A Novel Affine Projection Algorithm for Superdirective Microphone Array Beamforming

Danilo Comminiello, Michele Scarpiniti, Raffaele Parisi and Aurelio Uncini
INFOCOM Dpt., "Sapienza" University of Rome,
Via Eudossiana 18, 00184 Rome.
Email: comminiello@infocom.uniroma1.it

***Abstract*— This paper describes a new adaptive algorithm and assesses its effectiveness within speech enhancement applications. The proposed *variable step size block exact APA* (VSS-BEAPA) filtering is based on the *affine projection algorithm* (APA) and introduces a block processing with a variable step size that allows to consider under-modeling scenarios. The algorithm shows improved convergence performance and computational efficiency and its robustness is proved in a typical context of a hands-free teleconferencing application in a noisy environment. The experiments show that a microphone array system joined with VSS-BEAPA filtering is capable of both decreasing the noise level and enhancing the speech signal quality.**

I. Introduction

Most of hands-free mobile telephony and teleconferencing systems tend to amplify all environment sounds, without trying to detect the main speech signal. In an office environment the background noise typically arises from computer fans, traffic, audio equipment or other speakers present in the room (e.g. cocktail party noise). Background noise can compromise the intelligibility of recorded speech signals and resulting in binaural information loss. This is one of the reasons why speech enhancement is currently an important area of research.

Beamforming techniques intend to enhance the speech in teleconferencing applications. A classic beamforming system that enhances speech is the *generalized sidelobe canceller* (GSC) [1] composed of a fixed *delay-and-sum beamformer* (DSB) and adaptive noise cancelling path that enables the microphone interface to adapt to varying noise conditions, providing additional attenuation of undesired noise sources and leading to lower noise power in the beamformed output.

Teleconferencing applications, as well as other hands-free applications, require adaptive filters with hundreds or even thousands of taps. Their success depends on the nature of the acoustic impulse response. Generally the adaptation of the canceller uses classic least-mean-square-based algorithms, such as *least mean square* (LMS) and *normalized LMS* (NLMS); however, these algorithms display a very slow convergence for long filters [2] such that adaptation becomes unpractical in hands-free applications. The *affine projection algorithm* (APA) [3] and other APA-based algorithms were used in adaptive beamforming [4] showing better convergence rates and manageable computational complexity.

In this paper a *variable step size block exact affine projection algorithm* (VSS-BEAPA) is proposed. The VSS-BEAPA is an *exact* transposition in the frequency domain of a *block APA* [5] with a variable step size that allows to consider under-modeling situations [6] that occur when the length of the adaptive filter is shorter than the length of the impulse response, as is the rule in hands-free applications.

A GSC with a VSS-BEAPA filtering is evaluated in terms of speech quality measure, showing improved performance with respect to classic configurations. Furthermore, the addition of a post-filter is investigated.

This paper is organized as follows: in Section II a superdirective beamforming system is introduced. In Section III the VSS-BEAPA filtering is described. Section IV is devoted to the performance analysis of the proposed system. In Section V our conclusions are drawn.

II. Superdirective Beamforming Technique

The chosen superdirective beamforming system, depicted in Fig. 1, is derived from *near-field superdirectivity* (NFSD) [7] and consists of a DSB and an adaptive sidelobe cancelling path in typical GSC configuration. It is well known that standard DSB is not well suited for the task of speech enhancement because of its poor directivity index at low frequencies [7]. This shortcoming is covered by a proper microphone interface as well as by the adaptive noise cancelling path. The main assumption made is that the desired source is situated in the array's near-field while the dominating noise sources are located in the far-field, as is generally the case in the chosen application.

