A MULTI-CHANNEL MAXIMUM LIKELIHOOD APPROACH TO DE-REVERBERATION

Massimiliano Tonelli, Maria G. Jafari and Mike Davies

Centre for Digital Music
Queen Mary, University of London
Mile End Road, E1 4NS London, UK

ABSTRACT

Reverberation can severely degrade the intelligibility of speech. Blind de-reverberation aims at restoring the original signal by attenuating the reverberation without prior knowledge of the surrounding acoustic environment nor of the source. In this paper, single-channel and multi-channel de-reverberation structures are compared and the advantages of the multi-channel approach are discussed. We propose an adaptive multi-channel blind de-reverberation algorithm based on a maximum likelihood approach that exploits results relating to the multiple input/output inverse theorem (MINT). The performance of the algorithm is illustrated using an eight-channel linear microphone array placed in a real room. Simulation results show that the algorithm can achieve very good de-reverberation when the channels are time aligned.

1. INTRODUCTION

Reverberation, a component of any sound generated in a natural environment, can degrade speech intelligibility or more generally the acoustic quality. 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 [1]. More ambitious applications have addressed the improvement of the acoustics of theaters or even the creation of virtual acoustic environments [2]. In all these cases de-reverberation is critical.

If the system input is unknown, only blind techniques can be applied and the equalization becomes more complex. Different approaches have been proposed for blind de-reverberation, the core idea being the estimation of one or more inverse filters that are used as system equalizers. Although the evaluation of the inverse filter can be problematic for a realistic acoustic environment, multi-channel structures offer a viable solution. Intuitively, since they provide several measurements and thus a greater amount of information, they potentially lead to better de-reverberation performances.

In this paper we propose a modification based on a maximum likelihood (ML) approach of the multi-channel structure discussed in [3]. The problems connected to single-channel system inversion and the advantages of a multi-channel approach are explained in section 2. The benefits of the ML approach over the one based on kurtosis are discussed in section 3. The modification of the multi-channel de-reverberation system by the ML learning rule is reported in 3.1 and in section 4 is described its application to the de-reverberation of speech acquired by a linear microphone array.

1.1 Single channel de-reverberation

If the acoustic path is modeled as a linear-time invariant system characterized by an impulse response $h(n)$, the source signal, $s(n)$, and the reverberant signal, $x(n)$, are linked by the equation

$$x(n) = h(n) * s(n) \tag{1}$$

where $*$ denotes the discrete linear convolution.
De-reverberation is achieved by finding a filter with impulse response $g(n)$ so that

$$\delta(n - N_d) = g(n) * h(n) \tag{2}$$

where $g(n)$ is defined as the inverse filter of $h(n)$, $\delta(k)$ is the unit sample sequence and N_d a delay [4].

If the impulse response $h(n)$ is unknown and there is no information about the original source $s(n)$, de-reverberation is blind. The aim of blind de-reverberation is to estimate $s(n)$ by removing the reverberation components from the measured signal $x(n)$, without knowledge of the surrounding acoustic environment. Therefore, single channel blind de-reverberation is connected to the blind estimation of $g(n)$ so that

$$\hat{s}(n) = y(n) = g(n) * x(n) \tag{3}$$

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

Several approaches exist to address this problem, including a large class of blind identification methods that are based on higher order statistics [5].

1.2 Multi-channel de-reverberation, the MINT and the Bezout identity

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

$$x_i(n) = h_i(n) * s(n),\ 1 \leq i \leq N \tag{4}$$

$$\hat{s}(n) = y(n) = \sum_{i=1}^{N} g_i(n) * x_i(n) \tag{5}$$

where $x_i(n)$, $h_i(n), g_i(n)$ 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 $g_i(n)$ so that

