(n) = \sum_{p=0}^{P} w(p)e(n-p) \tag{38}$$

$$u(n) = \sum_{m=0}^{P} w(P-m)y(n-m) \tag{39}$$

$$w_{new}(p) = w(p) + \mu\Big(w(p) + f(y(n-P))u(n-p)\Big). \tag{40}$$

The standard gradient descent (as in the standard Bussgang algorithm) is most useful for cost functions that have a single minimum and whose gradients are isotropic in magnitude with respect to any direction away from this minimum. In practice, however, the cost function being optimized is multi–modal, and the gradient magnitudes are non–isotropic about any minimum. In such a case, the parameter estimates are only guaranteed to locally-minimize the cost function, and convergence to any local minimum can be slow. The natural gradient adaptation modifies the standard

[3]The - $\tanh(.)$ or the -sign. functions is usually assumed for super–Gaussian source signals.

gradient search direction according to Riemannian structure of the parameter space. While not removing local cost function minima, natural gradient adaptation provides isotropic convergence properties about any local minimum independently of the model parametrization and of the dependencies within the signals being processed by the algorithm. Moreover, natural gradient adaptation overcomes many of the limitations of Newton's method, which assumes that the cost function being minimized is approximately locally *quadratic*. By providing a faster rate of convergence, the natural gradient increases the usability of Bussgang type methods to non stationary environments.

In [Bell and Sejnowski, 1995] Bell and Sejnowski derive a self–organising learning algorithm which maximises the information transferred in a network of non–linear units to perform blind deconvolution cancellation of unknown echoes and reverberation in a speech signal. However the examples reported in their paper are restricted to the deconvolution of unrealistically short IRs.

It must be highlighted that a non Gaussian source signal $s(n)$ is required for all the HOS blind identification and deconvolution methods. In fact for a Gaussian distribution with expected value μ and variance σ^2, the cumulants are $k_1 = \mu$, $k_2 = \sigma^2$, and $k_3 = k_4 = \ldots = 0$. Therefore no information is present in higher order cumulants/moments.

4.1 Reverberation cancellation based on HOS methods

Single channel reverberation cancellation that can be referred to blind deconvolution HOS approaches are reported in [Gillespie et al., 2001],[Wu and Wang, 2005], [Fee et al., 2006]. All these methods are based on an LPC pre-whitening step followed by a blind deconvolution algorithm that performs the non-Gaussianity maximization of the received signal in a similar fashion to what suggested by Wiggins [Wiggins, 1978], and discussed in section 4.

In [Gillespie et al., 2001] the kurtosis of the reverberant residual was proposed by Gillespie et al. as a reverberation metric. It was observed that for clean voiced speech, LP residuals have strong peaks corresponding to glottal pulses, whereas for reverberated speech such peaks are spread in time. A measure of amplitude spread of LP residuals can serve as a reverberation metric. By building a filter that maximize the kurtosis of the reverberant residual it is theoretically possible to identify the inverse function of the RTF and thus to equalize the system. This approach is blind since it requires only the evaluation of the kurtosis of the reverberant residual of the system output. Gillespie's observation can be explained by considering that by its nature, reverberation is the process of summing a large number of attenuated and delayed copies of the same signal. Thus, by the central limit theorem [Donoho, 1981][Papoulis, 1984], the reverberated signal has a more Gaussian distribution in 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. Therefore reverberation causes this signal to assume a more Gaussian distribution. While the performance of these methods is quite limited in the SISO case, a considerable improvement is offered by their extension to a multichannel framework. This allows one the use of both spatial and temporal diversity to achieve better deconvolution [Gillespie et al., 2001].

5 Multi-channel SOS methods

If a multichannel system is considered, SOS can be used both for system identification and deconvolution.

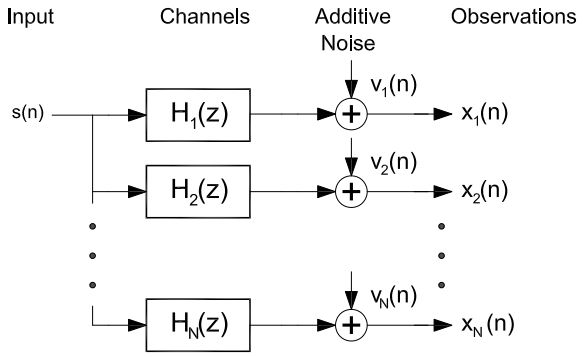


Figure 7: *Illustration of the relationships between the input $s(n)$ and the observations $x(n)$ in an M-channel SIMO system.*

5.1 SIMO identification

5.1.1 The cross relation and other subspace techniques

In [Xu et al., 1995], based on the results in [L.Tong et al., 1991], Xu, Tong et al. proposed a method to blindly identify a set of FIR filters. The method relies on the fact that all the outputs from a multiple channel FIR system are correlated if driven by the same input. Any pair of different noise free instantiation of the same source signal $s(n)$ is linked by the following relations

$$x_i(n) = h_i(n) * s(n) \quad x_j(n) = h_j(n) * s(n) \tag{41}$$

then

$$x_i(n) * h_j(n) = s(n) * h_i(n) * h_j(n) = x_j(n) * h_i(n); \quad i, j = 1, 2, ..., M;\ i \neq j. \tag{42}$$

From this relation, an overdetermined set of linear equations, with h_i, h_j as unknowns, can be written [Xu et al., 1995]. For $n = L, \ldots, N$, where N is the last sample index of the received data $x_i(n)$ and $x_j(n)$ and L is the maximum length for the channel impulse response, we have $N-L+1$ linear equations

$$\begin{bmatrix} \mathbf{X}_i(L) & \vdots & -\mathbf{X}_j(L) \end{bmatrix} = \begin{bmatrix} \mathbf{h}_j \\ \mathbf{h}_i \end{bmatrix} \tag{43}$$

where $\mathbf{h}_m = [h_m(L), \ldots, h_m(0)]^T$ and

$$\mathbf{X}_m(L) = \begin{bmatrix} x_m(L) & x_m(L+1) & \ldots & x_m(2L) \\ x_m(L+1) & x_m(L+2) & \ldots & x_m(2L+1) \\ \vdots & \vdots & \ddots & \vdots \\ x_m(N-L) & x_m(N-L+1) & \ldots & x_m(N) \end{bmatrix}. \tag{44}$$

Equation 43 can be written for each pair of channels (i, j). The equations of all channels can be combined and a larger set of linear equations in terms of $\mathbf{h}_1, ..., \mathbf{h}_L$ or simply $\mathbf{h} = \left[\mathbf{h}_1^T, \ldots, \mathbf{h}_L^T\right]^T$, so that all the channel impulse responses can be calculated simultaneously

$$\mathbf{X}(L)\mathbf{h} = 0 \tag{45}$$

where $\mathbf{h}$ is the matrix of the impulse responses, and $\mathbf{X}(L)$ the matrix containing the received signals.

$$\mathbf{X}(L) = \begin{bmatrix} \mathbf{X}^1(L) \\ \vdots \\ \mathbf{X}^{M-1}(L) \end{bmatrix} \tag{46}$$

with

$$\mathbf{X}^i(L) = \begin{bmatrix} 0 & \dots & 0 & X_{i+1}(L) & -X_i(L) & 0 & 0 \\ \vdots & & \vdots & 0 & \ddots & 0 & \\ 0 & \dots & 0 & X_M(L) & 0 & \dots & -X_i(L) \end{bmatrix}. \tag{47}$$

The necessary and sufficient conditions to ensure a unique solution to the above equation, or in other words to assure identifiability, are [Xu et al., 1995]:

1. the channel transfer functions do not share any common zeros;
2. the autocorrelation matrix of the source signal $R_{ss} = E\{s(k)s^T(k)\}$ is of full rank (such that the SIMO system can be fully excited).