Let us consider a microphone array composed of *M* sensors. The signal $u_m[n]$ acquired by the *m*-th microphone, with $m = 1, \ldots, M$, contains a delayed replica of the target signal $s[n]$ with the addition of background noise $v_m[n]$:

$$u_m[n] = s[n] + v_m[n] \tag{1}$$

As shown in Fig. 1, the DSB spatially aligns the microphone signals with reference to the speech source direction and generates the speech reference $d[n]$. The adaptive path receives the input signals and generates the noise references $x_i[n]$, with $i = 1, \ldots, M-1$, by means of the *blocking matrix* (BM). These signals are then filtered by the *adaptive noise canceller* (ANC) which removes the correlation of the residual noise component in the speech reference and the noise references, generating the beamformer output $e[n]$. This structure exploits the microphone array configuration further maximizing the

978-1-4244-5309-2/10/$26.00 ©2010 IEEE

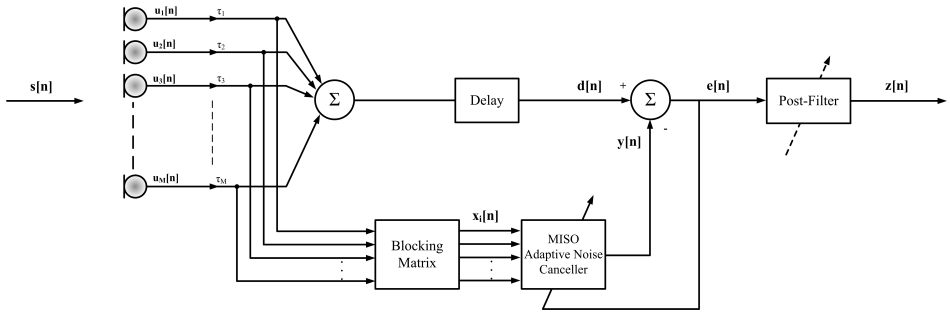

Fig. 1. Speech enhancement system.

directivity index in the desired source direction and reducing interfering signals derived from diffuse noise field.

In order to improve the performance of the system, a Zelinski post-filter [8] is added to the beamformer. The whole system benefits from some important properties of the post-filter. Firstly, a full cancellation of incoherent components of the signal is obtained, so that a high reverb reduction can be expected. Moreover, the time-varying nature of the post-filter allows to consider non-stationary acoustic environments. Zelinski post-filter estimation is based on the the cross- and auto-spectral densities of the microphone signals.

III. Variable Step Size Block Exact APA

The adaptive filtering occuring in the ANC block plays a fundamental role in the superdirective beamforming process. The ANC can be seen as a MISO (*multiple-input single-output*) system composed of a bank of adaptive filters, each relative to a microphone signal. The cancellation of each filter output from the speech reference $d\left[n\right]$ yields the estimate of the background noise $y\left[n\right]$ and the beamformer output $e\left[n\right]$.

In speech enhancement applications, the adaptive filter order can be large thus requiring a large computational cost. Computational complexity is proportional to the number of coefficients; therefore, adaptation can become prohibitively expensive, compromising real-time implementation. For this reason advanced adaptive algorithms with high complexity, such as the *recursive least squares* (RLS) algorithm [2], are not taken into consideration. On the other hand, cheaper algorithms, such as *least mean squares* (LMS) and *normalized LMS* (NLMS) [2] converge slowly, especially with speech signals.

The proposed VSS-BEAPA is derived from a frequency-domain implementation of the block APA [5] with a time-varying step size [6]. Each iteration provides a block of P samples of the beamformer output $e\left[n\right]$. Let us denote with $b = 1, \ldots, B$ the block index, where B is the number of blocks. The beamformer output relative to the block b is a $P \times 1$ vector defined as:

$$\mathbf{e}_n^{[b]} = \mathbf{d}_n^{[b]} - \mathbf{y}_n^{[b]} \tag{2}$$

where $\mathbf{d}_n^{[b]}$ is a selection of the $L \times 1$ DBS output vector $\mathbf{d}_n = \left[d\left[n\right], d\left[n-1\right], \ldots, d\left[n-L+1\right]\right]^T$. Similarly, $\mathbf{y}_n^{[b]}$ is an ANC output block, achieved by:

$$\mathbf{y}_n^{[b]} = \frac{1}{M-1} \sum_{i=1}^{M-1} \mathbf{X}_{\mathbf{i}n}^T \mathbf{w}_n \tag{3}$$

