Hindawi Publishing Corporation
EURASIP Journal on Applied Signal Processing
Volume 2006, Article ID 40960, Pages 1–12
DOI 10.1155/ASP/2006/40960

On Building Immersive Audio Applications Using Robust Adaptive Beamforming and Joint Audio-Video Source Localization

J. A. Beracoechea, S. Torres-Guijarro, L. García, and F. J. Casajús-Quirós

Departamento de Señales, Sistemas y Radiocomunicaciones, Universidad Politécnica de Madrid, 28040 Madrid, Spain

Received 20 December 2005; Revised 26 April 2006; Accepted 11 June 2006

This paper deals with some of the different problems, strategies, and solutions of building true immersive audio systems oriented to future communication applications. The aim is to build a system where the acoustic field of a chamber is recorded using a microphone array and then is reconstructed or rendered again, in a different chamber using loudspeaker array-based techniques. Our proposal explores the possibility of using recent robust adaptive beamforming techniques for effectively estimating the original sources of the emitting room. A joint audio-video localization method needed in the estimation process as well as in the rendering engine is also presented. The estimated source signal and the source localization information drive a wave field synthesis engine that renders the acoustic field again at the receiving chamber. The system performance is tested using MUSHRA-based subjective tests.

1. INTRODUCTION

The history of spatial audio started almost 70 years ago. In a patent filled in 1931 Blumlein [1] described the basics of stereo recording and reproduction which can be considered as the first true spatial audio system. At that time, the possibility of creating "phantom sources" supposed a major breakthrough over monaural systems. Some years later, it was finally determined that the effect of adding more than two channels did not produce so much better results to justify the additional technical and economical efforts [2]. Besides, at that time, it was very difficult and expensive to develop simultaneous recording of many channels so stereophony became the most used sound reproduction system in the world until our days.

In the 1970's some efforts tried to enhance the spatial quality by adding 2 more channels (quadraphony) but the results were so poor that the system was abandoned. Lately, we have seen the development of a number of sound reproduction systems that use even more channels to further increase the spatial sound quality. Originally designed for cinemas, the five-channel stereo (or 5.1) adds 2 surround channels and a center channel to enhance the spatial perception of the listeners. Although well received by industry and general public, results with these systems range from excellent to poor depending on the recorded material and the way of reproduction.

In general, all stereo-based systems suffer from the same problems. First of all, the position of the loudspeakers is very strict and any change in the setup distorts the sound field. Secondly, the system can only render virtual sources between loudspeaker positions or further but not in the gap between the listener and the loudspeakers. Finally, perhaps the most important problem is that the system suffers from the so-called "sweet spot" effect. That means that there is only a very particular (and small) area with good spatial quality (Figure 1).

In parallel with the development of stereophony some work to avoid this "sweet spot" effect was being investigated. In 1934 Snow et al. [3] proposed a system where the performance of an orchestra is recorded using an array of microphones and the recording is played back to an audience through an array of loudspeakers in a remote room (in what we could call a hard-wired wavefield transmission system, as we will see later). This way, we could produce the illusion that there is a real mechanical window, that he called "virtual acoustic opening," between two remote rooms (Figure 2). Unfortunately, the idea was soon abandoned due to the enormous bandwidth necessary to send the signals which was way beyond the realms of possibility at that time.

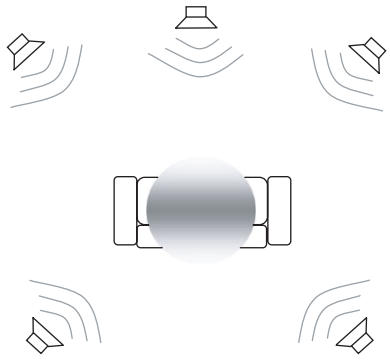

Figure 1: Sweet spot in 5.1 systems.

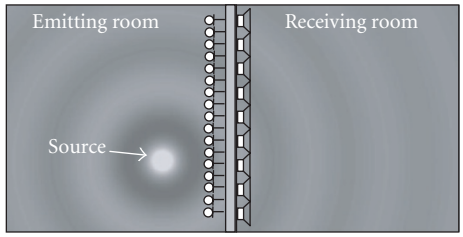

Figure 2: Acoustic opening concept.

Nowadays, with the advent of powerful multichannel perceptual coders, (like MPEG4) this kind of schemes is much more feasible and the "acoustic opening" concept is again being revisited [4].

Using as much as 64 Kbps/channel it is possible to transparently codify these signals before transmission, efficiently reducing the overall bandwidth. Furthermore, some recent work [5], that exploits the correlation between microphone signals, obtains a 20% reduction over those values. Clearly, when the number of sources is high (like in a live orchestra transmission) this is the way to go. However, the acoustic window concept can be used to build several other applications where the number of sources is low (or even one like in teleconference scenarios). In those speech-based applications, sending as many signals as microphones seems to be really redundant.

Over the last 5–10 years a new way of dealing with this problem has attracted the attention of the audio community. Basically the new framework [6, 7] explores the possibility of using microphone array processing methods to make an estimation of the original dry sources in the emitting room. Once obtained, the acoustic field is rendered again at reception using wave field synthesis (WFS) techniques.

WFS is a sound reproduction technique based on the Huygens principle. Originally proposed by Berkhout [8] the synthetic wave front is created using arrays of loudspeakers that substitute individual loudspeakers. Again, there is no "sweet spot" as the sound field is rendered all over the listening area (simulation in Figure 3). Being a well-founded wave theory, WFS replaces somehow the intuitive "acoustic opening" concept of the past.