The first condition is the coprimeness, or channel diversity, identifiability condition for a SIMO system [Miyoshi and Kaneda, 1988]. The second condition is relatively mild and does not imply the knowledge of the exact statistic of the input signal, nor even constrain it to be an i.i.d process. Therefore, theoretically, any input signal that can fully excite the SIMO system can be employed. These conditions are sufficient for the blind identification of any SIMO system.

This approach is known in the literature as the Cross Relation (CR) approach. The CR approach, a termed coined by Hua [Hua, 1996], was discovered independently and in different forms by several authors, among them, by Liu et al. [Liu et al., 1993], and Gurreli and Nikias [Gurelli and Nikias, 2001]. These algorithms, originally aimed at solving communication problems, are often referred to as *deterministic subspace methods*, since the statistical properties of the source are not exploited. Subspace algorithms are based on the idea that the channel (or part of the channel) vector is in a one-dimensional subspace of either the observation statistics or a block of noiseless observations.

Some of the subspace techniques, such as the EVAM algorithm proposed by Gurelli and Nikias [Gurelli and Nikias, 2001], have been used in dereverberation problems. Gurelli and Nikias showed that the null space of the correlation matrix of the received signals contains information on the transfer function relating the source and the microphones. This was extended by Gannot and Moonen [Gannot and Moonen, 2001], [Gannot and Moonen, 2003] to the speech dereverberation problem.

Even if these techniques are supported by theory, they have several drawbacks in real-life scenarios. The Generalized Eigenvalue Decomposition (GED) [Golub and Loan, 1996], which is used to construct the null space of the correlation matrix, is not robust enough, and quite sensitive to small estimation errors in the correlation matrix. Furthermore, the matrices involved become extremely large causing severe memory and computational requirements. Another problem arises from the wide dynamic range of the speech signal. This phenomenon may result in an erroneous estimate of the frequency response of the IRs in the low energy bands of the input signal.
In a following paper [Eneman and Moonen, 2007], Moonen and Eneman showed that, at the current state, even the more advanced subspace-based dereverberation techniques did not provide, in a real-life scenario, any signal enhancement. Furthermore, even if most subspace methods can converge quickly, they are difficult to implement in an adaptive mode and have a high-computational load [Tong and Perreau, 1998] [4].

[4] A review on other subspace methods can be found in the same paper.

5.1.2 Adaptive blind channel identification techniques

To overcome these limitations, Huang and Benesty [Huang et al., 2006b] proposed a set of adaptive algorithms to solve the set of linear equation obtained from the CR approach. Their work started from the formulation of the CR approach described in [Avendano et al., 1999], where the channel impulse responses of an identifiable system are blindly determined by calculating the null space of the cross-correlation like matrix of channel outputs

$$\mathbf{R}_x \mathbf{h} = 0 \tag{48}$$

with

$$\mathbf{R_x} = \begin{bmatrix} \sum_{i\neq 1} \mathbf{R}_{x_i x_i} & -R_{x_2 x_1} & \cdots & -R_{x_M x_1} \\ -\mathbf{R}_{x_1 x_2} & \sum_{i\neq 2} R_{x_i x_i} & \cdots & -R_{x_M x_2} \\ \vdots & \vdots & \ddots & \vdots \\ -\mathbf{R}_{x_1 x_M} & -R_{x_2 x_M} & \cdots & \sum_{i\neq M} R_{x_i x_i} \end{bmatrix} \tag{49}$$

where M is the number of channels

$$\mathbf{R}_{x_i x_i} = E\left\{x_i(n) x_j^T(n)\right\}; \quad i,j = 1,2,...,M \tag{50}$$

and

$$\mathbf{h} = \left[h_1^T, \ldots, h_M^T\right]^T \tag{51}$$

is the matrix of the impulse responses. For a blindly identifiable SIMO system, matrix R_x is rank deficient by 1. In the absence of noise, the channel impulse responses can be uniquely determined from R_x, which contains only the SOS of the system outputs.

By following the fact that

$$x_i(n) * h_j(n) = s(n) * h_i(n) * h_j(n) = x_j(n) * h_i(n); \quad i,j = 1,2,...,N;\ i \neq j; \tag{52}$$

we have, in the absence of noise, the following cross relation at time k

$$x_i^T(k) * \mathbf{h}_j = x_j^T(k) * \mathbf{h}_i; \quad i,j = 1,2,...,N;\ i \neq j. \tag{53}$$

When noise is present and/or the estimate of channel impulse responses deviates from the true value, an a priori error signal is produced

$$e_{ij}(k+1) = x_i^T(k+1)\hat{\mathbf{h}}_j(k) - x_j^T(k+1)\hat{\mathbf{h}}_i(k); \quad i,j = 1,2,\ldots,N; \tag{54}$$

where $\hat{\mathbf{h}}_i(k)$ is the model filter for the i-th channel at time k. The estimated channel impulse response vector is aligned to the true one, but up to a non-zero scale. This inherent scale ambiguity is usually harmless in most of acoustic signal processing applications. But in the development of an adaptive algorithm, attention needs to be paid to prevent it from converging to a trivial all–zero estimate. Therefore, a constraint can be imposed on the model filter. Two constraints can be found in the literature. The unit-norm constraint, i.e. $\left\|\hat{\mathbf{h}}\right\| = 1$, and the component normalization constraint [Avendano et al., 1999], i.e. $\mathbf{c}^T\hat{\mathbf{h}} = 1$, where c is a constant vector. The unit–norm constraint can be explained by the Parseval's theorem

$$\int_{-\infty}^{\infty} |x(t)|^2 dt = \int_{-\infty}^{\infty} |X(f)|^2 df \tag{55}$$

where $X(f) = \mathcal{F}\{x(t)\}$ represents the continuous Fourier transform (in normalized, unitary form) of $x(t)$ and f represents the frequency component of x. The interpretation of this form of the

theorem is that the total energy contained in a waveform $x(t)$ summed across all of time t is equal to the total energy of the waveform's Fourier Transform $X(f)$ summed across all of its frequency components f. For a discrete time system, the Parseval's theorem is

$$\sum_{n=0}^{N-1} |x(n)|^2 = \frac{1}{N} \sum_{k=0}^{N-1} |X[k]|^2 \tag{56}$$

where $X[k]$ is the Discrete Fourier Transform (DFT) of $x(n)$, both of length N. Therefore, the unitary–norm constraint

$$\left\|\hat{\mathbf{h}}\right\| = \sum_{n=0}^{N-1} |x(n)|^2 = 1 \tag{57}$$

imposes a unitary constraint on the filter power.

The component normalization constraint is useful when one coefficient of the model filter is know to be equal to α, which is not zero. Then the vector $\mathbf{c} = [0, \ldots, 1/\alpha, \ldots, 0]^T$ can be properly specified, so that $\mathbf{c}^T\hat{\mathbf{h}} = 1$. Even though the component normalization can be more robust to noise than the unit–norm constraint [Avendano et al., 1999], the knowledge of the location of the component to be normalized and its value α may not be available in practice. So the unit-norm constraint is more widely used.

With the unitary–norm constraint enforced on $\left\|\hat{\mathbf{h}}\right\|$, the normalized error signal is

$$\epsilon_{ij}(k+1) = e_{ij}(k+1) / \left\|\hat{\mathbf{h}}(k)\right\| \tag{58}$$

accordingly the cost function is formulated as

$$J(k+1) = \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \epsilon_{ij}^2(k+1) \tag{59}$$

the update equation of the algorithm is then given by

$$\nabla J(k+1) = \frac{\partial J(k+1)}{\partial \hat{\mathbf{h}}(k)} = \frac{2\left[\tilde{\mathbf{R}}_x(n+1)\hat{\mathbf{h}}(n) - J(n+1)\hat{\mathbf{h}}(n)\right]}{\left\|\hat{\mathbf{h}}(k)\right\}} \tag{60}$$

where $\tilde{\mathbf{R}}$ is a matrix with the same structure of the $\mathbf{R}_x$ matrix in equation 49, but built with the instantaneous values of the received signals

$$\tilde{\mathbf{R}}_{x_i x_j} = x_i(k+1)x_j^T(k+1); \quad i, j = 1, 2, \ldots, N. \tag{61}$$

This algorithm is known as the Multichannel LMS algorithm (MCLMS). Several algorithms based on the MCLMS algorithm, and that outperform it, have been proposed by the same authors. All of them are documented and discussed in [Huang et al., 2006b]. Among them:

- the Unconstrained Multichannel LMS algorithm with optimal step size (VSS-UMCLMS), that has faster convergence and similar computational load
- the Constrained Multichannel Newton (CMN) algorithm, that has much faster convergence but high computational load

- the normalized multichannel frequency domain LMS (NMCFLMS), that takes advantage of the computational efficiency of the FFT, which by orthogonalizing the data offers also faster convergence [Haykin, 2002].

