# SOURCE SEPARATION FOR MICROPHONE ARRAYS USING MULTICHANNEL CONJUGATE GRADIENT TECHNIQUES

*J.A Beracoechea, S.Torres-Guijarro,E.Terleira, L.Ortiz, F.J. Casajús*

Signal Proc. Applications Group, Dept. SSR
ETSI Telecomunicación-UPM, Madrid, Spain
berako@gaps.ssr.upm.es

*Lino García*

Dept. Ing. Eléctrica Electrónica
Universidad Europea de Madrid
lino.garcia@uem.es

**ABSTRACT**

This paper proposes a new scheme to improve the source separation problem aimed to microphone array applications like WFS based teleconference systems. A multichannel, sub-band approach to reduce computational complexity is presented. Also, instead of using the LMS adaptive algorithm, a new system based on hybrid Conjugate Gradient-nLMS techniques is developed to accelerate the convergence time. This adaptive algorithm is controlled by a voice activity detector block that basically detects double talk situations and freezes the adaptation process to avoid the appearance of sound artifacts which may cause a significant degradation of the recovered signals and have a great impact in the quality of the full system.

## 1. INTRODUCTION

Microphone array systems are quickly evolving over the last years. The ability of this kind of setups to enhance speech signals using spatial information has been very useful to develop a new kind of multichannel applications. One of the most promising is developing advanced teleconference systems using Wave Field Synthesis techniques to render acoustic fields [1]. The audio communication system known as acoustic opening uses arrays of microphones and arrays of loudspeakers to produce the illusion that between two remote rooms there is a mechanical opening [2]. Unfortunately, working with so many channels means that there is an enormous amount of information to deal with. One of the proposed ideas is to send only the dry sources and recreate the wave field at reception [3][4]. In figure 1 we can see the full scheme where a WFS system synthesizes the wave field produced by primary sources in the emitting room.

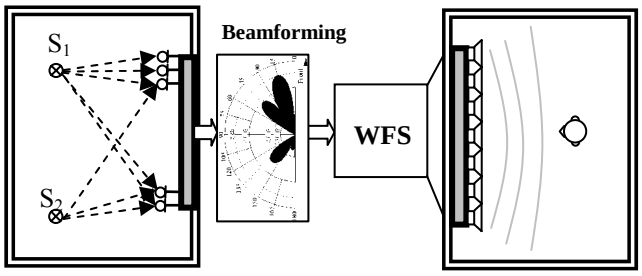


Figure 1: *Source Separation + WFS approach*

This scheme leads us to the problem of obtaining the dry sources given that we only know the signals captured with the microphone array and the position of the speakers. As you can see, basically, this is a source separation problem. The spatial diversity provided by the microphone array is often exploited to recover these dry signals. The scheme known as Generalized Sidelobe Canceller (GSC) [5] is widely used in these kind of situations because it can obtain a high interference reduction performance with a small number of microphones arranged in a small space [5].

In figure 2 we can see the block diagram of a GSC. The microphone signals are time delayed steered ($\tau_1 \ldots \tau_M$) to produce signals which ideally have the desired signal in phase with each other. If we add all these signals (d(n)) we have a fixed beamformer (FB). A delayed version of d(n): d'(n) (to keep causality) is used as reference for the multichannel canceller (MC). The delayed signals (before adding) are also sent to the blocking matrix. The purpose of the blocking matrix is to block out the desired signal from the lower part of the GSC. The idea is to adaptively cancel out noise and interference sources; therefore we only want noise to go into the MC.

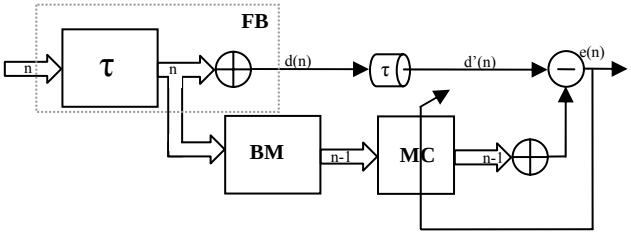


Figure 2: *GSC block diagram*

If the input signals of the MC contain only interferences it rejects the interferences and extracts the target signal (e(n)).

