BLIND DEREVERBERATION OF A SINGLE SOURCE BASED ON MULTICHANNEL LINEAR PREDICTION

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 LP will both equalize the reverberation filter and whiten the source. We exploit the observation that a multichannel reverberation filter tends to become allpass as the number of channels and/or the reverberation delay spread increases. As a result, the sum of the channel output correlations approximates the source correlation structure which can hence be used to determine a source whitening filter. Multichannel LP is then applied to the sensor signals filtered by the source whitenening filter, to obtain source dereverberation. 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.

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 confort 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. Speech enhancement for dereverberation and noise reduction in reverberant environments has been addressed extensively; but no adequate solution has yet been established [3, 2].

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 so as to focus in the direction of the desired speaker. The direction of arrival is generally adapted using a second-order statistic approach. In [4], the authors propose an alternative adaptive filtering approach using a kurtosis metric on the LP residual signal. They seek to find a blind deconvolution filter that makes the LP residual as non-Gaussian as possible. They show that the proposed technique achieves significant improvement in performance over the delay-and-sum beamformer.

A second class of speech dereverberation technques is based on source-filter speech production. The source-filter model describes speech production 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 propose using LP coefficients obtained by spatial averaging of the LP coefficients estimated on each microphone. In fact, by applying statistical room acoustic theory, one can show that the spatially expected values of the predicted LP coefficients obtained from the reverberant speech are equal to those obtained from clean speech.

Another way to address the problem is the use of an explicit model for the stationary channel impulse response. To avoid any channel-source identification ambiguities, each non-stationary source is modelled by a block stationary AR process; and each channel path by a stationary subband all-pole filter [6, 7]. Then using a Bayesian framework, the parameters of the distortion filter get estimated (source parameters are considered as nuisance parameters). In [3], the authors focus on the single-source two-microphone system; and solve the distortion due to the channel-source non-identification ambiguities using a common polynomial extraction technique: the common factor is extracted as a characteristic polynomial of the two-channel linear prediction matrix.

As we have seen, spatial-diversity and channel stationarity are two key ingredients in the multi-microphone speech dereverberation problem. This motivates us to propose a three-stage approach for speech dereverberation.

- 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

[*]Eurecom Institute's research is partially supported by its industrial members: Bouygues Télécom, Fondation d'entreprise Groupe Cegetel, Fondation Hasler, France Télécom , Hitachi, Sharp, ST Microelectronics, Swisscom, Texas Instruments, Thales. The research reported here was also sponsored in part by French RIAM project SIEPIA.

the previous step.

This paper is organized as follows. In section 2, the multichannel spatial diversity is investigated. The speech dereverberation procedure will then be derived in section 3. The performance of the algorithm is evaluated in section 4, and finally a discussion and concluding remarks are provided in section 5.

2. MULTICHANNEL SPATIAL DIVERSITY

We 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, $\mathbf{H}(q) = [H_1(q) \cdots H_M(q)]^T = \sum_i \mathbf{h_i} q^{-i}$ is the SIMO channel transfer function, and q^{-1} is the one sample time delay operator.

To investigate the effect of the multichannel spatial diversity, we consider a rectangular room with dimensions $L_x = 8m$ $L_y = 10m$ and $L_z = 4m$, and with wall reflection coefficients $\rho_x = 0.5$, $\rho_y = 0.5$, and $\rho_z = 0.2$. A speech signal with duration of 8.8s, and sampled at 8 kHz is used as the original source signal (figure 1). The reverberant speech signal is observed on 8 distinct

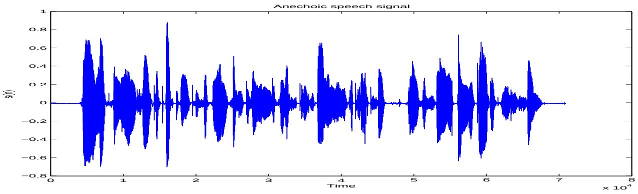

Fig. 1. Anechoic speech signal.

microphones. A computer implementation (graciously provided by Geert Rombouts from K.U. Leuven) of the image method as described in [8] is used to generate synthetic room impulse response for the microphones.

Figure 2(a) superposes the magnitudes of channel transfer functions $|\mathbf{H_k}(f)|^2$ $k = 1 : M$ between the source and the $M = 8$ microphones. The transfer function magnitude of the multichannel reverberation filter $\sum_{k=1}^{M} |\mathbf{H_k}(f)|^2$ is plotted in figure 2 (b). As can be seen, the multichannel transfer function tends to become flat, and the multichannel reverberation filter tends to be all-pass.

