ITERATED DELAY AND PREDICT EQUALIZATION FOR BLIND SPEECH DEREVERBERATION

Mahdi Triki, Dirk T.M. Slock *

Eurecom Institute
2229 route des Crêtes, B.P. 193, 06904 Sophia Antipolis Cedex, FRANCE
Email: {triki,slock}@eurecom.fr

ABSTRACT

In this paper, we consider the blind multichannel dereverberation problem for a single source. The multichannel reverberation impulse response is assumed to be stationary enough to allow estimation of the correlations it induces from the received signals. It is well-known that a single-input multi-output (SIMO) filter can be equalized blindly by applying multichannel linear prediction (LP) to its output when the input is white. When the input is colored, the multichannel linear prediction will both equalize the reverberation filter and whiten the source. We exploit the channel spatial diversity to estimate the source correlation structure, which can hence be used to determine a source whitening filter. Multichannel linear prediction is then applied to the sensor signals filtered by the source whitening filter, to obtain source dereverberation. We exploit the input signal time diversity to reduce the equalization noise. Due to the speech signal non-stationarity, averaging equalizers (which are computed on different frames) increases the dereverberation accuracy. Simulation results reveal that an iterated equalization scheme (based on frame-by-frame analysis) increases the dereverberation performance, and leads to better auditive results.

1. INTRODUCTION

The quality of speech captured in real-world environments is invariably degraded by acoustic interference. This interference can be broadly classified into two distinct categories: additive and convolutive. The convolutive interference (commonly referred to as reverberation) is due to sound wave reflections from surrounding walls and objects. It leads to a modification of the speech signal characteristics. Therefore, it constitutes a major problem in speech recognition, speaker verification, and general auditive comfort in "hands-free" telephony applications. Blind dereverberation is the process of removing the effect of reverberation from an observed reverberant signal. Reducing the distortion caused by reverberation is a difficult blind deconvolution problem, due to the broadband nature of speech and the length of the equivalent impulse response from the speaker's mouth to the microphone.

A simple multi-microphone speech dereverberation system is the delay-and-sum beamfomer [1, 2]. The dereverberation is performed by a simple averaging over the sensor outputs, delayed to focus in the direction of the desired speaker. Note that beamforming exploits only a partial spatial information (relative delays), and ignores the input signal characteristics.

Eurecom Institute's research is partially supported by its industrial members: BMW, Bouygues Télécom, Cisco Systems, France Télécom , Hitachi Europe, SFR, Sharp, STMicroelectronics, Swisscom, Thales

A first class of speech dereverberation techniques suggests exploiting the statistical and spectral models of the speech signal to improve the enhancement accuracy. In [3], the authors seek to find a blind deconvolution filter that makes the LP residual as non-Gaussian as possible (using a kurtosis-based metric). In this way, they exploit the a priori knowledge that the signal to be recovered (speech) is super-Gaussian. They show that the proposed technique achieves significant improvement in performance over the delay-and-sum beamformer. A generic approach is proposed in [4] exploiting simultaneously the non-gaussianity, non-whiteness, and non-stationarity of the speech signal. On the other hand, source production-based techniques are also proposed for blind dereverberation. The source model describes speech signal in terms of an excitation sequence exiting a time-varying all-pole filter. Dereverberation is achieved by attenuating the peaks in the excitation sequence (due to room reverberation), then synthesizing the enhanced speech using the enhanced LP residual on the all-pole filter (estimated from the reverberant speech). It is clear that an important assumption is made; that the LP coefficients are unaffected by reverberation. In [5], the authors show that spatial averaging of the LP coefficients (estimated on each microphone) is required to improve the accuracy of this type of algorithms. They also demonstrate in [6] that LP coefficients obtained from spatially averaged multichannel speech signals achieve equally satisfactory results.