The GSC approach can achieve good results; however it has to face several problems. If the target signal leaks through the blocking matrix, the adaptive algorithm cancels both the interference and the desired signal (the original dry source we want to recover). This leakage can be caused by a bad Direction of Arrival (DOA) tracking or a highly reverberant room. This problem can be partially solved using adaptive blocking matrixes [6][7] which may double the computational complexity of the system. And even in this case real may systems suffer from distortions at the output (which is highly undesirable) in double talk situations so some kind of voice activity detector is often needed to avoid these problems [8]. Another problem with the GSC approach is that most of

its implementations are based on Least Mean Square (LMS) methods which are very simple but suffer from a long convergence time. In a time varying environment, where small head movements of the speaker can change the response of the filter that we have to synthesize (because the impulse response of the room changes) convergence speed is something critical.

This paper tests an algorithm that combines elements of the GSC with new concepts that try to reduce the impact of these problems. This algorithm provides with multichannel subband adaptive filtering employing high speed conjugate gradient methods to reduce convergence time while keeping low computational time and latency. In section 2 we describe the proposed system and in section 3 we briefly describe the CG-based method developed for our multichannel adaptive system. Finally, experimental setup and results are described in section 4 followed by conclusions.

## 2. MULTICHANNEL ADAPTIVE FILTERING

In figure 3 we can see the block diagram of the proposed system. This figure shows a particular case with three sources speaking at the same time. One of the sources acts like the source we want to recover and the other two are interferers. We suppose that the position of the sources is known. The signals coming the microphones are time-delay steered using fixed beamformers (FB1,FB2,FB3) The difference from the GSC is that the interference references do not come from a blocking matrix but from these fixed beamformers steered to the position of those interferers. This means that the filters to synthesize are longer but we may not need to employ an adaptive blocking matrix. The adaptive BM can certainly reduce the target signal leakage but generally definition of constraints [7] (which are related to bad DOA estimation) to the LMS algorithm (to avoid the appearance of sound artefacts at the output), is needed.

At these point we have three signals: d(n) which is an early estimation of the desired target signal. And $i_1(n)$and $i_2(n)$ which act as the interfering signals entering the multichannel canceller (MC). The idea is to adaptively cancel out the noise and interferences still present in d(n) using $i_1(n)$and $i_2(n)$. The output of the system e(n) is the best estimation of the original dry signal $s_1(n)$.

As we are working with quite long impulse responses the output system delay and computational effort may make the adaptive filtering unfeasible. This effect is even worse taking into account that the computational requirements of the CG methods are higher than LMS algorithms. To overcome this issue a number of methods based on subband decomposition of system input signals have been developed. Subband decomposition allows reducing the complexity of the general system by a factor that approximately equals to the order of the number of subbands.

This subband decomposition can be seen in figure 3 too. First of all, the desired signal d(n) and the interfering signals $i_1(n)$ and $i_2(n)$ pass a GDFT based filter bank and a decimation block where the decimation factor K is 2/3 of the number of filters (M) to avoid aliasing problems. The adaptation process takes place here, where every FIR filter tries to minimize the output using each subband of the desired signal and the appropriate subbands of the reference signals. The problem using filter banks is that they usually introduce a significant delay to the system. To avoid this problem we have used the approach described in [9] where filter adaptation is made in the subband domain but the filtering of the full signal is made through a transformation (T block) that translates the subband adaptive filters into full band filters. The transformation can be seen in (1) where $\mathbf{h_m}$ represents the analysis filters (m=1..M), $\mathbf{f_m}$ the synthesis filters and $\mathbf{c_{mp}}$ represents the coefficients of the adaptive filter in subband m and channel p (p=1,2 as we just have two interferers).For our case we have employed a 16 filter bank, and decimation factor K=12.

$$\mathbf{w}_p = real\left\{\sum_{m=1}^{M}\left[\mathbf{h}_m\big|_{\downarrow K} * \mathbf{c}_{mp}\right]_{\uparrow K} * \mathbf{f}_m\right\} \quad (1)$$

The convolution in time domain has no inherent delay but it is computationally very costly so it is not a good choice to be used with large filters in real time as in our case. To make this process more efficient we have implemented a partitioned convolution algorithm (PBFDF block) [10] which can achieve better performance than time convolution with low delays. The subband decomposition added to the multichannel filtering and fast convolution techniques increase the performance of the system around 10 times. This effect can be seen in figure 4. This simulation shows the time needed to arrive to a particular filter in both scenarios. Both simulations have same conditions (600 tap filters, CG algorithm). The only difference is that in the first one we employed full band processing while in the second one we used subband processing (M=16, K=12). This configuration has, also, the advantage of being able to adapt with different parameters, and moreover, different filtering algorithms in each subband although this idea is still unexplored.