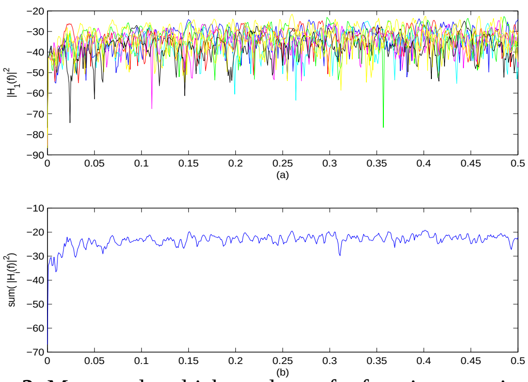

Fig. 2. Mono and multichannel transfer function magnitudes

By summing the spectra of the received signals, we get:

$$\sum_{k=1}^{M} S_{y_k y_k}(f) = \sum_{k=1}^{M} |H_k(f)|^2 \, S_{ss}(f) \approx c S_{ss}(f) \tag{2}$$

Then, due the multichannel spatial diversity, the superposition of the spectra of the received signals can estimate (up to a multiplicative constant c) the source spectrum (figure 3),

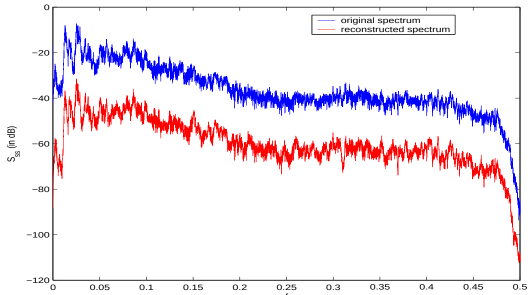

Fig. 3. Averaged periodogram of the original clean signal $P_{ss}^{welch}(f)$, and the reconstructed one $\sum_{k=1}^{M} P_{y_k y_k}^{welch}(f)$.

3. SPEECH DEREVERBERATION PROCEDURE

Motivated by the previous observations, we propose in this contribution a processing scheme that works with three cascade of stages:

- Source whitening stage: removes correlation due to the speech signal.
- Multichannel prediction stage: computes a blind multichannel predictor (using pre-processed reverberant speech).
- Dereverberation stage : equalizes the channel impulse response (using a zero-forcing equalizer based on the predictor computed in the previous step).

3.1. The source whitening stage

As we have seen previously, due to the multichannel spatial diversity, the superposition of the spectra of the received signals can estimate (up to a multiplicative factor) the source spectrum. This motivates us to remove correlation due to the source speech signal by compensating the common part on the multichannels impulse response. 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 prediction errors, averged over the microphones:

$$\begin{aligned} e &= \sum_{k=1}^{M} \sum_{n=0}^{\infty} e_k^2(n) \\ &= \sum_{k=1}^{M} \sum_{n=0}^{\infty} \left[y_k(n) - \sum_{j=1}^{l} a_j y_k(n-j) \right]^2 \end{aligned} \tag{3}$$

The previous optimization problem leads to the normal equations:

$$\begin{bmatrix} r_0 & r_1 & \cdots & r_{l-1} \\ r_1 & r_0 & \cdots & r_{l-2} \\ \vdots & & \ddots & \vdots \\ r_{l-1} & \cdots & r_1 & r_0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_L \end{bmatrix} = - \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_l \end{bmatrix} \tag{4}$$

where - $r_j = \sum_{k=1}^{M} r_{y_k y_k}(j)$

- $r_{y_k y_k}(j)$ represents the correlation at the time-lag j of the received signal at the k^{th} microphone

- $\{a_j\}$ are the common AR parameters.

In figure 4 we superpose the anechoic speech signal periodogram, and the AR spectral models estimated either using the source signal directly, or the sum of the correlation sequences of the M reverberant signals.

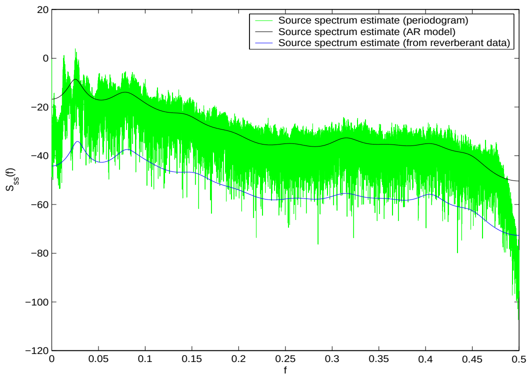

Fig. 4. Source periodogram, spectrums of AR processes estimated using the clean and the reverberant signals ($l = 20, M = 8$).