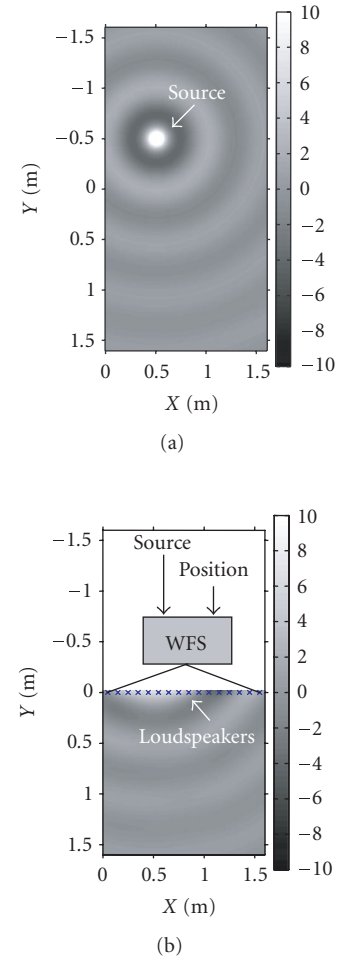

Figure 3: Wave field synthesis simulation. (a) Acoustic field primary monochromatic source. (b) Rendered acoustic field with WFS using a linear loudspeaker array.

The advantages of this scheme over the previous systems are enormous. First of all, the number of channels to be sent is dramatically reduced. Instead of sending as many channels as microphones we just need to send as many channels as simultaneous sources in the emitting room. Secondly, reverberation and undesirable noises can be greatly reduced in the estimation process as we will see in next sections. Finally, the ability of being capable of rebuilding with fidelity an entire acoustic field has enormous advantages for developing future speech communication systems [9, 10] in terms of overall quality and intelligibility.

This paper explores the possibility of building such kind of systems. The problems to be solved are reviewed and several solutions are proposed: microphone array methods are employed for enhancing and estimating the sources and providing the system with localization information. The impact of those methods after the sound field reconstruction (via WFS) has been also explored. A real system using two chambers and two arrays of transducers has been implemented to test the algorithms in real situations. The paper is organized as follows. Section 2 deals with the problems to be solved and

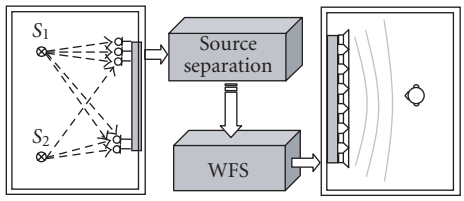

Figure 4: Source separation + WFS approach.

describes the different strategies we are using in our implementation. Sections 3 to 7 focus on the different blocks of our scheme. Section 8 shows some subjective tests of the system followed by conclusions and future work.

2. GENERAL FRAMEWORK

As mentioned in the previous section, within this approach, the idea is to send only the dry sources and recreate the wave field at reception. This leads us to the problem of obtaining the dry sources given that we only know the signals captured with the microphone array. As you can see, basically, this is a source separation problem (Figure 4)

From a mathematical point of view, the problem to solve can be resumed in expression (1). There are P statistically independent wideband speech sources $(S_1, \ldots, S_P)$ recorded from an M-microphone array ($P < M$). Each microphone signal is produced as a sum of convolutions between sources and H_{ij} which represent a matrix of z-transfer functions between P sources and M microphones. This transfer function set contains information about the room impulse response and the microphone response.

We make the assumption that source signals S are statistically independent processes, so the minimum number of generating signals Γ will be the same as the number of sources P. We need Γ to be as similar as possible to S. Ideally J would be the pseudo-inverse of H; however, we may not know the exact parameterization of H. In the real world spatial separation of sources from an output of a sensor array is achieved using beamforming techniques [11]:

$$\begin{bmatrix} X_1(z) \\ X_2(z) \\ \vdots \\ X_M(z) \end{bmatrix} = \begin{bmatrix} H_{11}(z) & \cdots & H_{1P}(z) \\ H_{21}(z) & \cdots & H_{2P}(z) \\ \vdots & \vdots & \vdots \\ H_{M1}(z) & \cdots & H_{MP}(z) \end{bmatrix} \begin{bmatrix} S_1(z) \\ S_2(z) \\ \vdots \\ S_P(z) \end{bmatrix}, \quad (1)$$

$$\mathbf{X} = \mathbf{HS},$$

$$\mathbf{\Gamma} = \mathbf{JHS}.$$

The fundamental idea of beamforming is that prior knowledge of the sensor and source geometry can be exploited in our favor. However, as we will see in Section 4 beamforming algorithms need localization and tracking of the sound sources in order to steer the array to the right position. Our solution (described in Section 5) employs a joint audio-video-based localization and tracking to avoid the inherent reverberation problems associated with acoustic-only source localization. The full block diagram of the system can be seen in Figure 5.

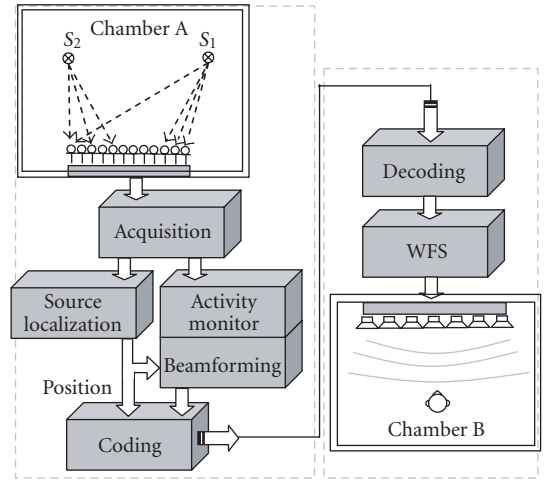

Figure 5: General architecture of the system.

The *acquisition* block receives the multichannel signals from the microphone array through a data acquisition (DAQ) board and captures digital audio samples to form multichannel audio streams.

The *activity monitor* basically consists in a vocal activity detector that readjusts to the noise level and stops the adaptation process when necessary to avoid the appearance of sound artifacts.

The *source localization* (SL) block uses both acoustical (steered response power-phase transform (SRP-PHAT)) and video (face tracking) algorithms to obtain a good estimation of the position of the source. This information is needed by the beamforming component and the WFS synthesis block.

