CANCELLING CONVOLUTIVE AND ADDITIVE COLOURED NOISES FOR SPEECH ENHANCEMENT

W. Bobillet[1], E. Grivel[1], R. Guidorzi[2] and M. Najim[1]
[1] Equipe Signal et Image, UMR 51 31 LAP,
351 Cours de la libération, 33405 Talence Cedex, France
[2] Dipartimento di Elettronica, Informatica e Sistemistica, Università di Bologna,
Viale del Risorgimento 2, 40136 Bologna, Italy
email: {bobillet, eric.grivel, najim}@tsi.u-bordeaux1.fr, rguidorzi@deis.unibo.it

***Abstract*—** In the framework of speech enhancement, many approaches have been developed when the speech signal is only corrupted by an additive noise. However, in an auditorium, when echoes appear, spatial transformations between the sources and the microphones must be considered. For this reason, we propose in this paper to deal with a speech contaminated both by convolutive and additive coloured noises.

The two-microphone based noise canceller we present operates as follows: firstly, a pre-whitening step is carried out. Secondly, the blind deconvolution method we use makes it possible to estimate the finite impulse responses (FIR), their orders and the variances of the additive noises, which is a great advantage. Then, the filtered versions of speech, estimated by means of a subspace method, are used to retrieve the original speech.

Index Terms — speech enhancement, convolutive noise, additive noise, blind deconvolution, subspace methods

I. INTRODUCTION

In most voice communication systems, speech is corrupted by a background noise. In the last two decades, several methods using one microphone have been developed for speech enhancement, as depicted in FIG 1.

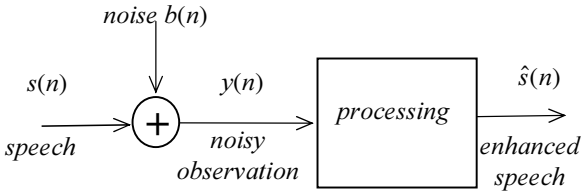

FIG 1: speech enhancement, speech contaminated by an additive noise

As an alternative to the non-parametric methods using short time spectral attenuation [2] [6], model-based speech enhancement can be considered.

On the one hand, when choosing an autoregressive (AR) model, Wiener [10] or Kalman filtering [8] can be performed to enhance speech. These methods usually require the estimation of the variance of the additive noise and the prediction coefficients from noisy observations, which is however a challenging issue.

On the other hand, the signal can be modelled, more realistically, as a sum of sinusoidal tracks. In [12], the authors propose to update, frame by frame, the estimation of the real magnitudes of the frequency components, by using a Wiener filter. The frequencies are then tracked by smoothing the spectral envelope of each analyzed frame. For sake of simplicity, a sum of complex exponentials is often preferred [16]. Thus, subspace methods have been investigated [11] [7]. The purpose is to separate the noisy observation subspace into two orthogonal subspaces:

- The signal subspace, estimated in the least square or minimum variance senses [11] or by introducing perceptually relevant estimation criteria [7], makes it possible to retrieve the original speech.
- The noise subspace provides information about the noise statistics.

This subspace decomposition is performed by using the Karhunen-Loève transform. The dominant eigenvalues of the noisy observation autocorrelation matrix correspond to the signal subspace while the lowest ones correspond to the variance of the additive noise. An alternative approach consists in only considering the dominant singular values of the Toeplitz noisy observation matrix. It should be noted that these methods have been also extended to the coloured additive noise case in [11].

When considering the block-diagram proposed in FIG 1, the speech and the additive noise are assumed to be uncorrelated. Therefore, speech enhancement consists in retrieving the first formants of speech by reducing the additive noise. However, in an amphitheatre or an auditorium where speech echoes may appear, these methods can not provide significant results since they do not take into account the spatial characteristics of both the source and the noise. For this reason, acoustic room FIR representing the spatial transformations between the sources and the microphones are introduced. In addition, several microphones must be used (Cf. FIG 2).

Few approaches have been recently developed to retrieve the source speech from the observations contaminated by convolutive and additive noises. In [4], the authors combine two methods: first, a NLMS-based multi-channel adaptive echo canceller is completed. Each "channel" resulting output is the sum of the source speech and the additive coloured noise. Then, a Generalized Singular Value Decomposition (GSVD)-based optimal filter, presented in [5], is completed to retrieve speech.