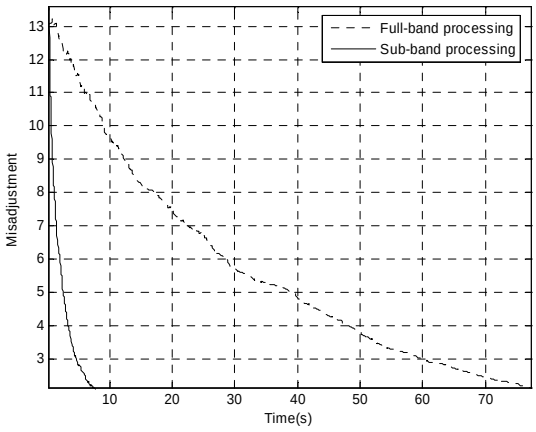


Figure 4: *Full-bad processing vs Sub-band processing*

## 3. ADAPTIVE ALGORITHM

With the general framework already clear we take a more detailed look to the adaptive algorithm used to actualize the filter coefficients. As we have mentioned before the Least Mean Square family algorithms are very well known and used methods because of its simplicity and speed but they suffer from bad convergence time and this is not desirable in a quickly time varying environment. Conjugate Gradient methods [11], originally designed for minimizing convex quadratic functions have been extended to the

general case with some variations. As an optimization strategy this method can be considered between the steepest descent (SD) and 2nd order Newton method. As you can see in figure 5 in a very simple simulation of typical noise cancelling situation (64 tab-long filter) it is much quicker than LMS method while keeping its computational complexity under the requirements of the Recursive Least Square (RLS) method.

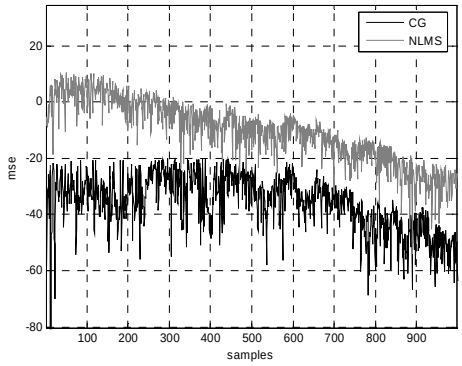


Figure 5: *Adaptive algorithm: CG vs NLMS.*

The problem with CG algorithm is that it is very sensitive to situations of low correlation between the upper branch signal d(n) and the reference signals that enter the MC. This can cause the appearance of highly undesirable sound artefacts at the output. A number of solutions have been proposed to avoid this problem including using constrained versions of the CG methods like in [12]. Our solution employs a hybrid algorithm CG-NLMS controlled by a voice activity detector (**vad** block) that changes the algorithm employed depending on the circumstances.

The **vad** block applied to d(n) decides if the adaptive algorithm is going to work with a CG algorithm when we are sure that d(n) contains only interferences, or freeze the adaptation process to avoid the appearance of sound artefacts. When the **vad** cannot decide in which situation we are, we continue the adaptation process using a standard NLMS algorithm.

## 4. EXPERIMENTAL RESULTS

### 4.1. Microphone Setup

To obtain the microphone signals we are employing the impulse-response recordings of two chambers. The first one is the varecoic chamber in Bell Labs [13], corresponding to different audio source locations in a chamber with a 22-microphone linear array and the second one corresponds to impulse response recordings of our teleconference chamber here at ETSIT-UPM. We employed pseudo-random sequences (MLS) to recover the impulse responses from 3 different audio locations. In figure (6) and (7) we can see the geometry of the chambers as well as the position of the three different source locations for each chamber used in our tests: v21,h2 (left side of the microphone arrays), v34,h0 (centre), v27,h1 (right side) .