The NMCFLMS algorithm has been documented as being able to identify rather short (256 taps) impulse responses by using a 100 second long speech voice as the source signal [Huang et al., 2006b].

One of the main challenges for NMCFLMS is that the algorithm suffers from a misconvergence problem. It has been shown through simulations presented in [Hasan et al., 2005], [Ahmad et al., 2006], that the estimated filter coefficients converge first toward the impulse response of the acoustic system but then misconverge. Under low signal-to-noise ratio (SNR) conditions, the effect of misconvergence becomes more significant and occurs at an earlier stage of adaptation. Possible solutions to this problem have been investigated by Gaubitch et al. in [Gaubitch et al., 2006], [Gaubitch et al., 2005] and by Ahmad et al. [Hasan et al., 2005] [Ahmad et al., 2007a],[Ahmad et al., 2006]. In [Ahmad et al., 2007b] a noise robust adaptive blind multichannel identification algorithm for acoustic impulse responses, called ext-NMCFLMS, based on a modification of the NMCFLMS has been proposed. No information about the performance with longer acoustic IR is available.

5.2 SIMO equalization

The previous approaches do not directly address the problem of system equalization. In fact, a further step is required to calculate the equalizer once the FIR filters have been estimated.

5.2.1 A fundamental theorem for multiple-channel blind equalization

The problem of SIMO system equalization has been investigated by Slock and Papadias in [Slock et al., 1995], where it is shown that a blind equalization can be achieved by linear prediction algorithm on the channel bank outputs.

A more general approach to blind multichannel equalization, that provides a unifying framework for multichannel blind equalization, is reported in Liu and Dong [Liu and Dong, 1997]. In this paper, it is proved that, if no common zeros are shared among channels transfer functions h_i and if the source signal $s(n)$ is zero-mean and temporally uncorrelated, then an FIR equalizer bank w_i equalizes the FIR channel bank if, and only if, the composite output of the equalizer bank $y(n)$ is temporally uncorrelated. It is interesting to note that the condition of existence for the filter bank is identical to the one requested by the multiple input /output inverse theorem (MINT) [Miyoshi and Kaneda, 1988] . While the MINT gives a closed formula to calculate the equalizer in a non blind framework, this theorem suggests how to calculate them in the blind case. The advantage of a multichannel structure in respect to a single channel one when a non minimum phase system is considered, is present in the blind case also.

Important aspects that stem out from this theorem are:

- the second–order statistics of the composite output of the equalizer bank has sufficient information for the blind equalization, though it does not have sufficient information for blind identification

- the hypotheses are very mild: the source signal needs not be stationary, nor i.i.d.

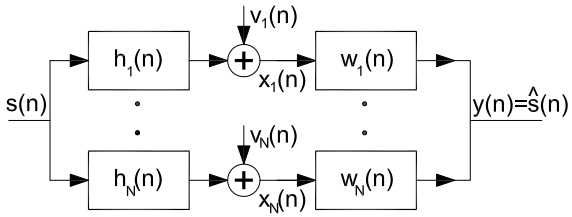


Figure 8: *Multichannel blind equalizer for a SIMO system.*

- since the linear prediction error is temporally uncorrelated, the method proposed by Slock and Papadias [Slock et al., 1995] that achieves direct blind equalization by multichannel linear predication follows immediately from the previous theorem
- no constraints on how to obtain the output decorrelation exist; therefore both linear and non-linear approaches can be used.

All the following blind speech dereverberation algorithms are based, or can be explained, by the Liu and Dong [Liu and Dong, 1997] theorem. In all of them in fact, the FIR equalizer bank is calculated to equalize the FIR channel bank by decorrelating the composite output of the equalizer bank $y(n)$. This decorrelation is achieved by multichannel linear prediction [Triki and Slock, 2006],[Delcroix et al., 2004], or by decorrelating the composite filter bank output both by SOS [Yoshioka et al., 2006a] and HOS [Yoshioka et al., 2006b] methods, or by shaping the correlation of the received reverberant signal [Gillespie and Atlas, 2003]. The main novelty of these contributions is on the techniques that are used to decouple the speech production system and the room response system to avoid ambiguous deconvolution.

In [Triki and Slock, 2006] Triki and Slock proposed a dereverberation technique based on the observation that a single-input multi-output (SIMO) system is equalized blindly by applying multichannel linear prediction (LP). This approach follows [Slock et al., 1995] and the Liu and Dong theorem [Liu and Dong, 1997]. However when the input is colored, the multichannel linear prediction will both equalize the reverberation filter and whiten the source. Therefore, it is critical to define a criterion to decouple the source and the reverberation. To address this problem, channel spatial diversity and the speech signal non-stationarity were employed to estimate the source correlation structure, which can hence be used to determine a source whitening filter. Multichannel linear prediction was then applied to the sensor signals filtered by the source whitening filter, to obtain source dereverberation. The algorithm was tested with synthetic impulse responses generated by a MISM algorithm, and in the simulation it is shown that the proposed equalizer outperforms the delay and sum beamformer . In a following work [Triki and Slock, 2007] the dereverberation in a noisy environment was considered and the robustness of the algorithm was improved by considering an equalizer based on an MMSE criterion instead of a simple ZF equalizer.

A similar approach based on multi–channel LP was considered by Delcroix et al. in [Delcroix et al., 2004],[Delcroix et al., 2005], where a two–channel dereverberation algorithm called LInear-predictive Multi-input Equalization (LIME) is proposed. In their approach multi–channel LP was used, but allowing the source to be whitened, while restoring the coloration in a final stage.

In [Yoshioka et al., 2006b] and in [Yoshioka et al., 2006a] by Yoshioka et al. proposed two similar methods, one HOS and the other SOS based, applied to the same multichannel structure, to calculate the dereverberation filters. Both the methods rely on the Liu and Dong [Liu and Dong, 1997] theorem since the dereverberation filter is calculated by uncorrelating the composite output of the equalizer bank. The fundamental issue addressed in both papers is how to estimate a channel's inverse filter separately from the inverse filter of the speech generating AR system, or in other words from the prediction error filter (PEF). The authors claim that by jointly estimating the channel inverse filter and the PEF, the channel inverse is identifiable due to the time varying nature of the PEF.

In [Gillespie and Atlas, 2002] Gillespie et al. showed that penalizing long-term reverberation energy is more effective than maximizing the signal-to-reverberation ratio (SRR) for improving audible quality and automatic speech recognition (ASR) accuracy. In a following paper [Gillespie and Atlas, 2003], they noticed that the energy in the tail of the autocorrelation sequence of the received reverberant signal is related to the amount of long–term reverberation. Based on these observations, a technique, called Correlation Shaping (CS), aimed to reshape the autocorrelation function of a signal by means of linear filtering was proposed. This has the intended effect of reducing the length of the equalized speaker–to–receiver impulse response to improve audible quality and ASR accuracy blindly. Dereverberation can, therefore, be achieved by reshaping the autocorrelation of the linear prediction residual to a Dirac δ function, that is, whitening it. Therefore, this multichannel algorithm also relies on the Liu and Dong [Liu and Dong, 1997] theorem.