In this paper, we propose an alternative method which combines a blind deconvolution technique and a subspace method for speech enhancement. We consider a two-microphone based device (Cf. FIG 2). For each path, the observation $y_i(n)$ corresponds to the sum of a filtered version $x_i(n)$ of the speech signal $s(n)$ and an additive coloured noise $b_i(n)$.

Thus, the signal received by the i^{th} microphone can be expressed as follows:

$$y_i(n) = x_i(n) + b_i(n) = \sum_{k=0}^{q} g_i(k)s(n-k) + b_i(n) \qquad .(1)$$

where $g_i(n)$ denotes the acoustic room FIR for the i^{th} microphone.

Instead of directly dealing with this coloured noise case, we propose to introduce a pre-whitening step (Cf. FIG 2). This has the advantage to view the coloured case as an easier case, i.e. the additive white case.

The additive coloured noise $b_i(n)$ is assumed to be stationary, and hence can be modelled by a p^{th} order auto-regressive process:

$$b_i(n) = -\sum_{k=1}^{p} a_i(k)b_i(n-k) + w_i(n) \qquad (2)$$

where $w_i(n)$ is a zero-mean white noise with variance σ_i^2 and $\left(a_i(n)\right)_{n=1,\ldots,p}$ the prediction coefficients.

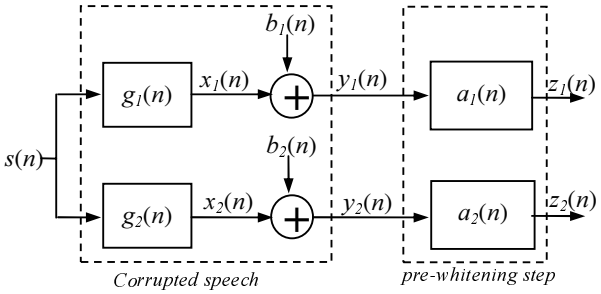

FIG 2: speech contaminated by both convolutive and additive coloured noises and pre-whitening step

Indeed, from relation (2), we obtain:

$$B_i(z) = \frac{W_i(z)}{A_i(z)} = \frac{W_i(z)}{1 + \sum_{k=1}^{p} a_i(k)z^{-k}} \qquad (3)$$

where $B_i(z)$ and $W_i(z)$ respectively denote the Z-transforms of $b_i(n)$ and $w_i(n)$.

The observations $y_1(n)$ and $y_2(n)$ are respectively filtered by the inverse filters $a_1(n)$ and $a_2(n)$, previously estimated from silent frames. Thus, given FIG 2, equations (2) and (3), one obtains:

$$\begin{aligned} Z_i(z) &= A_i(z)\times Y_i(z) = A_i(z)\times\left(X_i(z) + B_i(z)\right) \\ &= A_i(z)\times\left(G_i(z)\times S(z) + \frac{W_i(z)}{A_i(z)}\right) \\ &= A_i(z)\times G_i(z)\times S(z) + W_i(z) \end{aligned} \qquad (4)$$

$$Z_i(z) = H_i(z)\times S(z) + W_i(z) \qquad (5)$$

The resulting block-diagram is given in FIG 3, where $h_i(n) = (g_i * a_i)(n)$, which denotes the inverse Z-transform of $H_i(z)$, is a $(p+q) = L^{th}$ order FIR.

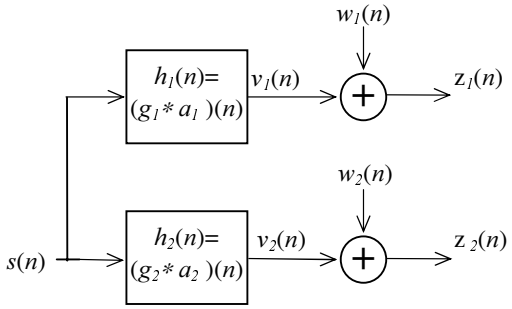

FIG 3 : equivalent scheme after the pre-whitening step

Let us introduce:

$$\underline{s}(n) = \left[s(n-L) \quad s(n-L+1) \quad \cdots \quad s(n+L)\right]^T \qquad (6)$$