Another way to address the problem is to consider the whole Acoustic Impulse Response (AIR). Matched Filter (MF) is proposed to equalize the room response [7]. In such a way, one increases the dereverberation SNR (compared to the Delay-and-Sum beamformer). However, MF equalization introduces a large equalization delay (of about the AIR length), and produces a pre-echo that is annoying in several applications (speech recognition...). On the other hand, SIMO channel can be perfectly equalized using multiple FIR filters (transverse filters) [8]. Let us consider a clean speech signal, $s(n)$, produced in a reverberant room. The reverberant speech signal observed on M distinct microphones can be written as:

$$\mathbf{y}(k) = \mathbf{H}(q)s(k) \tag{1}$$

where $\mathbf{y}(k) = [y_1(k) \cdots y_M(k)]^T$ is the reverberant speech signal, $\mathbf{H}(q) = [H_1(q) \cdots H_M(q)]^T = \sum_{i=0}^{L_h-1} \mathbf{h_i} q^{-i}$ is the SIMO channel transfer function, L_h is the channel length, and q^{-1} is the one sample time delay operator. According to the Bézout's identity, if the channels $H_1(q) \cdots H_M(q)$ does not have common zeros, then $\exists \mathbf{F}(q) = [F_1(q) \cdots F_M(q)]$ such that:

$$\mathbf{F}(q)\mathbf{H}(q) = \sum_{m=1}^{M} F_m(q)H_m(q) = 1 \tag{2}$$

If $\mathbf{H}(q)$ is known (or can be estimated), the coefficients of $F_m(q)$ can be computed by the well-known rules of matrix algebra. The AIR blind estimation should face the channel/speech identifiability problem. In fact, for any scalar filter $\alpha(q)$, $(H(q)/\alpha(q)$, $\alpha(q)s(k))$ is also an acceptable solution of (1).
In [9], Huang et al. focus on the single-source two-microphone system. They notice that the AIR can be estimated by minimizing the mean squared value of

$$e(k) = \widehat{H}_2(q)y_1(k) + \widehat{H}_1(q)y_2(k) \tag{3}$$

They show that the solution is obtained through the eigenvalue decomposition of the autocorrelation matrix of the observed signal. Generalization to an arbitrary channel number is presented in [10]. If the channel length is known, the AIR is well estimated. However for acoustic channels, the true impulse response length is generally unknown, or/and not defined. Therefore, it is frequently overestimated (let us denote by $\widehat{L}_h$ the overestimated length). In such a case, for any scalar filter $\alpha(q)/deg(\alpha(q)) < (\widehat{L}_h - L_h)$, $\alpha(q)\mathbf{H}(q)$ is also a solution of (3)

$$\begin{aligned} e(k) &= \alpha(q)H_2(q)y_1(k) + \alpha(q)H_1(q)y_2(k) \\ &= \alpha(q)\left(H_2(q)y_1(k) + H_1(q)y_2(k)\right) = 0 \end{aligned}$$

Hikichi et al. propose solve the identification ambiguities by post-processing the estimated channel, in order to estimate and compensate the common factor $\alpha(q)$ [11]. The common factor is extracted as a characteristic polynomial of the two-channel linear prediction matrix.
Another way to deal with identification ambiguities by exploiting a priori information on source spectrum. If the source is white, the channel can be equalized using multi-channels linear prediction [12]. For speech input, we take advantage of the spatial diversity to estimate the speech correlation; and we propose a tree-stage dereverberation procedure exploiting spatial, temporal, and spectral diversities [13, 14].
This paper is organized as follows. In section 2, the tree-stages speech dereverberation procedure will be reviewed. Next, multi-frame speech dereverberation will be investigated in section 3.

2. SPEECH DEREVERBERATION PROCEDURE

We have proposed in [14] a processing scheme that works with three cascades of stages:

- First, the colored non-stationary speech signal is transformed into an iid-like signal (by taking advantage of the spatial and temporal diversities).
- Then, a blind channel predictor is computed based on pre-processed reverberant speech.
- Finally, speech signal dereverberation is performed using a zero-forcing equalizer based on the predictor computed in the previous step.