In equation (3), M is the number of microphones and $\mathbf{X}_{\mathbf{i}n}$ is the $L \times P$ reference noise matrix, created using a projection order P for each reference noise signal, and defined as:

$$\mathbf{X}_{\mathbf{i}n} = \begin{bmatrix} \mathbf{x}_{\mathbf{i}n} \\ \mathbf{x}_{\mathbf{i}n-1} \\ \cdots \\ \mathbf{x}_{\mathbf{i}n-P+1} \end{bmatrix}^T \tag{4}$$

$$= \begin{bmatrix} x_i\left[n\right] & x_i\left[n-1\right] & \cdots & x_i\left[n-P+1\right] \\ x_i\left[n-1\right] & x_i\left[n-2\right] & \cdots & x_i\left[n-P\right] \\ \vdots & \vdots & \ddots & \vdots \\ x_i\left[n-L+1\right] & x_i\left[n-L\right] & \cdots & x_i\left[n-P-L+2\right] \end{bmatrix}$$

In (3), the $L \times 1$ vector $\mathbf{w}_n$ contains the coefficients of the adaptive filter. For each microphone, the resulting update equation of the VSS-BEAPA algorithm is:

$$\mathbf{w}_n = \mathbf{w}_{n-1} + \mu\left[n\right] \mathbf{X}_{\mathbf{i}n} \mathbf{R}_n^{-1} \mathbf{e}_n^{[b]} \tag{5}$$

where $\mathbf{R}_n = \mathbf{X}_{\mathbf{i}n}^T \mathbf{X}_{\mathbf{i}n} + \delta \mathbf{I}$ is the $P \times P$ input covariance matrix, δ is a regularization parameter and $\mu\left[n\right]$ is the time-varying step size. In equations (3) and (5) the matrix-vector products are computed using the fast Fourier transform (FFT), as done in [5]. Due to block processing, the inversion of $\mathbf{R}_n$ can be simplified in the following way [2]. Denoting with $\mathbf{\Gamma}_{\mathbf{U}n} = \mathbf{X}_{\mathbf{i}n}^{\lceil P \rceil T} \mathbf{X}_{\mathbf{i}n}^{\lceil P \rceil}$ and $\mathbf{\Gamma}_{\mathbf{D}n} = \mathbf{X}_{\mathbf{i}n}^{\lfloor P \rfloor T} \mathbf{X}_{\mathbf{i}n}^{\lfloor P \rfloor}$ the first and the last $P \times P$ sub-matrices of $\mathbf{X}_{\mathbf{i}n}^T \mathbf{X}_{\mathbf{i}n}$, respectively, we can write $\mathbf{R}_n$ in a recursive way:

$$\mathbf{R}_n = \mathbf{R}_{n-1} + \mathbf{\Gamma}_{\mathbf{U}n} - \mathbf{\Gamma}_{\mathbf{D}n-1} \tag{6}$$

Instead of compute the $(P \times L) \times (L \times P)$ product of $\mathbf{X}_{\mathbf{i}n}^T \mathbf{X}_{\mathbf{i}n}$, equation (6) computes $\mathbf{R}_n$ by means of the only updating matrices $\mathbf{\Gamma}_{\mathbf{U}n}$ and $\mathbf{\Gamma}_{\mathbf{D}n-1}$. It should be noted that the matrix $\mathbf{\Gamma}_{\mathbf{D}n-1}$ was already computed at time $n-1$, thus the computation of $\mathbf{R}_n$ requires only one $(P \times P) \times (P \times P)$ product per iteration. The initialization of $\mathbf{R}_n$ is $\mathbf{R}_0 = \delta \mathbf{I}$, where $\mathbf{I}$ is the $P \times P$ identity matrix.