As a final comment on SIMO equalization methods, it is interesting to notice that also some HOS multichannel methods can be explained by the Liu and Dong [Liu and Dong, 1997] theorem, except that they don't guarantee that the output is uncorrelated. For instance, the kurtosis maximization algorithm used in the multichannel dereverberation algorithm discussed in section 4.1 and described in [Gillespie et al., 2001] can be also explained as a way to uncorrelate the composite output of the equalizer bank $y(n)$.

5.2.2 A blind MINT approach to multiple-channel blind equalization

A different approach, that cannot be explicitly explained by Liu and Dong [Liu and Dong, 1997] theorem was proposed in [Furuya, 2001] by Furuya et al. This is a promising algorithm, based on the blind MINT approach. The conventional MINT [Miyoshi and Kaneda, 1988] requires the room impulse responses to calculate the inverse filters, so it cannot recover speech signals, when the room impulse responses are unknown. However, as suggested by other SOS methods, the inverse filters can be blindly estimated from the correlation matrix between input signals, that can be observed.

It is meaningful to point out that this is the only reverberation cancellation algorithm, among the ones described so far, that was tested in a realistic environment.

The deconvolution based on inverse filtering does not improve the tail of reverberation because impulse responses are always fluctuating in the real world and the estimation error of inverse filters is caused by deviation of the correlation matrix averaged for a finite duration. As a possible improvement, in [Furuya and Kataoka, 2007] an hybrid reverberation cancellation/suppression method is proposed. A modified spectral subtraction algorithm is cascaded to the blind MINT deconvolution algorithm. Spectral subtraction estimates the power spectrum of the reverberation and then subtracts it from the power spectrum of reverberant speech. Inverse filtering reduces early reflection, which has most of the power of the reverberation, and then, spectral subtraction suppresses the tail of the inverse–filtered reverberation. Inverse filtering reduces the power of the reverberation, so the nonlinear processing distortion of spectral subtraction is reduced using a small

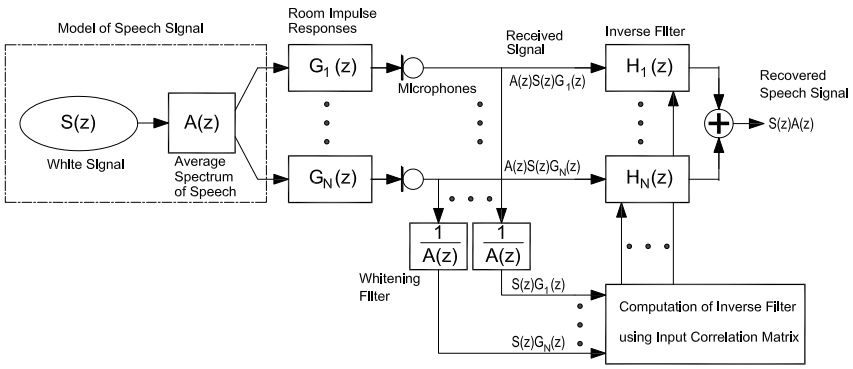


Figure 9: *Signal flow of the method proposed by Furuya et al.*

subtractive power. The authors claim superior dereverberation results for every reverberation time. On the other hand, even if the algorithm is effective and robust in situations requiring adaptation, the adaptation speed is still slow for practical applications.

5.3 HOS and SOS approaches, a comparison

According to Brillinger [Brillinger, 1975], the sample size needed to estimate the n^{th}-order statistics of a random process to prescribed values of estimation bias and variance, increases almost exponentially with order n. This is why often HOS-based blind deconvolution methods exhibit a slower rate of convergence in comparison to the SOS-based ones. This can be of concern in highly non-stationary environments where an HOS-based algorithm might not have enough time to track the statistical variations. On the other hand, when a method based on the kurtosis maximization is used, we are not necessarily trying to achieve an accurate estimate of kurtosis. So performance in one case may not translate into performance in the other case. HOS methods based on the implicit non-linear processing approach (i.e. the Bussgang equalizer) are less demanding from the complexity perspective, even if they might be prone to local minima [C. R. Johnson, 1991]. An advantage of HOS-based methods is that they can be employed in SISO systems, while SOS methods require a multichannel framework, however HOS methods require a non Gaussian received signal.

6 Novel HOS based blind dereverberation algorithm

This section describes a novel single and multichannel methods based on the natural gradient and of a new dereverberation structure that improves the speech and reverberation model decoupling [Tonelli and Davies, 2010].

6.1 The single channel dereverberation problem

Let us define:

. $s(n)$ - the clean speech signal

. $x(n)$ - the reverberant speech signal

. $y(n)$ - the estimate of the clean speech signal

. $h(n)$ - the speaker-to-receiver impulse response

. $w(n)$ - the equalizer impulse response

If the acoustic path is modeled as a linear–time invariant system $s(n)$ and $x(n)$, are linked by the equation

$$x(n) = h(n) * s(n) \tag{62}$$

where $*$ denotes the discrete linear convolution. Dereverberation is achieved by finding a filter with impulse response $w(n)$ so that

$$\delta(n - N_d) = w(n) * h(n) \tag{63}$$

where $\delta(k)$ is the unit sample sequence and N_d a delay. Typically in the single channel case $w(n)$ needs to be IIR.

If $h(n)$ and $s(n)$ are unknown, dereverberation is blind. Therefore, single channel blind dereverberation is connected to the blind estimation of $w(n)$ so that

$$\hat{s}(n) = y(n) = w(n) * x(n) \tag{64}$$

where $\hat{s}(n)$ is an estimate of $s(n)$.

6.2 The multi-channel dereverberation problem

The de–reverberation problem can be generalized for an arbitrary N-input channel system, leading to the following set of relations

$$x_i(n) = h_i(n) * s(n),\ 1 \leq i \leq N \tag{65}$$

$$\hat{s}(n) = y(n) = \sum_{i=1}^{N} w_i(n) * x_i(n) \tag{66}$$

where $x_i(n)$, $h_i(n), w_i(n)$ are respectively the i-th observation, transfer function and equalizer of the corresponding source-to-receiver channel. For a multi-channel structure, equalization is achieved by finding a set of filters with impulse response $w_i(n)$ so that

$$\delta(n - Nd) = \sum_{i=1}^{N} w_i(n) * h_i(n) \tag{67}$$

this expression is know as the MINT theorem [Miyoshi and Kaneda, 1988] and it is closely related to the Bezout identity [Huang et al., 2005]

$$1 = \sum_{i=1}^{N} W_i(z) H_i(z) \tag{68}$$

where $W_i(z)$ and $H_i(z)$ are polynomials with no common zeros. In the multi-channel case exact derverberation is possible with $w_i(n)$ being FIR if the Bezout identity holds.

The algebraic decomposition that satisfies the Bezout identity is in general not unique and the non blind algorithm reported in [Miyoshi and Kaneda, 1988] calculates one of the possible solutions for the equalizers $w_i(n)$.

6.3 Estimation of the $w(n)$ coefficients

The estimation of the $w(n)$ coefficients can be obtained by the algorithm proposed in [Gillespie et al., 2001]. This algorithm is based on the observation that the kurtosis of the linear prediction residual can serve as a reasonable reverberation metric. Low kurtosis values of the residual imply a highly reverberated speech signal, therefore enabling the inverse filter to be identified by kurtosis maximization. A time domain structure for this algorithm is reported in figure 10(a). The equalizer based on the kurtosis maximization update can be replaced by different blind equalization schemes. This can be desirable, since the estimation of the kurtosis and its derivative are prone to instability.

An alternative is the Natural Gradient Algorithm (NGA) [Amari et al., 1997], that for the single channel case is:

$$y(n) = \sum_{p=0}^{P} w(p)e(n-p) \tag{69}$$

$$u(n) = \sum_{m=0}^{P} w(P-m)y(n-m) \tag{70}$$

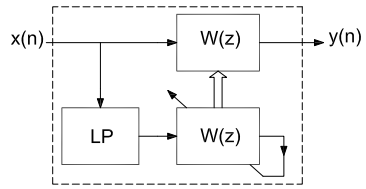
$$w_{new}(p) = w_{old} + \mu\Big(w(p) + f(y(n-P))u(n-p)\Big) \tag{71}$$

where P is the equalizer order and $f(.)$ the Bussgang nonlinearity (the $tanh(.)$ or the $sign(.)$ functions will be assumed).