2.1. The source whitening stage

From the Statistical Room Acoustics (SRA) theory, one can show that for frequencies $f > f_{sch} = 2000\sqrt{T_{60}/V}$, the average reverberation spectrum is flat [15], i.e.,

$$E\left\{\left|H\left(\exp^{2j\pi f}\right)\right|^2\right\} = \frac{1-\beta}{\pi A\beta} \tag{4}$$

where $E\{\}$ is the spatial expectation, β is the average wall absorption coefficients, A is the total wall surface area, f_{sch} is the "Schroeder frequency", T_{60} is the reverberation time and V is the room volume.
In [13], simulations shows that the superposition of the SIMO sub-channels spectrums tends to be flat as the number of microphones increase. Then, the superposition of the spectra of the received signals can estimate (up to a multiplicative factor) the source spectrum. As this common part is due to the anechoic speech signal, it can be modeled as an AR process. The common AR coefficients can be estimated as those that minimize the sum of the squares of the prediction errors, averaged over the microphones (which leads to the normal equations):

$$e = \sum_{k=1}^{M}\sum_{n=0}^{\infty} e_k^2(n) = \sum_{k=1}^{M}\sum_{n=0}^{\infty}\left[y_k(n) - \sum_{j=1}^{l} a_j y_k(n-j)\right]^2 \tag{5}$$

Once the source spetrum is estimated, the source whitened reverberant signal is computed as:

$$\mathbf{x}(k) = a(q)\mathbf{y}(k) \approx \mathbf{H}(q)\tilde{s}(k) \tag{6}$$

where $\mathbf{x}(k) = [x_1(k)\cdots x_M(k)]^T$, $a(q) = 1 + \sum_{j=1}^{l} a_j q^{-j}$ is the linear prediction error filter of the source signal (performed in the previous stage), $\tilde{s}(k)$ is the source prediction error.
A periodic input signal (which is perfectly predictible) may lead to identifiability problem for the SIMO channel: the predictor will have tendency to kill the signal rather than to whiten it. To alleviate this problem, we propose taking advantage from the signal non-stationarity (that can be interpreted as a form of temporal diversity). We suggest considering the totality of the speech signal in order to calculate the AR coefficients (which estimates the averaged speech spectrum). It is important to emphasize that non-stationarity of the source is irrelevant as long as the source correlations are estimated with the same temporal averaging as for the multichannel linear prediction. The temporal diversity becomes a byproduct of this requirement.

2.2. The multichannel prediction stage

In the previous section, we have shown that the channel spatial diversity and the speech non-stationarity can be exploited to estimate the source correlation structure, which can hence be used to compute a source whitening filter.
Consider now the problem of predicting $\mathbf{x}(k)$ from the L_p latest observations $\mathbf{X_{L_p}}(k-1) = [\mathbf{x}^T(k-1)\cdots\mathbf{x}^T(k-L_p)]^T$. The prediction error is given by:

$$\tilde{\mathbf{x}}(k) = \mathbf{x}(k) + \sum_{i=1}^{L_p} A_{L_p,i}\mathbf{x}(k-i) = A_{L_p}\mathbf{X_{L_p+1}}(k) \tag{7}$$

where $A_{L_p} = \left[I_m\, A_{L_p,1}\ \cdots\ A_{L_p,L_p}\right]$, $A_{L_p,i}$ are the linear prediction filter coefficient matrices that should be determined to minimize the mean squared value of $\tilde{\mathbf{x}}(k)$, L_p denotes the prediction order. Minimizing the energy of the prediction error leads to the system of equations (for large enough L [17]):

$$S_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}(z) = A_{L_p}(z)S_{\mathbf{xx}}(z)A_{L_p}^{\dagger}(z) = \mathbf{h_0}S_{\tilde{\mathbf{s}}\tilde{\mathbf{s}}}(z)\mathbf{h_0}^H \tag{8}$$

where - $S_{\tilde{x}\tilde{x}}(z)$, $S_{xx}(z)$, and $S_{\tilde{s}\tilde{s}}(z)$ denote respectively the spectrum of the reverberant signal prediction error, reverberant signal, and source prediction error signals.

- $A(z) = \sum_{i=0}^{L_p} A_{L_p,i} z^{-i}$ denotes the prediction error filter, computed by solving the well-known normal equations. $A^{\dagger}(z)$ is the matched filter associated to $A(z)$.

- $\mathbf{h_0} = \mathbf{H}(+\infty)$ represents the first vector coefficient of the SIMO channel filter, which can be estimated (up to a scalar) as the eigenvector corresponding to the maximum eigenvalue of the LP residual correlation matrix $r_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}(0)$.