According to [6], we choose a variable step size parameter $\mu[n]$ that allows to take into account an under-modeling scenario that occurs when the length of the filter L_F is shorter than the length L of the *acoustic impulse response* (AIR). Therefore the chosen variable step size is:

$$\mu[n] = \begin{cases} \mu_f, & n \leq L_F \\ \left| 1 - \frac{\sqrt{\left|\widehat{\sigma}_s^2[n] - \widehat{\sigma}_y^2[n]\right|}}{\widehat{\sigma}_e[n] + \xi} \right|, & n > L_F \end{cases} \tag{7}$$

where ξ is a small positive number that avoids division by zero. The general parameter $\widehat{\sigma}_\alpha^2[n]$, where $\alpha = \{s, y, e\}$, represents the power estimate of the sequence $\alpha[n]$. This can be computed as:

$$\widehat{\sigma}_\alpha^2[n] = \lambda \widehat{\sigma}_\alpha^2[n-1] + (1-\lambda)\alpha^2[n] \tag{8}$$

where λ is a weighting factor chosen as $\lambda = 1 - 1/(KN_F)$, with $K > 1$. The initial value is $\widehat{\sigma}_\alpha^2[0] = 0$. Due to the fact that for the first L_F iterations the filter is not under-modeled, we start the process using the original fixed step size μ_f for $n \leq L_F$, when the estimate of the coefficients is influenced only by the system noise $v[n]$. The computation of the power estimates in (7) could lead to minor deviations from the previous theoretical conditions, so that we can consider the absolute value of the step size parameter.

As it is evident from (5) and (7), the VSS-BEAPA algorithm uses parameters available exclusively from the adaptive filter, i.e. $s[n]$, $y[n]$, and $e[n]$. All the information concerning the acoustic nonstationarity is contained in the relation (7). This feature gives robustness to the VSS-BEAPA algorithm in very noisy environments.

In comparison with the NLMS or time-domain block APA, the VSS-BEAPA displays an improved convergence performance. It achieves a considerable cost reduction and reduced latency due to block processing in the frequency domain and to the possible choice of under-modeled filters.

IV. Simulation Results

A. Evaluation of the VSS-BEAPA filtering

In the following set of experiments we prove the effectiveness of the proposed algorithm in adverse environment conditions. In particular, we analyze a common scenario in teleconferencing applications, in which the acoustic environment changes due to a nonstationary source or to an alteration in the environmental conditions. The experiments take place in a $10 \times 6,6 \times 3$ m room with a reverberation time of $T_{60} = 300$ ms. The AIR is simulated by means of a Matlab tool (*Roomsim*[1]) and is measured by using an 8 kHz sampling rate with $L = 2048$ coefficients. The target signal is a white Gaussian noise. A further independent white Gaussian noise with zero mean and unit variance is added as background noise with a *signal to noise ratio* (SNR) of 20 dB. Both signals have a length of 10 seconds. In the parameter settings, we choose the following values: $K = 2$, $\xi = 0.0001$, $\delta = 30\sigma_x^2$, where σ_x^2 is the power of the filter input signal, and $\mu = \mu_f = 0.2$. For APA algorithms we choose a projection order of $P = 2$, and in VSS-BEAPA we set the length of the filter to $L_F = 1024$ samples. In order to measure the algorithms performance we use the normalized misalignment $\mathcal{M}_{um}$, expressed in dB, for under-modeling scenarios, defined as:

$$\mathcal{M}_{um} = 20\log_{10}\left(\frac{\|\mathbf{w}_{I_n} - \mathbf{w}_{F_n}\|_2}{\|\mathbf{w}_{I_n}\|_2}\right) \tag{9}$$

where $\mathbf{w}_{I_n}$ is the AIR column vector, and $\mathbf{w}_{F_n} = [w_0[n], w_1[n], \ldots, w_{L_F}[n], 0_{L_F+1}[n], \ldots, 0_L[n]]^T$ is the estimated filter in the under-modeling case.