The *beamforming* algorithm employs a robust generalized sidelobe canceller (RGSC) scheme. For the adaptive algorithms several alternatives have been tested including constrained-NLMS, frequency domain adaptive filters (xFDAF), and conjugate gradient (CG) algorithms to achieve a good compromise between computational complexity, convergence speed, and latency.

The *coding* block codifies the signal using two standard perceptual coders (MPEG2-AAC or G.722) to prove the compatibility between the estimation process and the use of standard codecs.

Finally, the acoustic field is rendered again in the receiving room using WFS techniques and a 10-loudspeaker array. Next sections give more details on the precise implementation of each of these blocks.

3. ACQUISITION

The acquisition block consists on a multichannel acquisition hardware (NI-4772 VXI board) and the corresponding software tool (NI-DAQ) responsible of retrieving the digital audio samples from the VXI boards. The acquisition tool has been implemented in Labview to facilitate the modification

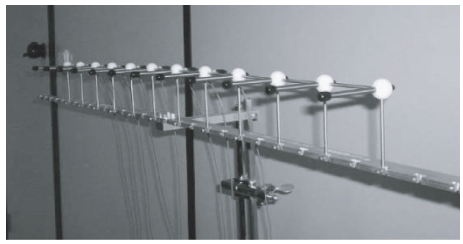

Figure 6: Microphone array.

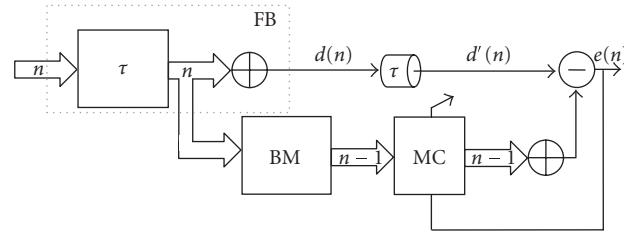

Figure 8: GSC block diagram.

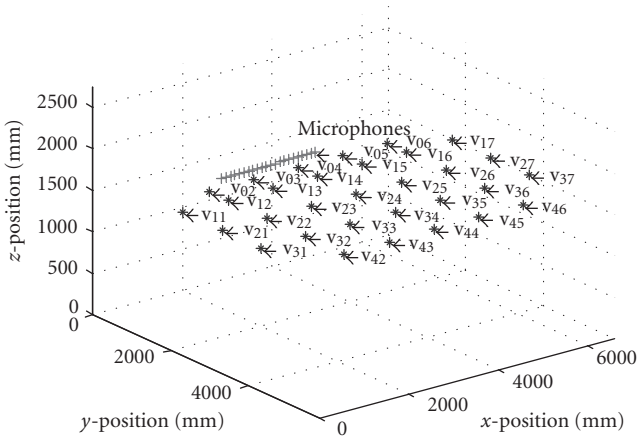

Figure 7: Bell labs chamber.

of several parameters such as sampling frequency and N^o points to capture. The microphone array (Figure 6) has 12 linearly placed (8 cm separation) PCB Piezotronics omnidirectional microphones (for our tests only eight were employed) with included preamplifiers. The test signals were recorded at midnight to avoid disturbing ambient sounds like the air conditioned system.

As the chamber used in our tests shows low reverberation (RT60 < 70 ms), to obtain the microphone signals we have also used some impulse response recordings of a varechoic chamber in Bell Labs [12] which offers higher reverberation values (RT60 = 380 ms). In that case the IRs were recorded from different audio locations (Figure 7) using a 22-linear omnidirectional microphone array (10 cm separation).

4. BEAMFORMING

4.1. Current beamforming alternatives

The spatial properties of microphone arrays can be used to improve or enhance the captured speech signal. Many adaptive beamforming methods have been proposed in the literature. Most of them are based on the linearly constrained minimum variance (LCMV) beamformer [11] which is often implemented using the generalized sidelobe canceller (GSC) developed by Griffiths and Jim [13]. The GSC (Figure 8) is based on three blocks: a fixed beamformer (FB) that enhances the desired signal using some kind of delay-and-sum strategy (and the direction of arrival (DOA) estimation provided by the SL block), the blocking matrix (BM) that blocks the desired signal and produces the noise/interference-only reference signal, and the multichannel canceller (MC) which tries to further improve the desired signal at the output of the FB using the reference provided by the BM.

The GSC scheme can obtain a high interference reduction with a small number of microphones arranged on a small space. However, it suffers from several drawbacks and a number of methods to improve the robustness of the GSC have been proposed over the last years to deal with the array imperfections.

Probably, the biggest concern with the GSC is related to its sensibility to steering errors and/or the effect of reverberation. Steering-vector errors often result in target signal leakage into the BM output. The blocking of the target signal becomes incomplete and the output suffers from target signal cancellation. A variety of techniques to reduce the impact of this problem has been proposed. In general, these systems receive the name of robust beamformers. Most approaches try to reduce the target signal leakage over the blocking matrix using different strategies. The alternatives include inserting multiple constraints in the BM to reject signals coming from several directions [14], restraining the coefficient growth in the MC to minimize the effect that eventual BM-leakage could cause [15], or using an adaptive BM [16] to enhance the blocking properties of the BM. Some recent strategies go even further, introducing a Wiener filter after the FB to try to obtain a better estimation [17]. Most implementations use some kind of voice activity detector [18] to stop the adaptation process when necessary and avoid the appearance of sound artifacts.

Apart from dealing with target signal cancellation, there are some other key elements to take into account for our application.

(i) Convergence speed. In a quick time varying environment, where small head movements of the speaker can change the response of the filter that we have to synthesize, the algorithm has to converge, necessarily, in a short period of time.

(ii) Computational complexity. The application is oriented towards building effective real-time communication systems so efficient use of computational resources has to be taken into account.

(iii) Latency: again, for building any communication system a low latency is highly desirable.