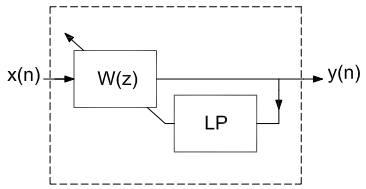
The application of such equalization schemes to blind dereverberation using the structure in figure 10(a), called from now on *forward structure* is relatively easy and consists of LPC pre–whitening followed by the equalization algorithm, e.g. (69)-(71). However better results can be achieved by using the structure shown in figure 10(b), called from now on the *reversed structure*.

7 On the correct structure for a single channel dereverberator

The order of two linear filters can be swapped only if they are time invariant. Since the vocal tract filter is not stationary, the forward and the reversed structure will lead to different results. It can be shown that the residual $e(n)$ calculated by the forward structure is a function of multiple quasi-stationary blocks, although it is being whitened by only a single LPC filter. This is particularly problematic in the dereverb setting since the duration over which speech is quasi-stationarity is usually significantly less than the room reverberation time. By performing the dereverberation before the LPC analysis (reversed structure), the modeled residual $e(n)$ is only a function of one quasi-stationary block. This observation led us to develop a new algorithm based on the natural gradient that exploits this second structure.



(a)



(b)

Figure 10: *(a) Diagram of the time domain dereverberation algorithm proposed in [Gillespie et al., 2001] (forward structure). (b) Diagram of the model proposed in this paper (reversed structure).*

7.1 Dereverberating speech

We begin by modeling the speech as a non-Gaussian i.i.d. source filtered by a non-stationary autoregressive filter. The correct relationship between $e(n)$ and $x(n)$ is:

$$\begin{aligned} e(n) &= \sum_l a_l(n) y(n-l) \\ &= \sum_l a_l(n) \sum_p w(p) x(n-l-p) \end{aligned} \tag{72}$$

where $a_l(n)$ is the $l-th$ time varying LPC filter coefficient at time n. Note that $e(n)$ is only a function of LPC coefficients associated with time n. For the moment we will assume that we know $a_l(n)$ for all n and l.

Our probability model for the reversed structure is:

$$\begin{aligned} J(w) &= -E\{\log p(x|w)\} \\ &= -\log\left\|\frac{\partial y}{\partial x}\right\| - E\{\log p(y|w)\} \\ &= -\log\left\|\frac{\partial y}{\partial x}\right\| - \log\left\|\frac{\partial e}{\partial y}\right\| - E\{\log p(e|w)\} \\ &= -\frac{1}{2\pi}\int_{-\pi}^{\pi} \log|W(\omega)| d\omega - E\{\log p(e(n))\} \\ &+ \text{constant} \end{aligned} \tag{73}$$

Differentiating this we get:

$$\begin{aligned} \frac{\partial J}{\partial w(p)} &= -h(-p) - f(e(n))\frac{\partial e(n)}{\partial w(p)} \\ &= -h(-p) - f(e(n)) \sum_l a_l(n) x(n-p-l) \end{aligned} \tag{74}$$

where $h(p)$ denotes the impulse response function for the inverse of w (i.e. $h * w = \delta_0$) and $f(.)$ a nonlinearity (the tanh(.) or the sign(.) functions will be assumed).

Now the Bussgang version (standard gradient) algorithm is:

$$w_{new}(p) = w_{old}(p) - \mu \left(\frac{\partial J}{\partial w(p)} \right) \tag{75}$$

It can be shown that an equivariant (natural gradient) form of the previous equation is:

$$\begin{aligned} w_{new}(p) = w_{old}(p) + \mu \Big(& w(p) + f(e(n)) \\ & \cdot \sum_l a_l(n) y(n-p+m-l) \Big) \end{aligned} \tag{76}$$

7.2 A causal normalized NGA dereverb algorithm

Finally we can address the causality problem, as in [Amari et al., 1997]. First we constrain the dereverb filter to be causal FIR:

$$y(n) = \sum_{p=0}^{P} w(p) x(n-p) \tag{77}$$

We then delay our update by P samples:

$$\begin{aligned} w_{new}(p) = w_{old}(p) + \mu \Big(& w(p) + f(e(n-P)) \\ & \cdot \sum_l a_l(n-P) \sum_m w(m) y(n-p+m-l-P) \Big) \end{aligned} \tag{78}$$

and introduce an auxiliary variable $u(n)$:

$$u(n) = \sum_{m=0}^{P} w(P-m) y(n-m) \tag{79}$$

We then have the simplified update rule for the reversed structure:

$$\begin{aligned} w_{new}(p) = w_{old}(p) + \mu \Big(& w(p) + f(e(n-P)) \\ & \cdot \sum_{l=0}^{L} a_l(n-P) u(n-p-l) \Big) \end{aligned} \tag{80}$$

It is of interest to compare equations (77), (79), (80), to the update equations for the forward structure, (69), (70), (71). The update equations for the reversed structure are more complex, as an additional step is required to calculate the last term in (80). However, in the forward structure, this additional filtering must be calculated outside the adaptation loop to obtain the speech residual. Furthermore, the reversed structure makes replicating the dereverberation filter unnecessary since it acts directly on the reverberated signal.

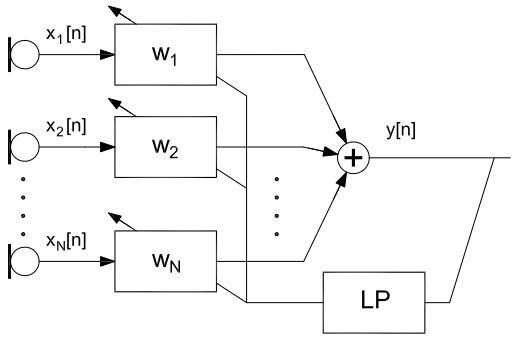


Figure 11: *proposed multi-channel de-reverberation structure.*

7.3 A multichannel de-reverberation of real speech

The benefit of the reverse algorithm becomes evident when a multichannel structure is employed. In fact, while the multichannel structure based on the forward algorithm proposed in [Gillespie et al., 2001] requires the calculation of the LP residual for each channel, the multichannel reversed structure, as shown in figure 11, requires only a single LP residual calculation, and no replication of the de-reverberation filters.

For an N-channel system the update equation becomes

$$e(n) = \sum_{l} a_l(n) y(n-l) \tag{81}$$

where $a_l(n)$ is the $l-th$ time varying LPC filter coefficient at time n and $y(n)$ is the de-reverberator output defined as

$$\hat{s}(n) = y(n) = \frac{1}{N} \sum_{i=1}^{N} y_i(n) \tag{82}$$

where $y_i(n)$ is the output vector of the i-th maximization filter

$$y_i(n) = w_i(n) * x_i(n) \tag{83}$$

and $x_i(n)$, $w_i(n)$ are respectively the i-th observation and equalizer of the corresponding source-to-receiver channel. The $p-th$ coefficient of the $i-th$ channel maximization filter, w_i, is give by

$$\begin{aligned} w_{i_{new}}(p) = w_i(p) + \mu \Big(& w_i(p) + f(e(n-P)) \\ & \cdot \sum_{l=0}^{L} a_l(n-P) u_i(n-p-l) \Big) \end{aligned} \tag{84}$$