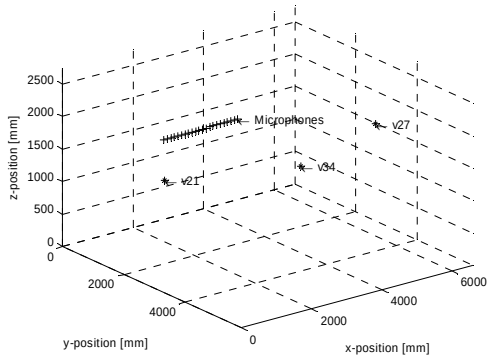


Figure 6: *Bell Labs chamber. Position of Sources and Interference*

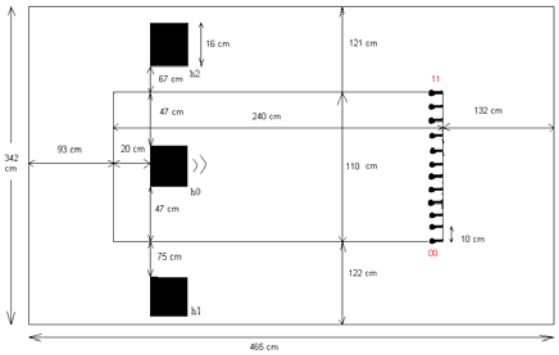


Figure 7: *ETSIT chamber*

Bell Lab's T60 reverberation time is around 0.28s while ours is lower as you can see in figure (8) where two representatives IR are shown. In both cases microphone separation is 10 cm. The tests presented in this paper use the Bell Lab's chamber as it is a worse scenario.

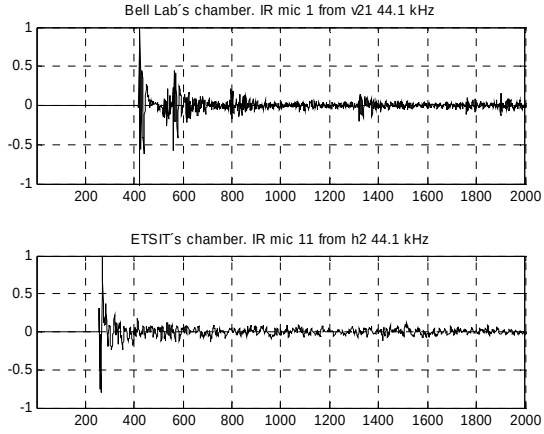


Figure 8: *Typical IR for both chambers*

### 4.2. Source Separation tests

We have considered several different scenarios. In the first one we just use as interferer one source of white noise situated in position v27 and we want to recover a male speech signal situated in position v21. In figure 9 you can see the results for this setup. On the upper part of the picture we see the original dry signal (male speech, $f_s$=16 kHz). In the middle, the signal recorded by the central microphone of the array (signal + noise). In the lower part we can see the system output (pseudo-dry signal).

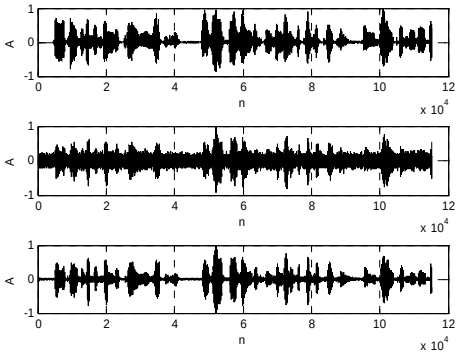


Figure 9: *a) original male dry signal b)signal on 11th microphone c)pseudo-dry signal*

Next scenario (figure 10) shows a more complicated environment, with two sources (male speech situated in v21 and female in v27) and one interferer (white noise in v34). The male pseudo-dry signal at the output shows a great improvement in SNR terms while avoiding noticeable sound artefacts. The recovered female speech shows similar performance.