Table 1

	NLMS	FDAF	PBFDAF	CG
Processing time (s)	< 0.70	< 0.09	< 0.19	> 5 s
Latency (samples)	1	128	32	1

The convergence speed problem is related to the kind of algorithm employed in the adaptive filters. Originally, typical GSC schemes use some kind of LMS filters due to its low computational cost. This algorithm is very simple but it suffers from not-so-good convergence time, so some GSC implementations use affine projection algorithms (APA) [19], conjugate gradient techniques [20, 21], or wave domain adaptive filtering (WDAF) [22] which speed up the convergence time at the cost of increasing the computational complexity. This parameter can be reduced using subband approaches [23], with efficient complex valued arithmetic [24] or operating in the frequency domain (FDAF) [25, 26].

4.2. Beamformer design: RGSC with mPBFDAF for MC

Figure 10 shows our current implementation which uses the adaptive BM approach to reduce the target signal cancellation problem and a VAD to control the adaptation process. After considering several alternatives we decided to develop multichannel partitioned block frequency domain adaptive filters (mPBFDAF) [27] for the MC (as they show a good tradeoff between convergence speed, complexity, and latency) and a constrained version of a simple NLMS filter for the BM. Subband conjugate gradient algorithms [28] were also tested but, although they showed really good convergence speed, they were discarded due to the enormous computational power they needed (two orders of magnitude higher compared to FDAF implementations, see Table 1 and Figure 9).

4.2.1. mPBFDAF (multichannel canceller)

PBFDAF filters take advantage of working in the frequency domain greatly reducing the computational complexity. Moreover, the filter partitioning strategy reduces the overall latency of the algorithm making it very suitable for our interests.

Figure 11 shows the multichannel implementation of the PBFDAF filter that we have developed for using in the MC. Assuming a filter with a long impulse response $h(n)$, it can be sectioned in L adjacent, equal length, and non-overlapping sections as

$$h_k(n) = \sum_{l=0}^{L-1} h_{k,l}(n), \tag{2}$$

where $h_{k,l}(n) = h_k(n)$ for $n = lN, \ldots, lN + N - 1$, L the number of partitions, k the channel number ($k = 0, \ldots, M - 1$), and N the length of the partitioned filter. This can be seen as a bank of parallel filters working in the full spectrum of the input signal.

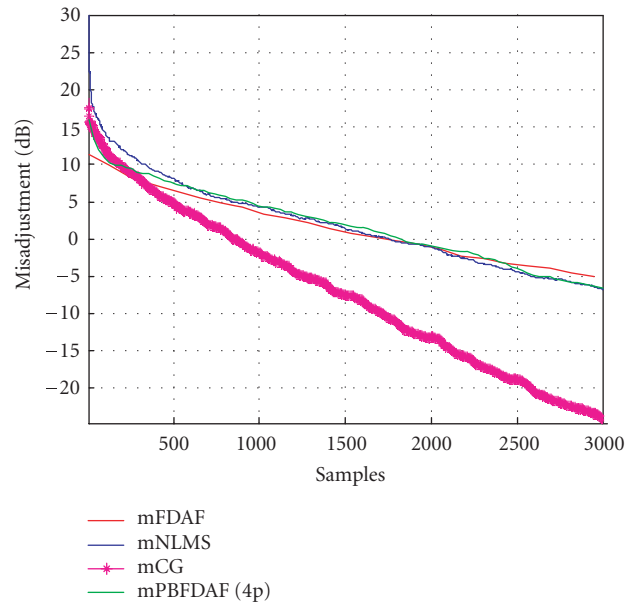

Figure 9: Convergence speed. System identification problem: 3 channels, 128 tap filters (PBFDAF using 4 partitions $L = 4, N = 32$).

The output, $y(n)$, can be obtained as the sum of L parallel N-tap filters with delayed inputs:

$$\begin{aligned} y_k(n) &= x_k(n) * \sum_{l=0}^{L-1} h_{k,l}(n) = \sum_{l=0}^{L-1} x_k(n) * h_{k,l}(n) \\ &= \sum_{l=0}^{L-1} x_k(n - lN) * h_{k,l}(n + lN) = \sum_{l=0}^{L-1} y_{k,l}(n). \end{aligned} \tag{3}$$

This way, using the appropriate data sectioning procedure the L linear convolutions (per channel) of the filter can be independently carried in the frequency domain with a total delay of N samples instead of the NL samples needed in standard FDAF implementations.

After a signal concatenation block (2 N-length blocks, necessary for avoiding undesired overlapping effects and to assure a mathematical equivalence with the time domain linear convolution), the signal is transformed into the frequency domain. The resulting frequency block is stacked in a FIFO memory at a rate of N samples. The final equivalent time output (with the contributions of every channel) is obtained as

$$y(n) = \text{IFFT}\left[\sum_{k=0}^{M-1}\sum_{l=0}^{L-1} \mathbf{X}_k^l(j - l)\mathbf{H}_k^l\right], \tag{4}$$

where "j" represents the time index. Notice that we have altered the order of the final sum and IFFT operations as

$$\begin{aligned} &\text{IFFT}\left[\sum_{k=0}^{M-1}\sum_{l=0}^{L-1} \mathbf{X}_k^l(j - l)\mathbf{H}_k^l\right] \\ &\quad = \sum_{k=0}^{M-1}\sum_{l=0}^{L-1} \text{IFFT}\left[\mathbf{X}_k^l(j - l)\mathbf{H}_k^l\right]. \end{aligned} \tag{5}$$