and

$$u_i(n) = \sum_{m=0}^{P} w_i(P-m) y_i(n-m) \tag{85}$$

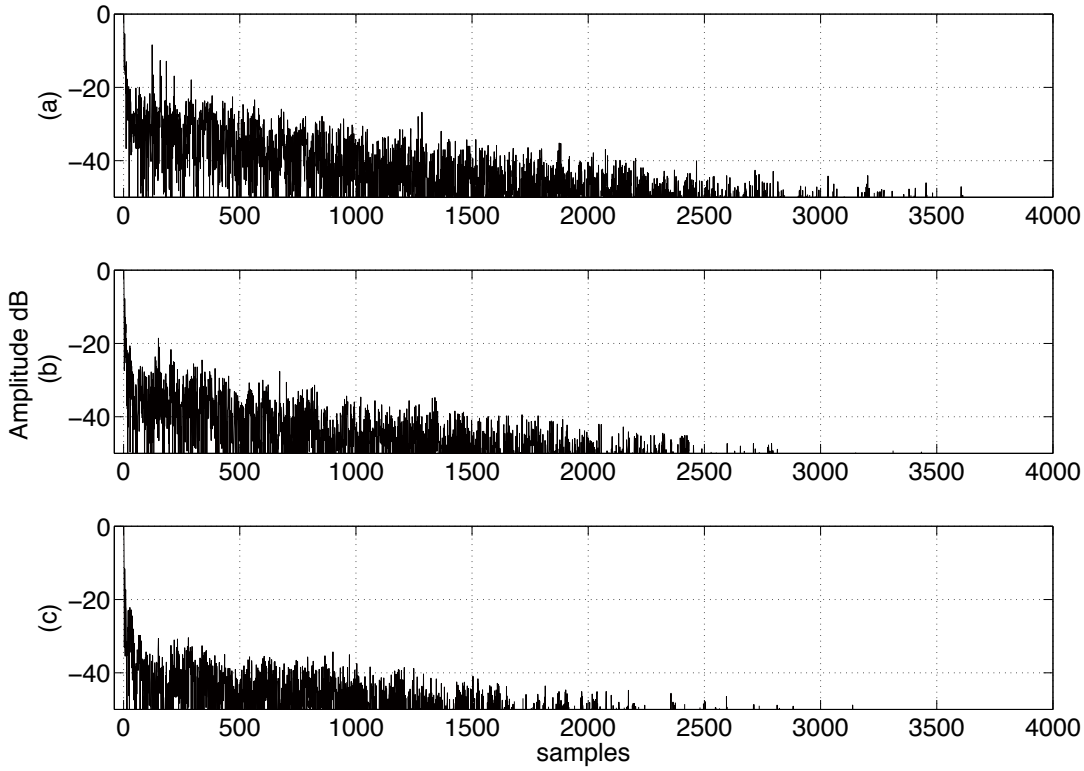


Figure 12: *(a) Reference echogram relating to the shortest source-to-receiver path, DRR=-2.9dB (b) 8-channel delay-sum beamfomer, DRR =-0.1dB. (c) proposed 8-channel structure de-reverberator, DRR =3.1dB*

8 Results

8.1 Blind deververberation of real speech

The proposed multichannel algorithm was used to dereverberate a speech signal convolved with 6000-tap impulse responses measured, by a linear array composed of eight microphones, from a real room characterized by a reverberation time of about 400ms. To acquire the impulse responses of the corresponding source-to-receiver acoustic paths, the technique reported in [Farina, 2000] was applied. A 4cm spacing between microphones was chosen, and to simulate a generic setup, the array was not placed orthogonal to the loudspeaker. The minimum *microphone to loudspeaker* distance was of 3 meters. The algorithm was applied to male and female speech files sampled at 22kHz and convolved with the resulting impulse responses. The following parameters and initializations were used for the algorithm. μ was set to 10^{-4}, the LP analysis order to 26 with an LP analysis frame length of $25ms$. The equalizers were $T = 1000$ taps long and initialized to $\mathbf{w}_i = [1, 0, 0, 0, 0, \ldots]$. This allows the filters to start the adaptation from a meaningful initial condition (the first tap set to one enables the signal to flow unaffected at the beginning of the adaptation). For speech input signals, where amplitude variations must be taken into account, the LPC coefficients must be divided by the standard deviation of the estimated residual. Therefore, the normalized coefficients and residual were used in the equation (80). The algorithm was left free to adapt also during unvoiced or silent periods as suggested in [Gillespie et al., 2001].

The performance of the de–reverberation algorithm reported in figure 12 have been evaluated

by the following formulation of the Direct to Reverberation Ratio (DRR):

$$DRR(dB) = 10 \log_{10} \frac{h^2(\delta)}{\sum_{k=0(k\neq\delta)}^{M-1} h^2(k)} \tag{86}$$

where $h(n)$ is the speaker-to-receiver impulse response, M its length in samples, and δ the time-index of the direct path in samples.

Figure 12(a) shows the echogram of the original impulse relating to the shortest source-to-receiver path. Figure 12(c) shows the equalized impulse response. A large amount of de-reverberation is already achieved by the delay and sum beamformer, figure 12(b), which however does not produce a consistent attenuation of the isolated early reflections. Similar results were obtained in several synthetic simulations where the impulse responses were calculated by the mirror image source method (MISM) [Allen and Berkley, 1979].

Similar performances were obtained by using different speech sources, as reported in table 1.

Table 1: DRR improvement in dB in respect to a DS beamformer.

Signal	DRR dB
Male 1 (Italian)	2.2
Male 2 (Portuguese)	2.9
Male 3 (English)	2.6
Male 4 (Italian)	3.1
Female 1 (Portuguese)	2.9
Female 2 (Russian)	2.3
Female 3 (Hebrew)	3.2
Female 4 (Dutch)	2.9

9 Conclusion

The dereverberation problem is still an open issue. In this article the main research directions have been described. Furthermore, it has been proposed a partial solution in an idealized framework where, in the absence of noise, the impulse responses of the source to receiver path are time-invariant. In this idealized case, the achieved dereverberation is relevant. The interest on the approaches based on the explicit channel inversion was led by the fact that, in theory, these methods can offer perfect dereverberation. In practice, even in the described idealized framework, this does not happen and the improvement, even if perceptually consistent, is limited. Why this discrepancy between theory and practice exists?

- A limiting factor of the performance of any dereverberation algorithm is how to decouple the speech and room contributions. This has been previously described as “the unambiguous deconvolution problem”. So far no perfect solution has been found in this sense, especially in the single channel case. Better approaches exist for multichannel systems. However, an high number of microphones spread in a large area of the room are necessary to obtain a convincing estimation of the speech LP coefficients. This is a severe constraint to practical applications. The unambiguous deconvolution problem is present for physical reasons, independently from the blind deconvolution algorithm that is used to perform the system inversion. This is a fundamental issue that is unlikely to be overcome.

- Another limiting factor is the performance of the blind deconvolution algorithm employed, both in terms of accuracy and speed of convergence, especially when the input signal fed to the algorithm is not perfectly white.
- Other issues come from all the systematic errors that might be present in the proposed dereverberation system, as for instance the NGA misconvergence described in[Douglas et al., 2005]. Of course these can be addressed and hopefully solved.

Departing from the idealized framework of a noiseless time-invariant acoustic systems, two problems are central and still need to be addressed:

- The speed of convergence of the algorithm must suffice to track the acoustic system variations.
- The sensitivity of the algorithm to noise might prevent the algorithm from working in real conditions.

Hopefully, both these issues are practically solvable.

References

R. Ahmad, A. W. H. Khong, M. K. Hasan, and P. A. Naylor. The extended normalized multichannel flms algorithm for blind channel identification. *in Proc. European Signal Processing Conf. (EUSIPCO)*, 2006.

R. Ahmad, N.D. Gaubitch, and P.A. Naylor. A noise-robust dual filter approach to multichannel blind system identification. *in Proc. European Signal Processing Conf. (EUSIPCO)*, 2007a.

R. Ahmad, A. W. H. Khong, and P. A. Naylor. A practical adaptive blind multichannel estimation algorithm with application to acoustic impulse responses. *Proc. IEEE Int. Conf. Digital Signal Processing*, 2007b.

J.B. Allen. Synthesis of pure speech from a reverberant signal. *U.S. Patent No. 3786188*, 1974.

J.B. Allen and D.A. Berkley. Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Am*, 65(4):943–948, 1979.