$$\underline{v}(n) = \left[v_1(n) \quad \cdots \quad v_1(n+L) \quad v_2(n) \quad \cdots \quad v_2(n+L)\right]^T \qquad (7)$$

$$\underline{z}(n) = \left[z_1(n) \quad \cdots \quad z_1(n+L) \quad z_2(n) \quad \cdots \quad z_2(n+L)\right]^T \qquad (8)$$

$$\underline{w}(n) = \left[w_1(n) \quad \cdots \quad w_1(n+L) \quad w_2(n) \quad \cdots \quad w_2(n+L)\right]^T \qquad (9)$$

and the observation matrix defined as follows:

$$\mathrm{H}_L = \begin{bmatrix} \mathrm{H}_L^1 \\ \mathrm{H}_L^2 \end{bmatrix} \qquad (10)$$

where

$$\mathrm{H}_L^i = \begin{bmatrix} h_i(L) & h_i(L-1) & \cdots & h_i(0) & & \\ & \ddots & \ddots & & \ddots & \\ & & h_i(L) & h_i(L-1) & \cdots & h_i(0) \end{bmatrix}_{(L+1)\times(2L+1)} \qquad (11)$$

From relations (1), (2) and (5), we obtain:

$$\underline{z}(n) = \mathrm{H}_L\,\underline{s}(n) + \underline{w}(n)\,. \qquad (12)$$

Given the data $\underline{z}(n)$, the method we propose operates in the three following steps, summarized as follows:

- Firstly, a blind deconvolution method is used to estimate the FIR $h_1(n)$ and $h_2(n)$. Among the blind deconvolution approaches developed the very last years such as the TSML [13] [14], subspace method for deconvolution [15], we propose to consider the method proposed in [3] since it has also the advantage to provide the estimations of FIR orders and the variances σ_1^2 and σ_2^2 of the additive white noises $w_1(n)$ and $w_2(n)$ (even in an unbalanced case).
- Secondly, given the noisy observations $z_1(n)$ and $z_2(n)$ and the estimations of σ_1^2 and σ_2^2, the filtered versions of the speech $v_1(n)$ and $v_2(n)$ are estimated by using a subspace method for speech enhancement [7].
- Thirdly, speech signal $s(n)$ can be estimated in the least square sense, from the estimations of $h_1(n)$, $h_2(n)$, $v_1(n)$ and $v_2(n)$. Indeed, we have :

$$\hat{\underline{s}}(n) = \hat{\mathrm{H}}_L^{+}\, \hat{\underline{v}}(n) \tag{13}$$

where $\hat{\mathrm{H}}_L^{+}$ denotes the pseudo-inverse of the matrix $\hat{\mathrm{H}}_L$.

The remainder of this paper is organized as follows: in section II, we recall the blind deconvolution method proposed by one of the author in [3]. In section III, we provide some simulation results.

II. FOCUS ONE THE ESTIMATIONS OF THE VARIANCES OF THE ADDITIVE NOISE AND THE FIR

A. Estimating the variances from the observation autocorrelation matrix

The deconvolution approach proposed in [3] is based on the positive definiteness property of the autocorrelation matrix R_{zz}^{L} of the data $\underline{z}(n)$ and the non-negative definiteness of the autocorrelation matrix R_{vv}^{L} of $\underline{v}(n)$. In [1] [3], the authors show that:

$$\ker(\mathrm{R}_{zz}^{L}) = \ker(\mathrm{H}_L^{T}) = Span(\underline{\mathrm{c}}_L) \tag{14}$$

where

$$\underline{\mathrm{c}}_L = \left[\underline{\mathrm{c}}_2^{T}, -\underline{\mathrm{c}}_1^{T}\right]^{T} = \left[h_2(L),...,h_2(0),-h_1(L),...,-h_1(0)\right]^{T} \tag{15}$$

The FIR, $h_1(n)$ and $h_2(n)$, can therefore be obtained providing R_{vv}^{L} is available. However, this matrix is unknown, but can be estimated from R_{zz}^{L}. Indeed, since $\underline{v}(n)$ and $\underline{w}(n)$ are uncorrelated, one obtains:

$$\mathrm{R}_{vv}^{L} = \mathrm{R}_{zz}^{L} - \mathrm{R}_{ww}^{L} \tag{16}$$

where

$$\begin{aligned}\mathrm{R}_{ww}^{L} &= diag(\sigma_1^2\, \mathrm{I}_{L+1}, \sigma_2^2\, \mathrm{I}_{L+1}) \\ &= \begin{bmatrix} \sigma_1^2\, \mathrm{I}_{L+1} & 0 \\ 0 & \sigma_2^2\, \mathrm{I}_{L+1} \end{bmatrix}\end{aligned} \tag{17}$$