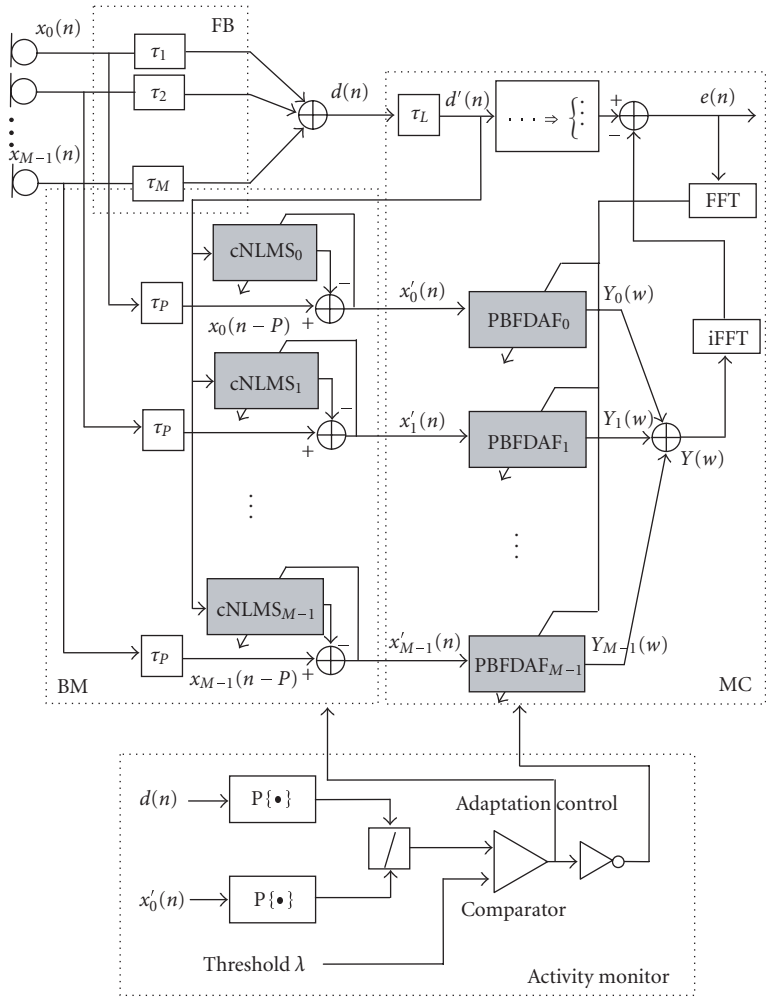

Figure 10: General diagram RGSC implementation.

This way, we save $(N-1)*(M-1)$ FFT operations in the complete filtering process.

As in any adaptive system the error can be defined as

$$e(n) = d(n) - y(n). \tag{6}$$

On the other hand, as the filtering operation is done in the frequency domain, the actualization of the filter coefficients is performed in every frequency bin ($i = 0, \ldots, 2*N-1$)

$$\begin{aligned} H^l_{k,i}(j+1) &= H^l_{k,i}(j) \\ &\quad + \mu^l_{k,i}(j)\,\mathrm{Prj}\left[E_i(j)\left[X_{k,i}(j-l+1)\right]^*\right], \end{aligned} \tag{7}$$

where E_i is the corresponding frequency bin, the asterisk denotes complex conjugation, and $\mu^l_{k,i}$ denotes the adaptation step. The "Prj" gradient projection operation is necessary for implementing the constrained version of the PBFDAF. This version adds two FFTs more (see Figure 11) to the computational burden but speeds up the convergence.

Finally, the adaptation step is computed using the spectral power information of the input signal:

$$\mu^l_{k,i}(j) = \frac{u}{\gamma + (L+1)P^i_k(j)}, \tag{8}$$

where u represents a fixed step size parameter, γ a constant to prevent the updating factor from getting too large, and P the power estimate of the ith frequency bin:

$$P^i_k(j) = \lambda P^i_k(j-1) + (1-\lambda)\left|X_{k,i}(j)\right|^2. \tag{9}$$

Being λ a small factor for the updating equation for the signal energy in the subbands.

4.2.2. cNLMS (blocking matrix)

For the BM filters, we are using a constrained version of a simple NLMS filter. BM filter length is usually below 32 taps so there was no real gain from using frequency domain adaptive algorithms like in the MC case. Each coefficient of the filter is constrained based on the fact that filter coefficients for target signal minimization vary significantly with the target DOA. This way we can restrict the allowable look-directions to avoid bad behavior due to a noticeable DOA error. The

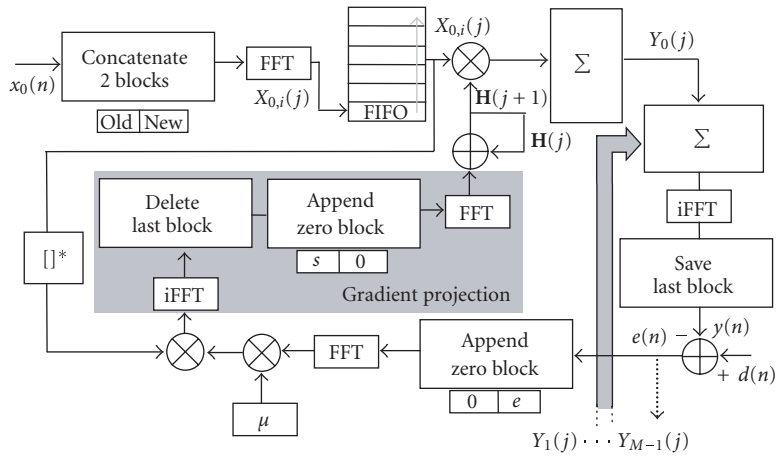

Figure 11: PBFDAF implementation.

adaptation process can be described as

$$h'_n(j+1) = h_n(j) + \mu \frac{x'_n(j)}{d(j)^T d(j)} d(j),$$

$$h_n(j+1) = \begin{cases} \phi_n & \text{for} \longrightarrow h'_n(j+1) > \phi_n, \\ \psi_n & \text{for} \longrightarrow h'_n(j+1) < \psi_n, \\ h'_n(j+1) & \text{otherwise}, \end{cases} \quad (10)$$

where ψ_n and ϕ_n represent the lower and upper vector bounds for coefficients.