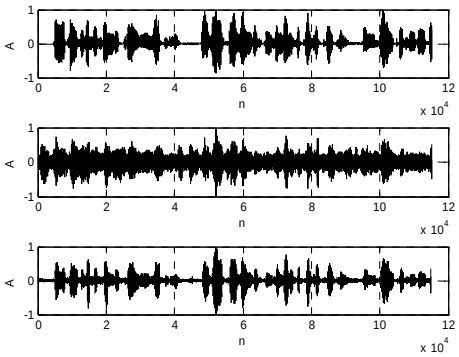


Figure 10: *a) original male dry signal b)signal on 11th microphone c)pseudo-dry signal recovered*

### 4.3. Sound Field reconstruction using WFS.

To see how similar our estimation is to the original dry signal we could use some kind of SNR measurements. However, we intend to use the separated signals as the input of a WFS system to recreate the acoustic field so some subjective tests using a real 10 speaker array (figure 11) and WFS synthesis had been carried out.

Figure 11: *Loudspeaker array*

The idea is to see how the imperfections of the recovered signals affect the reconstructed acoustic field. These preliminary tests try to have an idea about the loss of quality, intelligibility, and spatial localization properties suffered when we use our source separation + WFS approach against the hard-wired WFS approach where the signals captured at the microphones directly attack the loudspeaker array. Two very simple situations are tested. In the first one we try to see the noise reducing abilities of the system setting a male voice in 'v21' and a heavy noise in 'v27'. The noise power is calculated so listeners (situated in the centre of the receiving room) were unable to understand anything using the hard-wired WFS approach at reception. The same listeners were able to reproduce the sentence after the source separation (and noise reduction) and WFS synthesis. Moreover, the listeners were capable of situating the source correctly coming from the correct position.

The second situation uses two voice signals at the same time (male in 'v21' and female in 'v27') and tries to determine how well the spatial properties of the sound field are maintained when using the estimated sources to synthesize the acoustic field. After reconstruction the listeners were able to describe the male voice coming from the right and the female's voice coming from the left but they had more problems to precisely situate the source's origin than in the previous experiment were the interferer was white noise. This may be caused because in the source separation block, the imperfections in the male source (that is, the female voice we could not reject) are not perfectly aligned with the recovered female voice and vice versa. This fact (which does not happen with white noise) distorts the reconstructed field and affects the ability of the listeners to precisely describe the origin of the sources.

## 5. CONCLUSIONS

The proposed system that simplifies the blocking matrix approach, processes in the subband domain, uses fast convolution techniques and adds a voice activity detector to avoid sound artefacts can achieve quite good source separation. At the same time, the computational requirements have improved and the CG-NLMS hybrid algorithm which obtains better convergence times. And although this separation is not perfect, our preliminary subjective tests show

that the residual interferences are further attenuated once the sound field in synthesized again at reception using WFS. The information gathered using these subjective methods may point to further improvements in the source separation block which is the real bottle neck of the system.

## 6. ACKNOWLEDGEMENTS

This work was supported by Spanish Science and Technology Department trough projects TIC 2003-09061-C03-01 and "Ramón y Cajal"

## 7. REFERENCES

[1] M.M.Boone, Acoustic rendering with wave-field synthesis. ACM Siggraph and Eurographics Campfire. May 2001.

[2] Aki Härmä, Coding Principles for Virtual Acoustic Openings AES 22nd Internacional Conference.

[3] J.A Beracoechea, S.Torres-Guijarro, L.Ortiz, F.J. Casajús. "Source separation for WFS acoustic opening applications" 7th Int. Conference DAFX 2004 Naples

[4] H.Buchner, S.Spors, W. Kellermann, R.Rabenstein. "Full-duplex communications systems using loudspeaker arrays and microphone arrays" IEEE International Conference on Multimedia and Expo,ICME 2002 26-29 Aug. 2002

[5] L.J. Griffths, C. Jim, "An alternative approach to linearly constrained adaptive beamforming". IEEE Transactions on Antennas and Propagation Vol. AP-30 NO.1 January 1982

[6] H.Herbordt, W.Kellermann Efficient frequency-domain robust generalized sidelobe canceller Multimedia Signal Processing, 2001 IEEE Forth Workshop on, 3-5 Oct 2001 pp. 377-382.