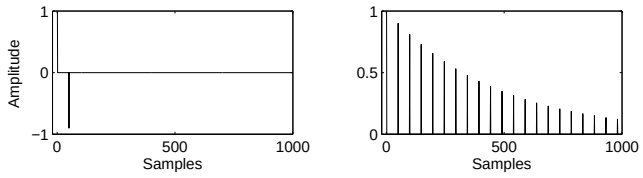

Figure 1: *Illustration of the inversion (a) Single echo impulse response; (b) Truncated inverse filter. A strong reflection requires a very long inverse filter.*

$$\delta(n-Nd) = \sum_{i=1}^{N} g_i(n) * h_i(n) \tag{6}$$

this expression is know as the MINT theorem [6] and it is closely related to the Bezout identity [7]

The Bezout identity states that there exists polynomials $G_i(z)$ such that equation

$$1 = \sum_{i=1}^{N} G_i(z) H_i(z) \tag{7}$$

holds if the polynomials $H_i(z)$ have no common zeros. The algebraic decomposition that satisfies the Bezout identity is in general not unique and the algorithm reported in [6] calculates one of the possible solutions for the equalizers $g_i(n)$.

2. SYSTEM INVERSION

2.1 Single-channel system inversion

When the speaker-to-receiver impulse response $h(n)$ is non-minimum phase (i.e. it has zeros outside the unit circle), the calculation of its inverse filter is problematic. In fact, the inverse of a non-minimum phase FIR system is an unstable IIR filter.

The acoustic signal-transmission channel is generally a non minimum phase function [8]. A possible solution is to consider a truncated FIR approximation of the inverse IIR filter, that is by definition always stable. Nevertheless, a very long filter might be necessary to attain a good de-reverberation. As an example, the inversion of a single echo impulse response is shown in figure 1.

2.2 Multi-channel system inversion

The MINT theorem offers a solution to the instability issue associated with the inversion of non-minimum phase transfer functions [6], by ensuring that the equalizers will be FIR filters if the channel transfer functions are FIR. Since reverberation is essentially due to energy that is decaying, every room transfer function can be approximated, down to a desired noise floor, by an FIR filter. Therefore, the inverse impulse responses are characterized by shorter lengths than in the single-channel case.

As was pointed out in [9] the MINT theorem also implies that the the number of taps needed for an N-channel system, considering an M-tap long impulse response, T is given by

$$T = \text{ceil}((M+1/N-1)-1) \tag{8}$$

therefore, by increasing the number of microphones, the filter length is reduced. This yields better statistical properties in the equalizer estimation, less computational demand and less memory requirement. Thus multi-channel based structures can potentially exploit these properties to provide better and more efficient de-reverberation.

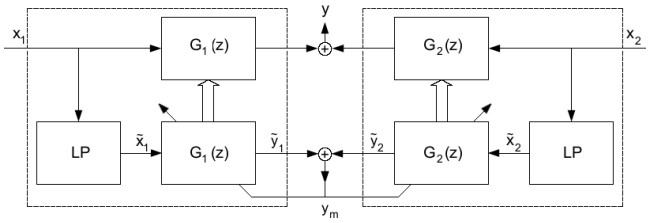

Figure 2: *Two-channel system composed by two single-channel structures*

3. A MAXIMUM LIKELIHOOD APPROACH TO DE-REVERBERATION

The observation that the kurtosis of the linear prediction (LP) residual can serve as a reverberation metric resulted in an algorithm based on higher order statistics introduced in [10]. Low kurtosis values of the residual imply an highly reverberated speech signal, thus enabling the inverse filter to be identified by kurtosis maximization. However, the calculation of the kurtosis and its derivative are prone to instability [3, 11]. In order to reduce this sensitivity, a single-channel blind de-reverberation algorithm that uses a ML approach to estimate the inverse filter $g(n)$ has been proposed [3].