A relevant issue with the linear prediction approach is the alignment of the received signals on the various microphones(delay compensation for direct path). This leads to an increase in the prediction performance, and allows the use of shorter predictor. Contrarily to the Delay-&-Sum beamformer, we would align first paths (not the most powerful paths). So that, correlation based techniques do not give always good results. In [14], we have proposed an iterative approach based on the analysis of the covariance matrix of the multi-channel prediction error. And, we have showed that the proposed scheme gives satisfactory results.

2.3. The dereverberation stage

Based on the predictor performed in the previous stage, the spatiotemporal zero-forcing equalizer (called Delay-and-Predict equalizer) can be computed as:

$$\mathbf{F_{D\&P}}(q) = \mathbf{h}_0^H A_{L_p}(q) D(q) \tag{9}$$

where $D(q)$ is a diagonal matrix of delays aligning the direct path contributions in the M reverberant signal.
Thus, the dereverberated speech signal can be computed as:

$$\widehat{s}(k) = \mathbf{F_{D\&P}}(q)\mathbf{y}(k) = \mathbf{h_0}^H A_{L_p}(q)\mathbf{y}(k) \tag{10}$$

Note that the delays in $D(q)$ are the same as in the delay-and-sum beamformer, in which $\mathbf{h_0}^H A_{L_p}(q)$ gets replaced by $[1 \cdots 1]$

3. MULTI-FRAME SPEECH DEREVERBERATION

The residual error in the delay-&-predict equalization can be broadly classified into two distinct categories: estimation and modelling errors. The estimation error is due to the use of the sampling covariance matrices. Whereas, the modelling error is due to the assumptions on the channel and the input signal structures. In fact, we suppose that we have enough spatial diversity such as the multichannel response becomes an all-pass filter; and that the averaged speech signal is an AR process with a given order. Of course, in practice this will never be the case; and some of the input signal correlations will remain on the output of the "source whitening stage". In general, the more data we have, the best we estimate the channel correlation; and the lower the estimation error will be. However, it has a very little influence on the error modelling.

On the other hand, the temporal diversity of the input signal (the speech signal non-stationarity) is used just to avoid singularities due to the prediction of the voiced frames. Therefore, a frame-by-frame based technique can be proposed to solve the problem. The received signal is first segmented into P frames (with or without overlapping). The frame length should be sufficient to have a good estimation of the channel correlations. At each frame, a channel equaliser is computed using the delay-&-predict technique. The equalized signal computed on the basis of p^{th} frame correlations is

$$z^{(p)}(k) = \widetilde{h}^{(p)} * s(k) \qquad p = 1 : P \tag{11}$$

where $\widetilde{h}^{(p)} = f^{(p)} * h$ is the equalized channel. In such a way, the M-SIMO channel equalization is transformed on P-SIMO channel equalization; with the advantage that $\widetilde{h}^{(p)}$ is less reverberant then the acoustic impulse responses.
The reverberation in $\widetilde{h}^{(p)}$ is mainly function of remaining source correlations in the whitened signal. Due to the speech non-stationarity, the average speech spectrum (next, the reverberation in $\widetilde{h}^{(p)}$) at each frame are not correlated. Thus, this reverberation can be reduced by averaging the computed equalizers (corresponding to delay-&-sum Beamforming on $z^{(p)}(k)$).

The most serious drawbacks of such approach are related to the process of segmentation of the received signal $y(k)$: a finite length segment $y^{(p)}(k)$ can only approximately be represented by the convolution of $h(k)$ with some clean-speech segment $s^{(p)}(k)$ [19]. In fact, each segment of the reverberant speech can be written as:

$$y^{(p)}(k) = s^{(p)}(k) * h(k) + v^{(p)}(k) - u^{(p)}(k) \tag{12}$$

where $v^{(p)}(k)$ is the "extra" echo which includes from the previous segment; and $u^{(p)}(k)$ is the "missing" tail of echo of the speech of the current segment. As $f_{D\&P}$ is a zero-forcing equalizer, it may amplify the additive noise $e^{(p)}(k) = v^{(p)}(k) - u^{(p)}(k)$; and it can reduce the dereverberation accuracy.