S. Amari, A. Cichocki, and H.H. Yang. A new learning algorithm for blind signal separation. *Advances in Neural Information Processing Systems 8. MIT Press.*, pages 752–763, 1996.

S. Amari, S. Douglas, A. Cichocki, and H. Yang. Novel on-line adaptive learning algorithms for blind deconvolution using the natural gradient approach, 1997.

C. Avendano, J. Benesty, and D. R. Morgan. A least squares component normalization approach to blind channel identification. *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 4:1797–1800, 1999.

A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

S. Bellini. *Bussgang techniques for blind deconvolution and equalizzation*. ed. S. Haykin, Prentice-Hall, Englewood Cliffs, NJ, 1994.

S. Bellini and F. Rocca. Blind deconvolution: polyspectra or bussgang techniques. *Digital Communications*, pages 251–263, 1986.

S. Bellini and F. Rocca. Near optimal blind deconvolution. *IEEE*, pages 2236–2239, 1988.

Benesty, Jacob, Makino, Shoji, Chen, and Jingdong Eds. *Speech Enhancement*. Springer, New Jersey, 2005.

Benesty, Jacob, Chen, Jingdong, Huang, and Yiteng. *Microphone Array Signal Processing*. Springer Topics in Signal Processing , Vol. 1, New Jersey, 2008.

S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. on ASSP*, 27(2)(3):113–120, 1979.

D. R. Brillinger. *Time series: Data analysis and theory*. Holt, Rinehart and Winston, New York, 1975.

J.r. C. R. Johnson. Admissibility in blind adaptive channel equalization. *IEEE Contr. Syst. Mag.*, pages 3–15, January 1991.

J. A. Cadzow. Blind deconvolution via cumulant extrema. *IEEE Signal Process.Mag.*, 13(3): 24–42, May 1996.

J.F. Cardoso and B. Laheld. Equivariant adaptive source separation. *IEEE Trans. Signal Processing. MIT Press.*, 43:3017–3030, 1996.

M. Delcroix, T. Hikichi, and M. Miyoshi. Dereverberation of speech signals based on linear prediction. *in Proc. of the 8th International Conference on Spoken Language Processing ICSLP04, Jeju Island, Korea*, 2:877–881, October 2004.

M. Delcroix, T. Hikichi, and M. Miyoshi. Blind dereverberation algorithm for speech signals based on multi-channel linear prediction. *Acoustical Science and Technology*, 26(5):432–439, October 2005.

D. L. Donoho. On minimum entropy deconvolution. *Applied Time Series Analysis, D. F. Findley, Ed. New York: Academic Press*, 1981.

S.C. Douglas, H. Sawada, and S. Makino. Natural gradient multichannel blind deconvolution and speech separation using causal FIR filters. *IEEE Transactions on Speech and Audio Processing*, 13(1):92–104, January 2005.

G. W. Elko. Microphone array systems for hands-free telecommunication. *Speech Commun.*, 20 (3-4):229–240, Dec. 1996.

K. Eneman and M. Moonen. Multimicrophone speech dereverberation: Experimental validation. *EURASIP Journal on Audio, Speech, and Music Processing*, 2007.

N. W. D. Evans, J. S. Mason, W. M. Liu, and B. Fauve. On the fundamental limitations of spectral subtraction: an assessment by automatic speech recognition. *in Proc. European Signal Processing Conf. (EUSIPCO)*, 2005.

A. Farina. Simultaneous measurements of impulse response and distortion with a swept-sine technique. *108th AES Convention*, 2000.

D.T. Fee, C.F.N Cowan, S. Bilbao, and I. Ozcelik. Predictive deconvolution and kurtosis maximization for speech dereverberation. *in Proc. European Signal Processing Conf. (EUSIPCO)*, 2006.

M. Ferras. Multi-microphone signal processing for automatic speech recognition in meeting rooms. Master's thesis, ICSI, Berkeley, California, 2005.

L.D. Fielder. Analysis of traditional and reverberation-reducing methods of room equalization. *J. Audio Eng. Society*, 51(1/2):3–26, January/February 2003.

K. Furuya. Noise reduction and dereverberation using correlation matrix based on the multiple-input/output inverse-filtering theorem (MINT). *Proc. of International Workshop on Handsfree Speech Communication*, 15(5):59–62, 2001.

K. Furuya and A. Kataoka. Robust speech dereverberation using multichannel blind deconvolution with spectral subtraction. *IEEE Transactions on audio, speech and languange processing*, 15 (5):1579–1591, 2007.

S. Gannot and M. Moonen. Subspace methods for multimicrophone speech dereverberation. *in Proceedings of the 7th IEEE/EURASIP International Workshop on Acoustic Echo and Noise Control (IWAENC 2001)*, 1:47–50, September 2001.

S. Gannot and M. Moonen. Subspace methods for multimicrophone speech dereverberation. *EURASIP Journal on Applied Signal Processing, vol. 2003*, 11:1074–1090, 2003.

N. D. Gaubitch. *Blind Identification of Acoustic Systems and Enhancement of Reverberant Speech*. PhD thesis, Imperial College, University of London, London, 2006.

N.D. Gaubitch, J. Benesty, and P.A. Naylor. Adaptive common root estimation and the common zeros problem in blind channel estimation. *in Proc. European Signal Processing Conf. (EUSIPCO)*, September 2005.

N.D. Gaubitch, Hasan M.K., and P.A. Naylor. Generalized optimal step-size for blind multichannel lms system identification. *Signal Processing Letters, IEEE*, 13(10):624–627, October 2006.

G.B. Giannakis and J.M. Mendel. Identification of nonminimum phase systems using higher order statistics. *IEEE Transaction on Acoustics, Speech and Signal Processing*, 37:360–377, March 1989.

B. Gillespie and L. Atlas. Strategies for improving audible quality and speech recognition accuracy of reverberant speech. *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003.

B. W. Gillespie, D. A. F. Florencio, and H. S. Malvar. Speech de-reverberation via maximum-kurtosis subband adaptive filtering. *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3701–3704, 2001.

Bradford W. Gillespie and Les E. Atlas. Acoustic diversity for improved speech recognition in reverberant environments. *in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 557–560, 2002.

G. H. Golub and C. F. Van Loan. *Matrix Computations (3rd ed.)*. Baltimore: Johns Hopkins University Press., 1996.

M. I. Gurelli and C.L. Nikias. Evam: an eigenvector-based algorithm for multichannel blind deconvolution of input colored signals. *in Proceedings of the 7th IEEE/EURASIP International Workshop on Acoustic Echo and Noise Control (IWAENC 2001)*, 43(1):134–149, Jannuary 2001.

E. A. P. Habets. Single-channel speech dereverberation based on spectral subtraction. *Proc. 15th Annual Workshop Circuits, Syst., Signal Process. (ProRISC04)*, 7(2):250–254, Nov. 2004.

E.A.P. Habets. *Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement*. PhD thesis, Technische Universiteit Eindhoven, Eindhoven, 2007.

M. K. Hasan, J. Benesty, P. A. Naylor, and D. B. Ward. Improving robustness of blind adaptive multichannel identification algorithms using constraints. *Proc. 13th European Signal Processing Conf.*, 2005.

D. Hatzinakos and C.L. Nikias. Blind equalisation based on higher order statistics (H.O.S.). *Haykin [1994]*, pages 181–258, 1994.

S. Haykin. *Blind Deconvolution*. Prentice-Hall, New Jersey, 1994.

S. Haykin. *Unsupervised adaptive filtering*. John Wiley & Sons, New Jersey, 2000.

S. Haykin. *Adaptive Filter Theory*. Prentice-Hall, New Jersey, 2002.

T. Hikichi, M. Delcroix, and M. Miyoshi. On robust inverse filter design for room transfer function fluctuations. *in Proc. European Signal Processing Conf. (EUSIPCO)*, 2006.

H.Kutruff. *Room Acoustics*. Taylor & Francis; 4th edition, New York, USA, 2000.

Y. Hua. Fast maximum likelihood for blind identification of multiple FIR channels. *IEEE Trans. Signal Processing*, 44:661–672, Mar. 1996.