In this paper, $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$ are respectively the residual vector of the LP analysis of x and the output vector from the kurtosis maximization filter. The method is based on the idea of building an FIR filter $g(n)$ so that the output $\tilde{y}(n)$

$$\tilde{y}(n) = g(n) * \tilde{x}(n) \tag{9}$$

has any desired probability density function. The probability density function of $\tilde{y}$ is chosen to have high kurtosis and bounded derivative. A popular probability density function with these properties is

$$P(\tilde{\mathbf{y}}) = \frac{1}{\cosh(\tilde{\mathbf{y}})} \tag{10}$$

this leads to an adaptive algorithm based on the update equation [3].

$$g(k+1) = g(k) - \mu E\{\tanh(\tilde{\mathbf{y}})\tilde{\mathbf{x}}\} \tag{11}$$

where μ is the adaptation step of the adaptive algorithm. In [3] it is shown how this solution can provide a better de-reverberation result for a single-channel structure compared to the kurtosis based method.

The stochastic gradient version of equation (11) is of course, a Bussgang-type equalizer.

3.1 The multi-channel ML de-reverberation algorithm

In blind de-reverberation, the transfer functions $h_i(n)$ are unknown. This implies that the equalizers $g_i(n)$ must be blindly estimated. To achieve this goal, we propose a time domain multi-channel structure similar to the two-channel system shown in figure 2. This is inspired by the method proposed in [10] but exploits the ML filter update rule (11) for each channel. In the N-th dimensional case (11) becomes

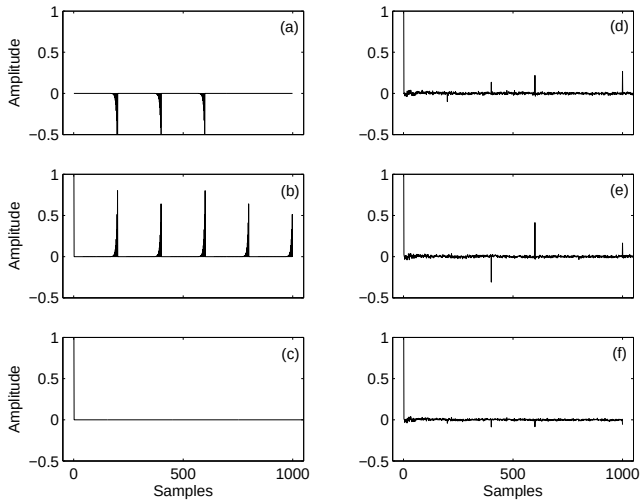

Figure 3: *Comparison of the MINT and multi-channel equalizers. (a), (b) Inverse filters calculated by MINT. (d), (e) Inverse filters calculated by the multi-channel algorithm. (c), (f) Equation (6) evaluated in both cases.*

$$g_i(n+1) = g_i(n) - \mu E\{\tanh(\tilde{\mathbf{y}}_m)\tilde{\mathbf{x}}_i\} \tag{12}$$

where $\tilde{\mathbf{x}}_i$ is the output vector of the *i*-th LP analysis filter and $\tilde{\mathbf{y}}_m$ is defined as

$$\tilde{\mathbf{y}}_m = \frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{y}}_i \tag{13}$$

where $\tilde{\mathbf{y}}_i$ is the output vector of the *i*-th maximization filter. The filters are jointly optimized to maximize the output $\tilde{\mathbf{y}}_m$ and the de-reverberator output y is defined in (5).

Extending our results from the single-channel ML technique, we expect the multi-channel structure to benefit by improved stability and less noise in the convergence compared to the kurtosis approach. Conversely, the use of the LP residual to decouple the harmonic structure of speech and reverb introduces ambiguity, since both the LP and the de-reverb FIR filter are convolutional operators. This problem has been mitigated as proposed in [12].