It can be seen that the AR Model estimated using reverberant signals gives a good estimation (up to a scalar) of the clean speech spectrum. Thus, it can be used to pre-process the reverberant speech in order to transform the colored source speech signals into a white signal.

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.

3.2. The multichannel prediction stage

The fundamental concept of blind estimation of SIMO linear systems arises from the observation that a rank one vector MA process can also be fully described as a rank one AR process under appropriate channel conditions. This observation allows prediction based algorithms to be developed for blind channel deconvolution.

The source whitened reverberant signal observed on M distinct microphones can be written as:

$$\mathbf{x}(k) = a(q)\mathbf{y}(k) \approx \mathbf{H}(q)\widetilde{s}(k) \tag{5}$$

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), $\widetilde{s}(k)$ is the source prediction error.

Consider now the problem of predicting $\mathbf{x}(k)$ from the L latest observations $\mathbf{X_L}(k-1) = [\mathbf{x}^T(k-1) \cdots \mathbf{x}^T(k-L)]^T$. The prediction error is given by:

$$\widetilde{\mathbf{x}}(k) = \mathbf{x}(k) + \sum_{i=1}^{L} A_{L,i}\mathbf{x}(k-i) = A_L \mathbf{X_{L+1}}(k) \tag{6}$$

where $A_L = [I_m \; A_{L,1} \; \cdots \; A_{L,L}]$, $A_{L,i}$ are the linear prediction filter coefficient matrices that should be determined to minimize the mean squared value of $\widetilde{\mathbf{x}}(k)$, L denotes the prediction order. Minimizing the energy of the prediction error leads to the system of equations (for large enough L [9]):

$$S_{\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}}(z) = A_L(z) S_{\mathbf{xx}}(z) A_L^{\dagger}(z) = \mathbf{h_0} S_{\widetilde{\mathbf{s}}\widetilde{\mathbf{s}}}(z) \mathbf{h_0}^H \tag{7}$$

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

- $A(z) = \sum_{i=0}^{L} A_{L,i} z^{-i}$ 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 $r_{\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}}(0)$.

Note that the proposed approach can be easily extended to the presence of an additive white noise, since the white noise variance can be easily identified and compensated for in the reverberant signal covariance matrix.

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. To estimate delays, a brut force method can be used by performing LP for all delay combination and choosing the one that maximizes the largest eigenvalue λ_{max} of the prediction error covariance matrix $r_{\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}}(0)$.

3.3. 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:

$$\mathbf{F_{D\&P}}(q) = \mathbf{h}_0^H A_L(q) D(q) \tag{8}$$

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:

$$\widehat{s}(k) = \mathbf{F_{D\&P}}(q)\mathbf{y}(k) = \mathbf{h_0}^H A_L(q)\mathbf{y}(k) \tag{9}$$

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

4. EXPERIMENTAL RESULTS

To analyze the validity of the proposed technique, we consider the reverberation scenario described in section 2. The delays for first paths are assumed to be known. Figure 5 (a) plots the equalized channel ($F_{D\&P} * H$) impulse response and spectrum, and the spectrum of the whitened source speech signal (preprocessed using a 20-order linear predictor).

We remark that due to the fact that the speech signal is a bandpass signal (observe values on very high and low frequencies), the

Delay-and-Predict equalizer has a tendency to amplify the missing frequency components (as it is a zero-forcing equalizer). To minimize this side effect, a larger order source whitening LP filter can be used (figure 5 (b)).

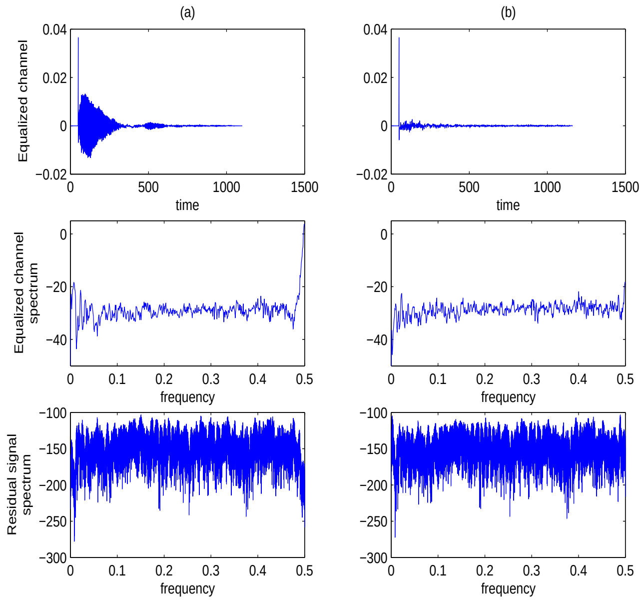

Fig. 5. Equalized channel impulse response and spectrum, and the source preprocessed speech signal. (a) $l = 20$. (b) $l = 100$

Figure 6 shows the increase of the dereverberation performance in terms of Signal-to-Echo Ratio ($SER = \frac{\sigma_s^2}{MSE}$), as a function of the source whitening LP order.