4.2.3. Activity monitor

The activity monitor is based on the measure of the local power of the incoming signals and tries to detect the pauses of the target speech signal. The MC weightings are estimated only during pauses of the desired signal and the BM weightings during the rest of the time. Basically, the pause detection is based on the estimation of the target signal-to-interference ratio (SIR). We are using the approach presented in [29] where the power ratio between the FB output and one of the outputs of the BM is compared to a threshold.

4.3. Source separation evaluation results

The full RGSC algorithm has been implemented in Matlab and *C* and runs in real time (8 channels, Fs = 16 kHz, BM = 32 taps, MC = 256 taps) in a 3.2 GHz Pentium IV. The behavior of the adaptive algorithm was tested in a real environment.

Two signals (Fs = 16 kHz, 4 s excerpts) were placed in positions v21 (speech signal) and v27 (white noise) (see Figure 7) to see the performance of the algorithm in recovering the original dry speech signal.

Figure 12 shows the SNR gain of each algorithm once the convergence time is over. The RGSC uses 16 tap filters at BM and 128 or 256 at the MC (2 configurations). As expected the longer the filter at the MC is, the better the results are; at SNR (input) = 5 dB more than 20 dB of gain is achieved in contrast with the mere 9 dB gain with a standard fixed beamformer.

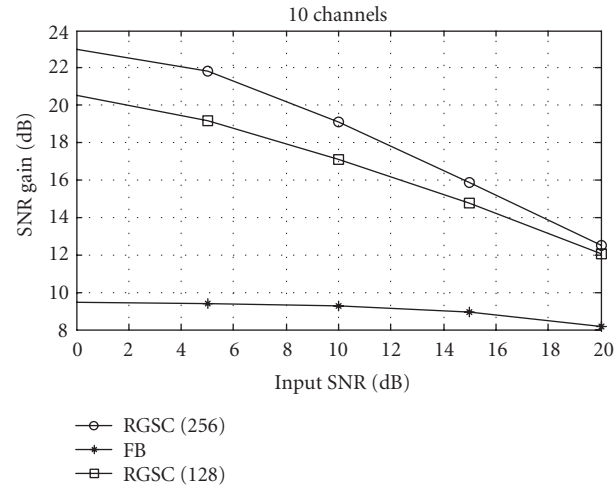

Figure 12: SNR gain versus input SNR using 10 microphones.

5. SOURCE LOCALIZATION

As mentioned in previous sections, source localization is necessary in the source separation process as well as in the sound field rendering process. From an acoustical point of view, there are three basic strategies when dealing with the source localization problem. Steered response power (SR) locators basically steer the array to various locations and search for a peak in the output power [30]. This method is highly dependant on the spectral content of the source signal; many implementations are based on a priori knowledge of the signals involved in the system making the scheme not very practical in real speech scenarios.

The second alternative is based on high resolution spectral estimation algorithms (such as MUSIC algorithm) [31]. Usually, these methods are not as computationally

demanding as the SR methods but tend to be less robust when working with wideband signals although some recent work has tried to address this issue [32].

Finally, time-difference-of-arrival- (TDOA-) based locators use time delay estimation (TDE) of the signals in different microphones usually employing some version of the generalized cross correlation (GCC) function [33]. This approach is computationally undemanding but suffers in high reverberant environments. This multipath channel distortion can be partially solved making the GCC function more robust using a phase transform (PHAT) [34] to deemphasize the frequency dependant weightings.

We have decided to use the SRP-PHAT method described in [35] that combines the inherent robustness of the steered response power approach with the benefits of working with PHAT transformed signals. The method is quite simple and starts with the computation of the generalized cross correlations between every microphone-pair signals:

$$R_{12}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \psi_{12}(\omega) X_1(w) X_2^*(w) e^{jw\tau} d\omega, \tag{11}$$

where $X_1(\omega)$ and $X_2(\omega)$ represent the signals in the microphones 1 and 2 and ψ_{12} the PHAT weighting defined by (12).

The PHAT function emphasizes the GCC function at the true DOA values over the undesirable local maximums and improves the accuracy of the method,

$$\psi_{12}(\omega) = \frac{1}{|X_1(w)X_2^*(w)|}. \tag{12}$$

After computing the GCC of each microphone pair, as in any steered response method, a search between potential source location starts. For every location under test, the theoretical delays of each microphone pair have been previously calculated. Using those delay values, for each position, the contribution of cross correlations is accumulated. The position with the highest score is chosen.

Figure 13 shows the method in action. Using the Bell chamber environment, a male speech (Fs = 16 kHz, 4 s excerpt, 8 microphones → 28 pairs) was placed in v46. Candidate positions were selected using a 0.01 m^2 resolution. Figures 13(a) and 13(b) (2D projection) show the result of running the SRP-PHAT algorithm (whiter → higher values, window → 512 taps ≈ 30 ms) where the "+" symbol marks the correct position and the "∗" the estimated one. As you can see, in these single speaker situations the DOA estimation is good but the problems arise when working in multiple source environments. In the test shown in Figure 13(c) a second (white noise) source was placed in v42 and the algorithm clearly had problems to identify the target source location. In those heavy competing noise situations acoustical methods (especially SRP-PHAT) suffer from high degradation.

To circumvent this problem we have used a second source of information: video-based source localization. Video-based source localization is not a new concept and has been extensively studied, especially in three-dimensional computer vision [36]. Recently, we have seen an effort to mix the audio and video information for building robust location systems in low SNR environments. Those systems relay on Kalman filtering [37] or Bayesian networks [38] for effective data fusion. We propose a very simple approach where video localization is used as a first rough estimation that basically discards nonsuitable positions. The remaining potential locations are tested using the SRP-PHAT algorithm in what we could call a visually guided acoustical source localization system. This position-pruning scheme is, most of the time, enough for rejecting problematic second source situations. Besides, the computational complexity associated to video signal processing is somehow compensated with a smaller search space for the SRP-PHAT algorithm.

Our video source location system is a real-time face tracker using detection of skin-color regions based on the machine perception toolbox (MPT) [39]. A sample result of face detection can be seen in Figure 14.