The estimation of the FIR, $h_1(n)$ and $h_2(n)$, is based on the preliminary estimations of σ_1^2 and σ_2^2. At that stage, one can pay attention to the method proposed in [1] where the authors introduce the matrix $\tilde{\mathrm{R}}^{l}(P) = diag\left(\alpha\, \mathrm{I}_{l+1}, \beta\, \mathrm{I}_{l+1}\right)$, for $l>0$, satisfying:

$$\hat{\mathrm{R}}^{l}(P) = \mathrm{R}_{zz}^{l} - \tilde{\mathrm{R}}^{l}(P) \geq 0\,. \tag{18}$$

The set of solutions (α, β) to equation (18) corresponds to a convex curve, in the plan (α, β), denoted $S\left(\mathrm{R}_{zz}^{l}\right)$. Then, estimating the variances of the additive noise (σ_1^2, σ_2^2) consists, in theory, in searching one common point P^* to the curves $S\left(\mathrm{R}_{zz}^{l}\right)$ with $l \geq L$ [1]. Cf. example in FIG 4.

However, in practical applications, P^* does not exist. For this reason, various criteria have been proposed to estimate σ_1^2, σ_2^2, and $\underline{\mathrm{c}}_L$ [3]. In the following, we will consider one of them.

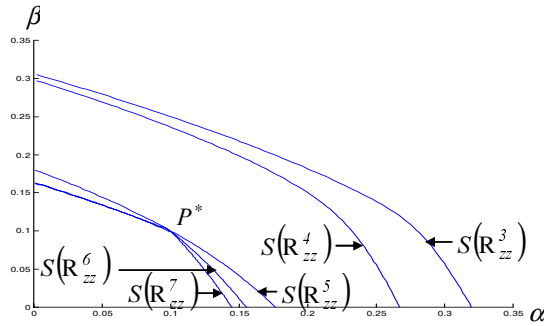

FIG 4: convex curves $S\left(\mathrm{R}_{zz}^{l}\right)$ and exhibition of a common point P^* corresponding to the variances of the additive noises. Synthetic case.

B. Implementing the deconvolution approach in real case

The criterion we consider is based on the following idea. First, one must notice that:

$$\begin{aligned}\hat{\mathrm{R}}^{L+1}(P^*) &= \mathrm{R}_{zz}^{L+1} - \tilde{\mathrm{R}}^{L+1}(P^*) \\ &= \mathrm{H}_{L+1}(P^*)\, \mathrm{R}_{ss}^{L+1}(P^*)\, \mathrm{H}_{L+1}^{T}(P^*)\end{aligned} \tag{19}$$

where

$$\mathrm{H}_{L+1}(P^*) = \begin{bmatrix} \mathrm{H}_{L+1}^{(1)}(P^*) \\ \mathrm{H}_{L+1}^{(2)}(P^*) \end{bmatrix} \tag{20}$$

However, when $P = P^*$, $h_i(l)$=0 for $l \geq L+1$. So we obtain:

$$\mathrm{H}_{L+1}^{(i)}(P^*) = \begin{bmatrix} \underline{0}_{(L+1)\times 1} & \mathrm{H}_{L}^{(i)}(P^*) & \underline{0}_{(L+1)\times 1} \\ 0 & 0...0\; h_i(L)...h_i(1) & h_i(0) \end{bmatrix} \tag{21}$$

In addition, we have:

$$\mathrm{R}_{ss}^{L+1}(P^*) = \begin{bmatrix} r_s(0) & \cdots & r_s(2L+2) \\ \vdots & \mathrm{R}_{ss}^{L}(P^*) & \vdots \\ r_s(2L+2) & \cdots & r_s(0) \end{bmatrix} \tag{22}$$

Therefore, the variances of the additive noises can be obtained by minimizing the following criterion:

$$J_2(P, \overline{P}) = \left\| \mathrm{R}_{ss}^{L}(P) - \mathrm{R}_{ss}^{L}(\overline{P}) \right\|_F \tag{23}$$

where $\|.\|_F$ denotes the Frobenius norm. (Cf [3] for more details).