To illustrate this problem, we consider a rectangular room with dimensions $L_x = 8m$, $L_y = 10m$, and $L_z = 4m$; and with wall reflection coefficients $\rho_x = \rho_y = \rho_z = 0.9$ (a reverberant room). A speech signal with duration of 8.8s, and sampled at 8 kHz is used as the original source signal. The reverberant speech signal is observed on a microphone array formed by 8 distinct microphones (spaced by 0.5m). A computer implementation (graciously provided by Geert Rombouts from K.U. Leuven) of the image method as described in [16] is used to generate synthetic room impulse response for the microphones.
As an evaluation criterion, we consider the Direct to Reverberant energy Ratio (DRR), defined as:

$$DRR = 10\log_{10}\left\{\frac{\sum_{n=0}^{\tau-1}\widetilde{h}^2(n)}{\sum_{n=\tau}^{L-1}\widetilde{h}^2(n)}\right\} \quad dB \tag{13}$$

where $\widetilde{h}(n) = h * f(n)$ denotes the equalized channel, τ is the number of samples to include as the direct component, and $L = T_{60} f_s$ is the length of the impulse response (T_{60} is the reverberation time, and f_s is the sampling frequency).

The table 1 gives the equalization DRR of the Delay-&-Sum beamformer, the classic Delay-&-Predict equalizer, and the windowed Delay-&-Predict equalizer (using a rectangular window, with an overlap of 50%). We consider the cases of 2, 4, and 8 microphones. We take $\tau = 10ms$.

	M=2	M=4	M=8
D & S	-3.81	0.8	1.94
classic D & P	7.43	9.34	10.71
windowed D & P	5.5	9.5	10.8

Table 1: Equalization DRR of the Delay-&-Sum beamformer, the classic and windowed Delay-&-Predict equalizers.

We notice that in all cases the Delay-&-Predict equalization outperforms the Delay-&-Sum beamforming, and that, due to the windowing effect, the frame-by-frame approach does not usually improve the equalization DRR.

In [18], the authors face the same problem when they estimate the channel cepstrum coefficients. They suggest using a smoothing window $w(k)$ to segment the speech signal. The goal of the windowing would be to reduce the error components by smoothly tapering the segment boundary. If $w(k)$ is sufficiently smooth (with respect to the channel), the received signal can be expressed approximately as:

$$\begin{aligned} y^{(p)}(k) &= (s(k) * h(k)) \, .w(k) \\ &= (s(k).w(k)) * h(k) \\ &= s^{(p)}(k) * h(k) \end{aligned}$$

The effects of time-domain windowing are investigated in [19]. Windowing should be chosen to reduce the error component in (12); while at the same time not introducing distorsion into the computed equalizer.
As in [19], we consider rectangular, Hamming, and exponential windows as candidates. The corresponding equalization DRR are given in table 2.

	M=2	M=4	M=8
Rectangular window	6.5	9.5	10.48
Hamming window	7.1	10.16	12.1
Exponential window	8	10.1	11.9

Table 2: Equalization DRR of the Delay-&-Predict equalizer using Rectangular, Hamming, and Exponential windows.

As expected, we see that Hamming and Exponential windows outperform the rectangular one. On the other hand, the exponential window surpasses the Hamming window only for a small number of microphones. In fact, exponential windows do not destroy the convolutional combination between the signal and the channel, i.e., $\lambda^k y(k) = \lambda^k s(k) * \lambda^k h(k)$. On the other hand, Hamming window is tailor-made for reduction of truncation error, but its effect upon the convolutional combination of signal is not known. As the number of microphones increases, the channel will be better equalized, and $\widetilde{h}^{(p)}$ tends to the Dirac pulse. Hence, the effect of the Hamming window on the convolutional combination is less perceptible.

Another way to address the problem is the use of an iterated dereverberation procedure:

1. an equaliser is computed (using the classic, or the windowed D-&-P.
2. using this equaliser, we cancel the error in (12); and we recomputed the D-&-P equaliser.
3. we iterate until convergence

The table 3 gives the equalization DRR of Iterated Delay-&-Predict equalizer. Simulation results reveal that this iterated equalization scheme (based on frame-by-frame analysis) increases the dereverberation performance, and leads to better auditive results.

	M=2	M=4	M=8
Iterated $D\&P$	8.5	11.3	15

Table 3: Equalization DRR of the Iterated Delay-&-Predict equalizer.