6. CODING/DECODING

After the estimation process, the signal must be codified prior to be sent. We have tested two different codification schemes, MPEG2-AAC (commonly used for wideband audio) and G-722 (very used in teleconference scenarios), to see if the estimation process has any impact in the behavior of these algorithms. Luckily, in the informal subjective test comparing the original estimated signal (the same work situation as in Section 4) with the coded/decoded signal (Figure 15), the listeners were unable to distinguish between both situations neither when using AAC (64 kbps/channel) nor when working with G.722 (64 kbps/channel).

7. WAVE FIELD SYNTHESIS

The last process involves rebuilding the acoustic field again at reception. The sound field rendering process is based on well-known WFS techniques. We are using a 10-loudspeaker array situated in a different chamber than the ones used for signal capturing. The synthesis algorithm is based on [40], although no room compensation was applied. Derivation of the driving signals for a line of loudspeakers is found in [41] and can be summarised with the expression:

$$Q(r_n, \omega) = S(\omega)\frac{\cos\theta_n}{G(\phi_n)}\sqrt{\frac{jk}{2\pi}}\sqrt{\frac{1}{2}}\frac{e^{-jkr_n}}{\sqrt{r_n}}, \tag{13}$$

where $Q(r_n, \omega)$ is the driving signal of the loudspeaker, $S(\omega)$ the virtual estimated source, θ_n the angle between the virtual source and the main axis of the nth loudspeaker, and $G(\phi_n, \omega)$ the directivity index of the virtual source (omnidirectional in our tests). Also notice that no special method was applied to override the maximum spatial aliasing frequency problem (around 1 kHz). However, it seems [42] that the human auditory system is not so sensitive to these aliasing artifacts.

8. SUBJECTIVE EVALUATION

The evaluation of the system is, certainly, not an easy task. Our aim was to prove that the system was able to significantly reduce the noise at the same time that the spatial properties were maintained. For that purpose, subjective

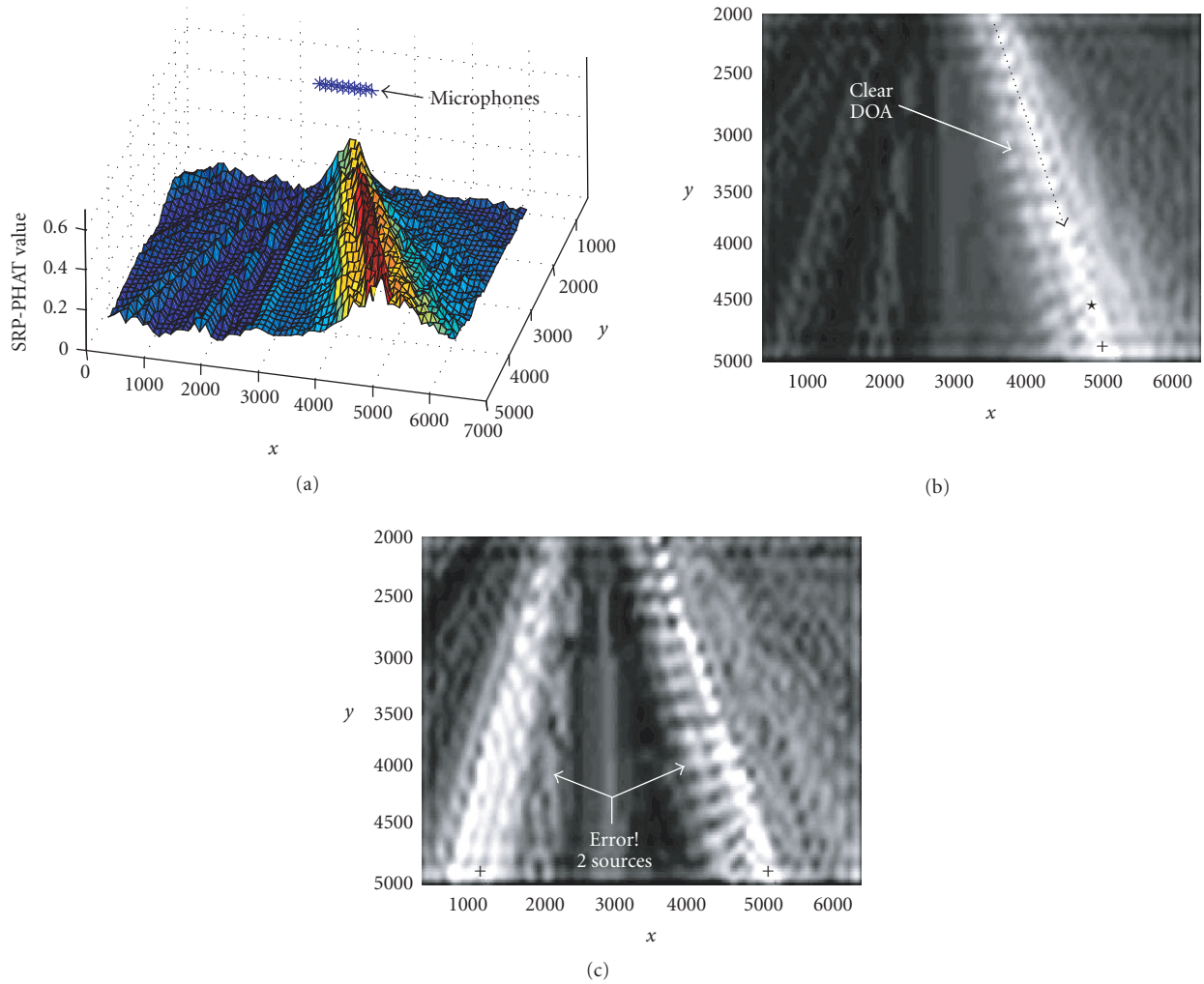

Figure 13: Source localization using SRP-PHAT. (a) Single source, (b) single source (2D projection), and (c) multiple sources.