3.2 A comparison between ML de-reverberation and the MINT

To highlight the affinity of the algorithm above with the MINT a two-channel system has been used to equalize a speech signal sampled at 22050Hz that has been processed with the following two long echoes:

$$\begin{cases} h_1(n) = \delta(0) - 0.8\delta(n-600) \\ h_2(n) = \delta(0) - 0.8\delta(n-1000) \end{cases} \tag{14}$$

The inverse filters have been calculated by the MINT inverse formula given in [6] with the explicit use of the impulse responses of the system, and this result has been compared to the filters that have been blindly identified by the multi-channel structure. This second approach directly estimates the inverse filters,which is statistically better than estimating the impulse responses of the system and then inverting. Note that the lengths of the inverse filters are comparable to the length of the longest delay present. This is in contrast to the single-channel case, which would require a much longer filter. The results are shown in figure 3 where the three leftmost plots relate to the MINT method while the rightmost three plots show the algorithm performance. The inverse filters have similar placement of the taps but different gains; both inverse filters provide an equalization for the system (figures 3*c*, 3*f*); the multi-channel structure does not converge to the solution calculated by the MINT but to a similar noisier one. For longer tap filters this solution is non-unique since it only needs to satisfy the Bezout identity.

4. EXPERIMENTAL RESULTS

An eight-channel system that uses a one-point sample mean version of the adaptation rule of equation (11) was evaluated. A linear array composed of eight microphones was used in a room to measure the impulse responses of the corresponding source-to-receiver acoustic paths. To acquire these impulse responses, the technique reported in [13] was applied. A 4cm spacing between microphones was chosen for the first experiment and 45cm for the second, and to simulate a generic setup, the array was not placed orthogonal to the loudspeaker. The minimum microphone-to-loudspeaker distance was of 2.5 meters. The impulse responses measured for the 4cm array configuration did not exhibit significant delay misalignment among the channels, while the impulse responses for the 45cm array were not aligned. The algorithm was applied to male and female speech files sampled at 22050hz and convolved with the resulting impulse responses.

The following parameters and initializations were used for the algorithm. μ was set to $5\cdot 10^{-5}$, the LP analysis order to 26 with an LP analysis frame length of $25ms$. The equalizers were $T = 1000$ taps long and initialized to $\mathbf{g}_i = [1,0,0,0,0,...]$. To obtain a more uniform convergence, the residual of each channel was normalized to a zero mean, unit variance process. The algorithm was left free to adapt also during unvoiced or silent periods as suggested in [10].

To solve the problem of gain uncertainty, a normalization of the filter coefficients was performed at every update cycle [10]. It should be noted that normalizing all the channels makes the problem over-constrained if we wish to take advantage of the Bezout inverse solution. Identifying the best form of constraint should therefore be the subject of future work.

The algorithm was found to provide de-reverberation in the case of the 4cm spacing, but not in the 45cm configuration, due to the time-misalignment among the channels. After investigation, it was understood that a meaningful convergence cannot be achieved for this algorithm when the channels are time-misaligned. Therefore the algorithm cannot identify the inter-channel delay. Note that this problem equally applies to the use of kurtosis within this framework [10]. Conversely, when the impulse responses were aligned manually, the algorithm converged for the 45 centimeter setup, and provided de-reverberation, although its operation was no longer blind. Figure 4(a) shows the echogram of the original impulse relating to the shortest source-to-receiver path. Figure 4(c) shows the equalized impulse response obtained with equation (11), in the case of a one-point sample mean.

The process of alignment is the same that is required for a delay and sum beamformer. In this sense it is worthy to observe that the algorithms proposed here and in [10] require a preprocessing stage. A large amount of de-reverberation is already achieved by the delay and sum beamformer, which however does not produce a consistent attenuation of the isolated early reflections. Figure 4(b) shows the echogram of a delay and sum beamformer using the same delays used to align the channel in the proposed method. Similar results were obtained in several synthetic simulations where the impulse responses were calculated by the image source method [2].