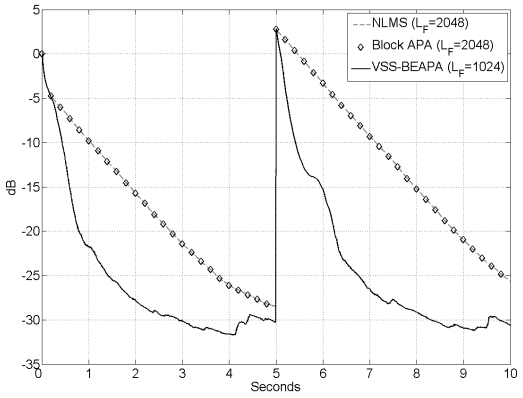

Fig. 2. Misalignment comparison. The impulse response changes after 5 seconds. VSS-BEAPA uses a filter length of $L_F = 1024$ samples.

In order to introduce an abrupt change in the acoustic environment we shift the acoustic impulse response circularly to the right by 20 samples, 5 seconds after the start of the adaptive process. Figure 2 shows that while NLMS and time-domain block APA display roughly the same convergence rate, the VSS-BEAPA has a lower misalignment; however, VSS-BEAPA reacts faster than other algorithms when the impulse response changes.

B. Evaluation of the Superdirective Beamformer with the VSS-BEAPA filtering

In order to evaluate the beamforming system described above, we consider the same configuration of the previous simulations, but in this case the source of interest is a male speaker located 70 cm from a microphone array.

The choice of the microphone array geometry plays an important role in recovering the binaural perception. An optimal array for speech enhancement applications should possess a large aperture in order to achieve a good spatial resolution and at the same time it should avoid spatial aliasing. The chosen microphone interface derives from [7] and is composed of 11 elements, consisting of 9 microphones looking at the source of interest, with 2 further *endside* microphones behind the two

[1]Roomsim is a MATLAB simulation of shoebox room acoustics for use in teaching and research. Roomsim is available from http://www.mathworks.com/ matlabcentral/fileexchange/authors/14085.

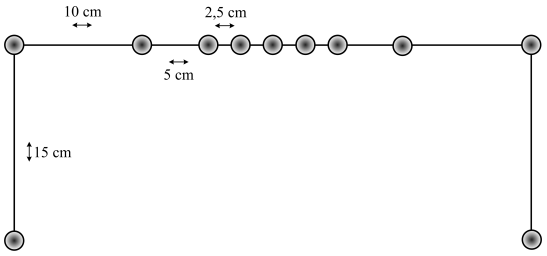

Fig. 3. Microphone interface geometry.

end microphones, as shown in Fig. 3. The two endside microphones are used in order to achieve a larger information at low frequencies. All the 11 microphones are cardioid elements. The choice of adopting cardioid microphones derives from the need to minimize undesired signals coming from side and back of the array.

The enhancement of speech and noise reduction is usually associated with a SNR improvement, defined as:

$$SNR = 10 \log \left[\frac{E\left\{ r_{in}\left[n\right]^2 \right\}}{E\left\{ r_{in}\left[n\right]^2 \right\} - E\left\{ r_{out}\left[n\right]^2 \right\}} \right] \qquad (10)$$

where $r_{in}\left[n\right]$ is the generic input clean signal and $r_{out}\left[n\right]$ is the processed signal. The operator $E\left\{\cdot\right\}$ is the mathematical expectation.