Huang, Yiteng, Benesty, Jacob, Chen, and Jingdong. *Acoustic MIMO Signal Processing*. Springer Topics in Signal Processing, Vol. 1, New Jersey, 2006a.

Y. Huang, J. Benesty, and J. Chen. *Acoustic MIMO Signal Processing (Signals and Communication Technology)*. Springer-Verlag New York, Siracuse, NJ, USA, 2006b.

Y. A. Huang, J. Benesty, and J. Chen. A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment. *IEEE Transactions on speech and audio processing*, 13(5):882–895, September 2005.

M. Joho. *A Systematic Approach to Adaptive Algorithms for Multichannel System Identification, InverseModeling, and Blind Identification*. PhD thesis, Swiss Federal Institute of Technology, Zrich, 2000.

O. Kirkeby and P. A. Nelson. Digital filter design for inversion problems in sound reproduction. *J. Audio Eng. Soc.*, 47(7/8):583–595, 1999.

K. Lebart, J.M. Boucher, and P.N. Denbigh. A new method based on spectral subtraction for speech dereverberation. *Acta Acoustica*, 87(3):359–366, 2001.

H. Liu, G. Xu, and L. Tong. A deterministic approach to blind equalization. *Proc. 27th Asilomar Conf., Pacific Grove, CA*, pages 751–755, 1993.

R. Liu and G. Dong. A fundamental theorem for multiple-channel blind equalization. *IEEE Trans. on circuits and systems*, 44(5):472–473, may 1997.

L.Tong, G. Xu, and T. Kailath. A new approach to blind identification and equalization of multipath channels. *conference record of the 25th Asilomar Conference on signals systems and computers*, 2:856–860, november 1991.

J.M. Mendel. Tutorial on higher-order statistics (spectra) in signal processing and system theory: Theoretical results and some applications. *Proc. IEEE*, 79(3):278–305, March 1991.

S.K. Mitra. *Digital Signal Processing*. Mc Graw Hill, New Jersey, 2002.

M. Miyoshi and Y. Kaneda. Inverse filtering of room acoustics. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 36(2):145–152, 1988.

J. Mourjopoulos. On the variation and invertibility of room impulse response functions. *Journal of Sound and Vibration*, 102(2):217–228, sept 1985.

J. Mourjopoulos, P. M. Clarkson, and J .K. Hammond. A comparative study of least-squares and homomorphic techniques for the inversion of mixed-phase signals. *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 7:1858–1861, May 1982.

T. Nakatani, M. Miyoshi, and K. Kinoshita. Implementation and effects of single channel dereverberation based on the harmonic structure of speech. *Proc. of the International Workshop on Acoustic Echo and Noise Control (IWAENC03)*, pages 91–94, 2003.

P.A. Naylor and N.D. Gaubitch Eds. *Speech Dereverberation*. Springer, New Jersey, 2008.

P.A. Naylor and N.D. Gaubitch. Speech dereverberation. *Proc. of the International Workshop on Acoustic Echo and Noise Control (IWAENC 2005)*, 2005.

C. L. Nikias and M. R. Raghuveer. Bispectrum estimation: A digital signal processing framework. *Proc. IEEE*, 75(7):869–891, July 1987.

C.L. Nikias and J.M. Mendel. Signal processing with higher order spectra. *IEEE Signal Processing Magazine*, 10(3):10–37, July 1993.

A. V. Oppenheim and R. W. Schafer. *Discrete-Time Signal Processing*. Prentice-Hall, Englewood Cliffs, New Jersey, 1989.

F.S. Pacheco and R. Seara. Spectral subtraction for reverberation reduction applied to automatic speech recognition. *Telecommunications Symposium, 2006 International*, 7(2):795–800, Sept. 2006.

A. Papoulis. *Probability, Random Variables and Stochastic Processes, 2nd ed.* McGraw-Hill, Singapore, 1984.

J.D. Polack. *La transmission de l'energie sonore dans les salles*. PhD thesis, Universite' du Maine, La mans, 1988.

B. D. Radlovic, R.C. Williamson, and R. A. Kennedy. Equalization in an acoustic reverberant environment: Robustness results. *IEEE Trans. Speech and Audio Processing*, 8(3):311–319, 2000.

R. Ratnam, D.L. Jones, and Jr. W.D. O'Brien. Fast algorithms for blind estimation of reverberation time. *Signal Processing Letters, IEEE*, 11(6):537–540, June 2004.

O. Shalvi and E. Weinstein. New criteria for blind deconvolution of nonminimum phase systems (channels). *IEEE Trans. on information theory*, 36(2):312–321, march 1990.

D.T.M. Slock, A.K. Meraim, P. Duhamel, D. Lesbert, P. Loubaton, S. Mayrargue, and E. Moulines. Prediction error methods for time-domain blind identification of multichannel FIR filters. *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1968–1971, May 1995.

M. R. P. Thomas, N. D. Gaubitch, J. Gudnason, and P. A. Naylor. A practical multichannel dereverberation algorithm using multichannel dypsa and spatiotemporal averaging. *In Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-07)*, Oct. 2007.

M. Tonelli and M. E. Davies. A blind multichannel dereverberation algorithm based on the natural gradient. *IWAENC*, 2010.

L. Tong and S. Perreau. Multichannel blind identification: From subspace to maximum likelihood methods. *Proc. IEEE*, 86(10):1951–1968, November 1998.

M. Triki and T.M.D. Slock. Iterated delay and predict equalization for blind speech dereverberation. *IWAENC 2006, Paris*, September 2006.

M. Triki and T.M.D. Slock. AR source modeling based on spatiotemporally diverse multichannel outputs and application to multimicrophone dereverberation. *DSP 2007, 15th International Conference on Digital Signal Processing*, July 2007.

M. Unoki, M. Toi, and M. Akagi. Refinement of an MTF-based speech dereverberation method using an optimal inverse-MTF filter. *Proc. SPECOM*, 7, jun 2006.

B.D. Van Veen and K.M. Buckley. Speech de-reverberation via maximum-kurtosis subband adaptive filtering. *IEE ASSP Magazine*, pages 4–24, 1988.

N. Virag. Single channel speech enhancement based on masking properties of the human auditory system. *IEEE Trans. Speech Audio Processing*, 7(2):126–137, Mar. 1999.

D.B. Ward, R.A. Kennedy, and R.C. Williamson. Theory and design of broadband sensor arrays with frequency invariant far-field beam patterns. *J. Acoust. Soc. Amer.*, 97(2):1023–1034, Feb. 1995.

J.Y.C. Wen, E.A.P. Habets, and P. A. Naylor. Blind estimation of reverberation time based on the distribution of signal decay rates. *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 329–332, 2008.

R. A. Wiggins. Minimum entropy deconvolution. *Geoexploration*, 16:21–35, 1978.

Mingyang Wu and DeLiang Wang. A two-stage algorithm for enhancement of reverberant speech. *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1:1085–1088, March 2005.

G. Xu, H. Liu, L. Tong, and T. Kailath. A least-squares approach to blind channel identification. *IEEE Trans. Signal Processing*, SP43(12):2982–2993, December 1995.

B. Yegnanarayana and P.S. Murthy. Enhancement of reverberant speech using lp residual signal. *IEEE Trans. Speech Audio Processing*, 8(3):267–281, 2000.

T. Yoshioka, T. Hikichi, and M. Miyoshi. Second-order statistics based dereverberation by using nonstationarity of speech. *IWAENC 2006, Paris*, September 2006a.

T. Yoshioka, T. Hikichi, M. Miyoshi, and H.G. Okuno. Robust decomposition of inverse filter of channel and prediction error filter of speech signal for dereverberation. *in Proc. European Signal Processing Conf. (EUSIPCO)*, 2006b.

Y. Zhang, J. A. Chambers, F.F. Li, P. Kendrick, and T.J. Cox. Blind estimation of reverberation time in occupied rooms. *in Proc. European Signal Processing Conf. (EUSIPCO)*, 2006.