The performance of the de-reverberation algorithm reported in figure 4 have been evaluated by the Signal-to-Reverberation Ratio (SRR)

$$SRR(dB) = 10\log_{10}\frac{h^2(\delta)}{\sum_{k=0(k\neq\delta)}^{M-1} h^2(k)} \qquad (15)$$

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 [1].

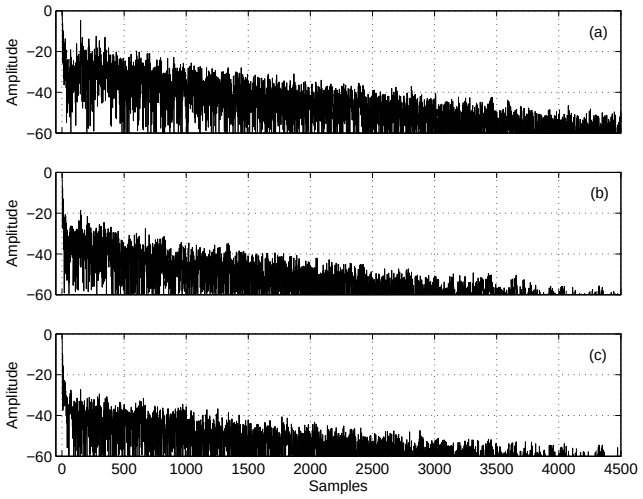

Figure 4: *(a) Reference echogram relating to the shortest source-to-receiver path, SRR = -6.3dB, (b) 8-channel delay-sum beamfomer, SRR =-0.1dB, (c) 8-channel ML dereverberator, SRR =2.1dB*

5. CONCLUSIONS

In this paper a novel modification of a multi-channel structure based on maximum likelihood has been proposed, which was used to de-reverberate signals recorded in a real room. The proposed algorithm allows the use of shorter filter lengths than single channel systems by exploiting the MINT. Our simulation results have shown that good dereverberation is achieved even in real room, although a pre-processing might be necessary, particularly for widely spaced microphones. Both the kurtosis and ML based algorithm suffer from this drawback.

We are currently investigating different schemes for the initialization and/or in the normalization to provide a blind estimation of the delay among channels, or otherwise, to use a different blind deconvolution algorithm that can address this problem.

6. ACKNOWLEDGMENTS

The authors wish to thank Nicolas Chétry for the interesting discussions and suggestions.

REFERENCES

[1] M. Ferrás, "Multi-microphone signal processing for automatic speech recognition in meeting rooms," M.S. thesis, ICSI, Berkeley, California, 2005.

[2] W. G. Gardner, "The virtual acoustic room," M.S. thesis, MIT Media Lab, Cambridge, 1992.

[3] M. Tonelli, N. Mitianoudis, and M. E. Davies, "A maximum likelihood approach to blind audio dereverberation," *Proc. Digital Audio Effects Conference (DAFx'04)*, pp. 256–261, 2004.

[4] B. D. Radlovic and R. A. Kennedy, "Iterative cepstrum-based approach for speech de-reverberation," *Proc. of ISSPAA*, vol. 1, pp. 55–58, 1999.

[5] C. L. Nikias and J. M. Mendel, "Signal processing with higher order spectra," *IEEE Signal Processing Magazine*, vol. 10, no. 3, pp. 1037, 1993.

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

[7] 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*, vol. 13, no. 5, pp. 882–895, September 2005.

[8] S.T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Amer.*, vol. 66, pp. 165–169, 1979.

[9] T.S. Bakir and R.M. Mersereau, "Blind adaptive dereverberation of speech signals using a microphone array," 2003.

[10] B. W. Gillespie, D. A. F. Florencio, and H. S. Malvar, "Speech de-reverberation via maximum-kurtosis subband adaptive filtering," *Proc. of ICASSP*, pp. 3701–3704, 2001.

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

[12] N. D. Gaubitch, P. A. Naylor, and D. B. Ward, "On the use of linear prediction for dereverberation of speech," *IWAENC*, pp. 99–102, 2003.

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