4. REFERENCES

[1] J. Flanagan, J. Johnston, R. Zahn, and G. Elko."Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Amer.*, pp.1508-1518, Nov. 1985.

[2] B.W. Gillespie, L.E. Atlas. "Acoustic Diversity for Improved Speech Recognition in Reverberant Environments," *In Proc. of ICASSP*,Vol.1, pp.557-560, May 2002.

[3] B.W. Gillespie, H.S. Malvar, D.A.F. Florencio. "Speech Dereverberation via Maximum-Kurtosis Subband Adaptive Filtering," *In Proc. of ICASSP*, Vol.6, pp.3701-3704, May 2001.

[4] H. Buchner, R. Aichner, W. Kellermann. "TRINICON: a Versatile Framework for Multichannel Blind Signal Processing," *In Proc. of ICASSP*, Vol.3, pp. 889-892, May 2004.

[5] N. Gaubitch, P.A. Naylor, and D.B. Ward. "On the Use of Linear Prediction for Dereverberation of Speech," *In Proc. of IWAENC*, pp.99-102, Sept. 2003.

[6] N. Gaubitch, P.A. Naylor, and D.B. Ward. "Multi-microphone speech dereverberation via spatio-temporal averaging," *In Proc. of EUSIPCO*, pp.809-812, Sept. 2004.

[7] J.L. Flanagan, A.C. Surendran, and E.E. Jan, "Spatially selective sound capture for speech and audio processing," *on Speech Communication*, Vol.13, pp.207-222, Oct. 1993.

[8] M. Miyoshi, and Y. Kaneda. "Inverse Filtering of Room Acoustics," *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol.36, Feb. 1988.

[9] Y. Huang, J. Benesty, and G. W. Elko. "Adaptive Eigenvalue Decomposition Algorithm for Real-Time Acoustic Source Localization System," *In Proc. of IEEE ICASSP*, Vol.2, pp.937-940, Mar. 1999.

[10] J. Chen, Y. Huang, and J. Benesty. "An Adaptive Blind SIMO Identification Approach to Joint Multichannel Time Delay Estimation," *In Proc. of IEEE ICASSP*, Vol.4, pp.53-56, May 2004.

[11] T. Hikichi, M. Delcroix, and M. Miyoshi. "Blind dereverberation based on estimates of signal transmission channels without precise information of channel order," *In Proc. of IEEE ICASSP*, Mar. 2005.

[12] C.B. Papadias, and D.T.M. Slock. "Fractionally Spaced Equalization of Linear Polyphase Channels and Related Blind Techniques Based on Multichannel Linear Prediction," *IEEE Trans. on Signal Processing*, Vol. 47, pp.641-654, Mar. 1999.

[13] Mahdi Triki and Dirk T.M. Slock. "Blind Dereverberation of a Single Source Based on Multichannel Linear Prediction," *In Proc. of IWAENC*, Sept. 2005.

[14] Mahdi Triki and Dirk T.M. Slock. "Delay and Predict Equalization for Blind Speech Dereverberation," *In Proc. of ICASSP*, May 2006.

[15] B. D. Radlovic, R. C. Williamson, and R. A. Kennedy. "Equalization in an Acoustic Reverberant Environment: Robustness Results," *IEEE Trans. Speech Audio Processing*, Vol.8, pp.311-319, May 2000.

[16] P.M. Peterson. "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," *J. Acoust. Soc. Amer.*, pp.1527-1529, Nov. 1986.

[17] D.T.M. Slock. "Form Sinusoids in Noise to Blind Deconvolution in communications". In Kailath, A. Paulraj, V. Roychowdhury, and C.D. Shaper, editors, *Communications, computation, control and signal processing*, Kluwer Academic Publishers, 1997.

[18] A. Oppenheim, R. Schafer, and T. Stockham. "Nonlinear Filtering of Multiplied and Convolved Signals," *IEEE Trans. on Audio and Electroacoustics*, Vol.16, pp.437-466 Sept 1968.

[19] D. Bees, M. Blostein, P. Kabal. "Reverberant Speech Enhancement Using Cepstral Processing," *In Proc. of ICASSP*, Vol.2, pp.977-980, April 2006.