C. Comparative study with TSML [13] and the Cramer-Rào lower bound (CRLB)

As an illustration, we propose to complete a comparative study using 5th order filters, for various SNR.

$$\begin{aligned}h_1(z) = &-1.1836 + 0.4906z^{-1} - 0.3093z^{-2} + 0.4011z^{-3} \\ &+ 0.1269z^{-4} - 1.8522z^{-5}\end{aligned}$$

$$\begin{aligned}h_2(z) = &\;0.8221 + 0.0333z^{-1} + 0.2162z^{-2} - 0.0165z^{-3} \\ &+ 0.2531z^{-4} - 0.5591z^{-5}\end{aligned}$$

Results are based on the following criterion:

$$NRMSE = \frac{1}{\|\underline{\mathrm{h}}\|} \sqrt{\frac{1}{R} \sum_{k=1}^{R} \left\| \hat{\underline{\mathrm{h}}}_{(k)} - \underline{\mathrm{h}} \right\|^2} \tag{24}$$

with $\underline{h} = [h_1(0)\cdots h_1(L)\ h_2(0)\cdots h_2(L)]^T$, $\hat{\underline{h}}_{(k)}$ denotes the estimation of the RIF for the k^{th} run and R the number of runs.

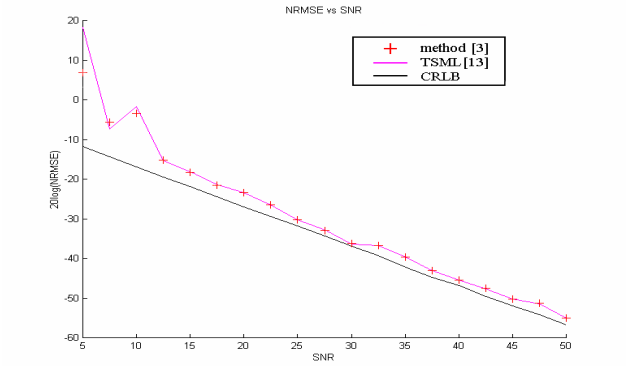

FIG 5: NRMSE(dB) vs input SNR(dB) for various deconvolution approach

III. SIMULATION RESULTS, CONCLUSION AND PERSPECTIVES

The speech enhancement approach is exercised with a speech signal, sampled at 8 KHz, then filtered by synthetic FIR filters whose orders q are equal to 100; $x_i(n)$ are contaminated by an additive 5th order auto-regressive process. Cf. Table 1.

X_iNR (dB)	$=10\log_{10}\left(\frac{\sum_k x_i(k)^2}{\sum_k (y_i(k)-x_i(k))^2}\right)$	10	20	30	40	50
Input SNR 1 (dB)	$=10\log_{10}\left(\frac{\sum_k s^2(k)}{\sum_k (y_1(k)-s(k))^2}\right)$	3.34	4.39	4.53	4.55	4.55
Input SNR 2 (dB)	$=10\log_{10}\left(\frac{\sum_k s^2(k)}{\sum_k (y_2(k)-s(k))^2}\right)$	1.49	1.85	1.88	1.89	1.89
Output SNR (dB)	$=10\log_{10}\left(\frac{\sum_k s^2(k)}{\sum_k (\hat{s}(k)-s(k))^2}\right)$	3.93	11.96	27.05	36.54	40.01

Table 1: average Output SNR vs X_iNR, based on 100 runs

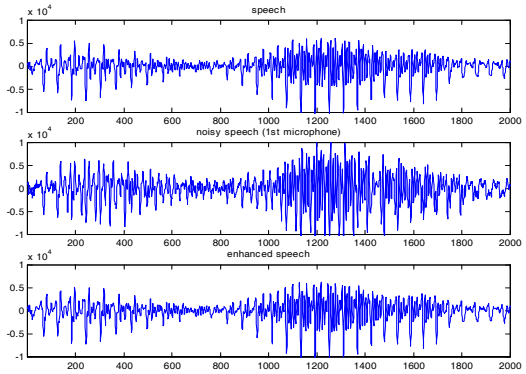

FIG 6: speech, speech contaminated both by an echo and an additive coloured noise (SNR=20, 1st microphone) and enhanced speech

In addition to SNR improvement criterion, informal subjective tests have been performed and show significant results, especially when the so-called X_iNR is higher than 20dB.