The speech spectral content is time-varying; this is the reason why the background noise affects the desired speech signal in different ways according to the variant nature of speech phonemes. In order to get a better evaluation of the achieved enhancement, we join the SNR measure with the *log area ratio* (LAR) measure [9]. The LAR is an objective measure of the dissimilarity between the original and processed speech signals; it derives from the *linear prediction coefficients* (LPC), a highly effective representation of the speech signal. In LAR analysis, the vocal tract of a person is modelled as a non-uniform acoustic tube formed by cascading a number of q cylindrical tubes of uniform equal length with different cross-section areas. The LAR coefficients are formed by the ratio between the cross-section areas of every two connected tubes. Hence, the LAR measure is defined as [9]:

$$LAR = \left[\frac{1}{q} \sum_{k=1}^{q} \left(\log \frac{1 + r_{in}\left[k\right]}{1 - r_{in}\left[k\right]} - \log \frac{1 + r_{out}\left[k\right]}{1 - r_{out}\left[k\right]} \right)^2 \right]^{\frac{1}{2}} \qquad (11)$$

The shorter this value is, the better the speech quality of the enhancement.

In our experiments, we measure a 5 dB SNR input level. We calculate SNR and LAR distances considering the following four structures: the first is the only delay-and-sum beamformer (DSB); the second is a GSC with the DSB and an adaptive noise canceller path whose adaptive filter is the NLMS algorithm (GSC NLMS). The third is the GSC with the proposed VSS-BEAPA filtering algorithm (GSC VSS-BEAPA) and the last is the GSC VSS-BEAPA with the addition of a Zelinski post-filter [8] (GSC VSS-BEAPA + Post-Filter). Speech analysis has been carried out on 20 ms speech frames with a 5 ms overlap. The results are collected in Table I where it is evident that the proposed system with the post-filter is consistently superior to standard configurations, such as the DSB and classic GSC beamformers, in terms of noise reduction as well as perceptual distortion.

TABLE I
SPEECH QUALITY COMPARISON.

	SNR (dB)	**LAR**
DSB	9.3	5.6
GSC NLMS	17.4	4.1
GSC VSS-BEAPA	24.6	3.5
GSC VSS-BEAPA + Post-Filter	27.1	3.1

V. CONCLUSION

This paper introduced a novel adaptive algorithm, VSS-BEAPA, that shows improved convergence performance and computational efficiency with respect to existing techniques. Its effectiveness is assessed in a speech enhancement system that combines a delay-and-sum beamformer with an adaptive noise canceller and a post-filter. The VSS-BEAPA exploits the noise reference signals derived from the blocking matrix and subtracts them from the output of the fixed path of the beamformer, bringing about a strong reduction of background noise. This system also determines an improvement of speech enhancement in terms of SNR and LAR measures. Performance advantages stand out especially in the case of high level of background noise, which is typical in such applications.

ACKNOWLEDGEMENT

This work has been partially supported by the Italian National Project: Wireless multiplatfOrm mimo active access netwoRks for QoS-demanding muLtimedia Delivery (WORLD), under grant number 2007R989S.

REFERENCES

[1] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. on Antennas and Propagation*, vol. 30(1), p. 27 34, 1982.
[2] A. H. Sayed, *Fundamentals of Adaptive Filtering*. Wiley, 2003.
[3] K. Ozeki and T. Umeda, "An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties," *Electron Commun. Japan*, vol. 67-A, pp. 19–27, 1984.
[4] Y. R. Zheng and R. A. Goubran, "Adaptive beamforming using affine projection algorithms," *WCCC-ICSP 2000*, vol. 3, pp. 1929–1932, 2000.
[5] M. Tanaka, S. Makino, and J. Kojima, "A block exact fast affine projection algorithm," *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 1, pp. 79–86, Jan. 1999.
[6] C. Paleologu, S. Ciochina, and J. Benesty, "Variable step-size nlms algorithm for under-modeling acoustic echo cancellation," *IEEE Signal Processing Letters*, vol. 15, pp. 5–8, 2008.
[7] W. Tager, "Near field superdirectivity (nfsd)," *Proc. of ICASSP '98*, pp. 2045–2048, 1998.
[8] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," *Proc. of ICASSP '88*, pp. 2578–2581, Apr 1988.
[9] S. Quackenbush, T. Barnwell, and M. Clements, *Objective Measures of Speech Quality*. NJ: Prentice-Hall, 1988.