[7] O. Hoshuyama, A. Sugiyama, A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters. IEEE Transactions on Signal Processing". Vol. 47, NO. 10. October 1999

[8] Z.M. Saric, S.T. Jovicic. "Adaptive microphone array based on pause detection" Acoustics Research Letters Online. 3 March 2004.

[9] J.P. Reilly, M.Wilbur, M.Seibert and N.Ahmadvand: "Complex subband decomposition and its applications to the decimation of large adaptive filtering problems" Submitted for publication in IEEE Transactions on Signal Processing

[10] E.R. Ferrara "Fast Implementation of LMS filters" IEEE Transactions on Acoustics, Speech and Signal Processing. Vol. ASSP-28 NO.4 August 1980

[11] L.García, M.Torres-Guijarro, J.A.Beracoechea, F.J.Casajús Quirós. "Conjugate gradient techniques for multichannel adaptive filtering" Submitted for publication DAFX2005

[12] J.A. Apolinario Jr., M.L.R. de Campos, C.P.Bernal "The constrained conjugate gradient algorithm" IEEE Signal Processing Letters Vol. 7 Nº 12 December 2000

[13] Bell Labs varecoic chamber: http://www.bell-labs.com/org/1133/Research/Acoustics/VarecoicChamber

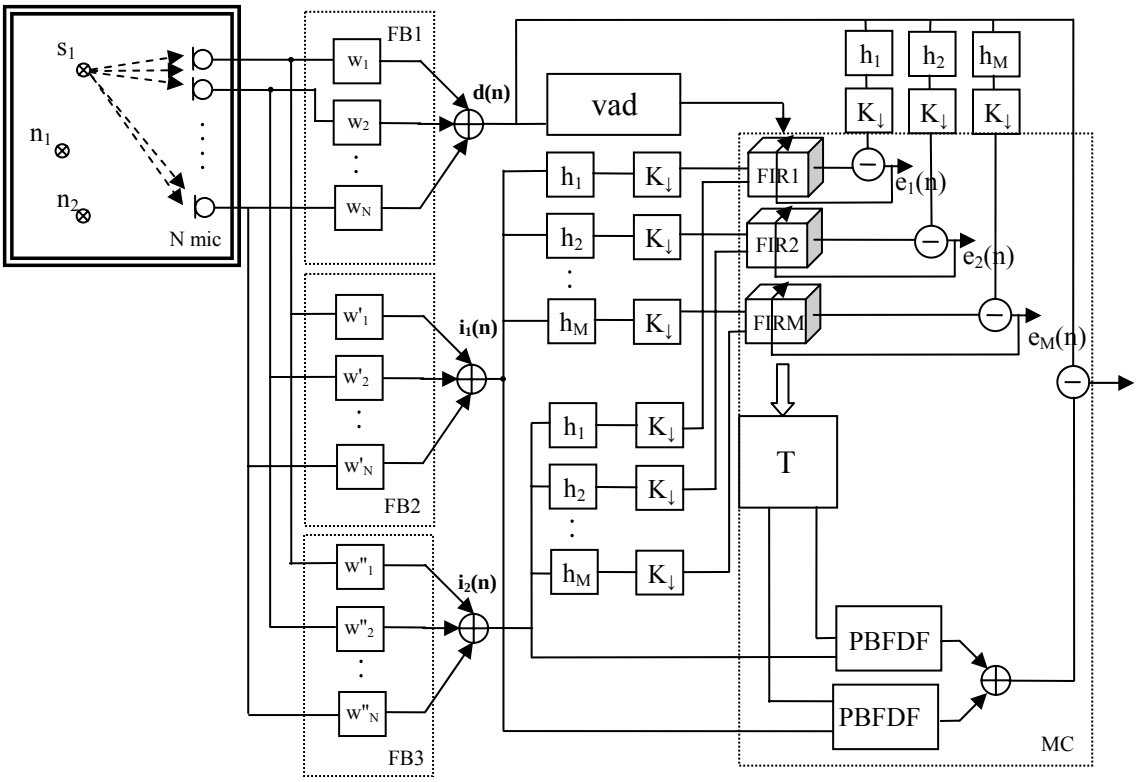


Figure 3: *Block diagram of the proposed system*