In this paper, our purpose was to deal with a more realistic speech enhancement situation. For this reason, we have taken into account the influence of the spatial features of the room. When using subspace methods for speech enhancement, estimating the variances of the additive noises is problematic. Here, we take advantage of the deconvolution step to obtain the noise statistics.

We are currently working on an alternative to the intermediate application of the subspace filtering technique that would consist in directly applying a minimal variance estimation approach. In addition, brand new simulations are in progress with FIR order much higher.

REFERENCES

[1] S. Beghelli, R.P.Guidorzi and U.Soverini, The Frisch Scheme in Dynamic System Identification, Automatica 26, 1990.

[2] M. Berouti, R. Schwartz and J. Makhoul, Enhancement of Speech Corrupted by Acoustic Noise, Proc. of the IEEE ICASSP 79, pp. 208-211.

[3] P. Castaldi, R. Diversi, R.P Guidorzi and U. Soverini, Blind Estimation and Deconvolution of Communication Channels with Unbalanced Noise, 12th IFAC Symposium on System Identification, Santa Barbara, California, June 2000.

[4] S. Doclo and M. Moonen, GSVD-Based Optimal Filtering for Single and Multimicrophone Speech Enhancement, IEEE Trans. on Signal Processing, vol. 50, n°. 9, pp. 2230-2244, Sept. 2002.

[5] S. Doclo, E. de Clippel and M. Moonen, Combined Acoustic Echo and Noise Reduction using GSVD-based optimal filtering, ICASSP 00, Istambul, Turkey, vol. 2, pp. 1061-1064.

[6] Y. Ephraim and D. Malah, Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator, IEEE Trans. Acoust., Speech, Signal Processing, vol. 32, n°. 6, 1984, pp. 1109-1121.

[7] Y. Ephraim and H. L. Van Trees, A signal Subspace Approach for Speech Enhancement, IEEE Trans. Speech Audio Processing, vol. 3, n°4, pp 251-266, July 1995.

[8] J. D. Gibson, B. Koo and S. D. Gray, Filtering of Colored Noise for Speech Enhancement and Coding, IEEE Trans. on Signal Processing, vol. 39, n°8, pp. 1732-1742, August 1991.

[9] E. Grivel, M. Gabrea and M. Najim, Speech Enhancement as a realization issue, Signal Processing, vol. 82, n°12, pp. 963-978, Dec. 2002.

[10] J. H. L. Hansen and M. A. Clements, Constrained Iterative Speech Enhancement with Application with to speech recognition, IEE Trans. Signal Processing, vol. 39, n°4, pp. 795-805, Apr. 1991.

[11] S. H. Jensen, P. C. Hansen, S. D. Hansen and J. Sorensen, Reduction of Broad Band Noise in Speech by Truncated QSVD, IEEE Trans. On Speech and Audio Processing, vol. 3, n°6, 1995, pp. 439-448.

[12] J. Jensen and J. L. Hansen, Speech Enhancement Using a Constrained Iterative Sinusoidal Model, IEEE Trans. On Speech and Audio Processing, vol. 9, n°7, Oct. 2001.

[13] Y. Hua, Fast Maximum Likelihood for Blind Identification of Multiple FIR Channels, IEEE Trans. on Signal Processing, vol. 44, no 3, March 1996.

[14] Y. Hua and M. Wax, Strict Identifiability of Multiple FIR Channels Driven by an Unknown Arbitrary Sequence, IEEE Trans. on Signal Processing, vol. 44, n° 3, March 1996.

[15] E. Moulines, P. Duhamel, J. Cardoso and S. Mayrargue, Subspace Methods for the Blind Identification of Multichannel FIR Filters, IEEE Trans. on signal processing, vol. 43, no 2, February 1995.

[16] T.F. Quatieri and R.J McAulay, Noise reduction using a soft-decision sine-wave vector quantizer, ICASSP 90, Albuquerque, New Mexico.