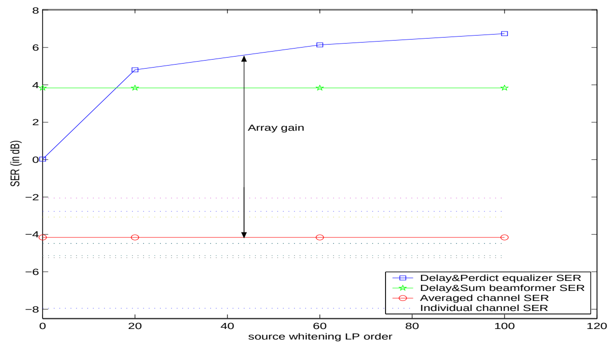

Fig. 6. delay-&-predict equalizer vs. delay-&-sum beamformer ($M = 8, L = 100$)

We remark also that the SER gain ($G = \frac{SER_{D\&P}}{SER_{D\&S}}$) is especially important if only few microphones are available (see figure 7). This is due to the fact that multichannel linear prediction performs well even using only two microphones; whereas the beamforming technique becomes an equalizer as the number of microphones increases.

5. CONCLUSION

In this paper, a linear prediction based dereverberation technique was proposed. The multichannel reverberation impulse response is assumed stationary enough to allow estimation of the correlations it induces in the received signals. Spatial, temporal, and spectral diversities are exploited to transform the source speech signal into

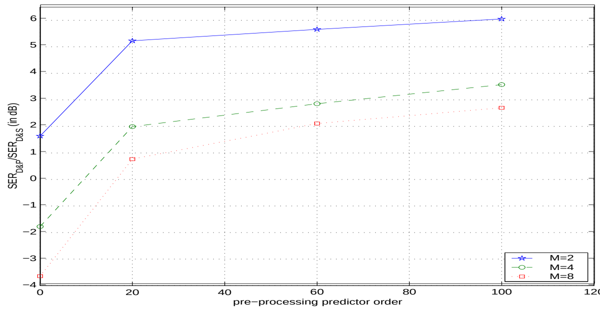

Fig. 7. The SER gain $G = \frac{SER_{D\&P}}{SER_{D\&S}}$ for 2, 4, and 8 microphones.

an whiter signal. An equalizer is then computed based on a multichannel linear prediction technique. Simulations shows that the Delay-and-Predict equalizer performs better than the delay-and-Sum beamformer, specially if only few microphones are available.

6. 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.*, Pages: 1508-1518, November 1985.

[2] B.W. Gillespie, L.E. Atlas. "Acoustic Diversity for Improved Speech Recognition in Reverberant Environments," *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*,Vol. 1, Pages: 557-560, May 2002.

[3] T. Hikichi, M. Delcroix, and M. Miyoshi. "Blind dereverberation based on estimates of signal transmission channels without precise information of channel order," *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, March 2005.

[4] B.W. Gillespie, H.S. Malvar, D.A.F. Florencio. "Speech Dereverberation via Maximum-Kurtosis Subband Adaptive Filtering," *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 6, Pages: 3701-3704, May 2001.

[5] N. Gaubitch, P.A. Naylor, and D.B. Ward. "On the Use of Linear Prediction for Dereverberation of Speech," *In Proceedings of the International Workshop on Acoustic Echo Noise Control* , pages 99-102, September 2003.

[6] J.R. Hopgood. "Bayesian Blind MIMO Deconvolution of Nonstationary Autoregressive Sources Mixed Through All-Pole Channels," *In Proceedings of the IEEE International Workshop on Statistical Signal Processing*, pages 422-425, September 2003.

[7] J.R. Hopgood, P.J.W. Rayner. "Blind Single Channel Deconvolution using Nonstationary Signal Processing," *IEEE Transactions on Speech and Audio Processing*, Vol. 11, Issue 5, pages 476-488, September 2003.

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

[9] 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.