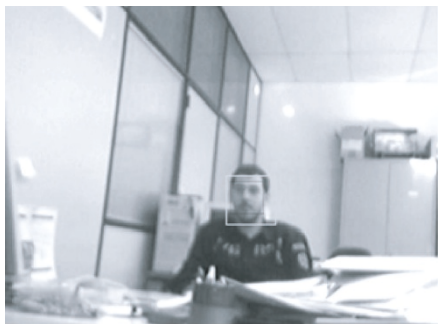

Figure 14: Face tracking.

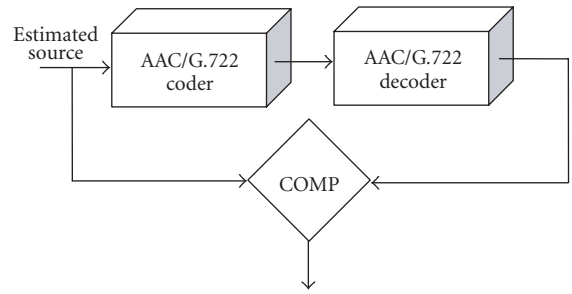

Figure 15: Comparison: estimated signal versus coded/decoded signal.

MOS experiments have been carried out to see how well the system performed. Two signals, speech in v21 and white noise in v27 (SNR_{in} = 5 dB), were recorded by the microphone array in the emitting room. After the beamforming process the estimated signal was used to render again the acoustic field at the receiving room. The subjective test is based on a slightly modified version of the MUSHRA standard [43]. This standard was originally designed to build a less sensitive but still reliable implementation of the BS.1116 recommendation [44] used to evaluate most high quality

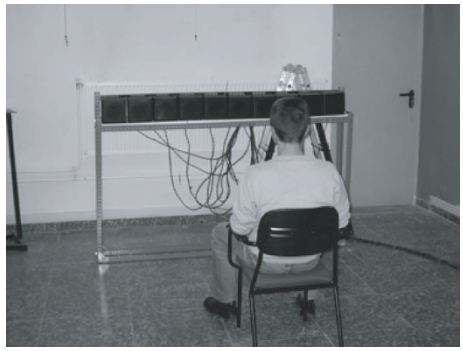

Figure 16: Loudspeaker array.

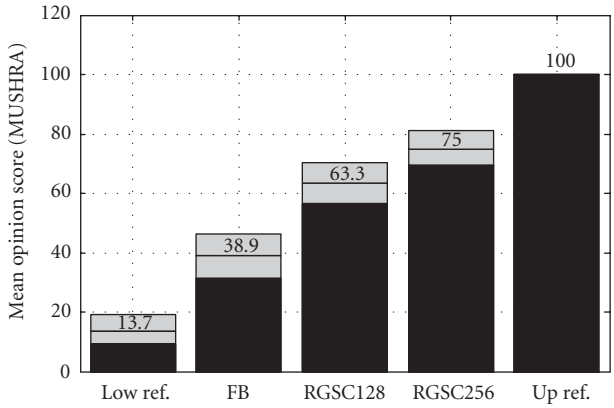

Figure 17: Mean opinion score (MUSHRA test) after WFS.

codification schemes. Fifteen listeners took part in the test; Figure 16 shows the relative position of the subjects to the array (centred position distance: 1.5 m).

In this kind of tests, the listener is presented with all different processed versions of the test item at the same time. This allows the subject to easily change between different versions of the test item and to come to a decision about the relative quality of the different versions. The original, unprocessed version (identified as the reference version) of the test item is always available to the subject to give him the idea how the item should really sound. In our case, the reference version was the sound field recreated (via WFS) using the original dry signal (as if all the noise had disappeared and the estimation of the source was perfect). This version is also presented to the subject as a hidden upper reference to ensure that the top of the scale is used. On the other side, to ensure that the low part of the scale is used, the standard proposes to employ a 3.5 kHz filtered version of the original reference which is not applicable to our situation as it lacks from the effect of the ambient noise. In our case we decided to use the sound field rendered using the sound captured by the central microphone of the array (without any noise reduction). We refer to this version as the hidden lower reference. Using both hidden anchors, we ensure that the full range of the scale is used and the system obtains more realistic values.

The subjects are required to assign grades giving their opinion of the quality under test and the hidden anchors. In our case, the subjects were instructed to pay special attention not only to overall quality, intelligibility, signal cancellation, or sound artifact appearance but they were also asked to concentrate on any displacements of the localization of the source. Any source movement should obtain a low score. The scale is numerical and goes from 100 to 0 (100–80: excellent, 80–60 good, 60–40 fair, 40–20 poor, 20–0 bad). Subjects were instructed to score 30 audio excerpts (6 different sentences, 5 situations per sentence: hidden upper reference, RGSC (256 taps in the MC), RGSC (128), fixed beamformer, hidden lower reference). The original dry sentences were selected from the Albayzin speech database [45] (Fs = 16 kHz, Spanish language). As the way the instructions are given to the listeners can significantly affect the way a subject performs the test, all the listeners were instructed the same way (using a 2-page documentation).

The results are shown in Figure 17 where the number on each bar represents the mean score obtained by each method and the vertical hatched box indicates a 95% confidence interval. Nearly all the listeners were able to describe the desired source coming from the right position and almost none of them described any target signal cancellation or the appearance of disturbing sound artifacts.

9. CONCLUSIONS AND FUTURE WORK

In this paper we have seen some of the challenges that future immersive audio applications have to deal with. We have presented a range of solutions that behave quite well in nearly every area. Partitioned block frequency domain-based robust adaptive beamforming significantly enhances the speech signals at the same time that keeps low computational requirements allowing a real time implementation.

On the other side, visually guided acoustical source localization is capable of dealing with not-so-low reverberation chambers and multiple source situations and provides with good localization estimations both the beamforming block and the WFS block. The WFS-based rendered acoustical field shows good spatial properties as the MUSHRA-based subjective tests have assessed. However, there is margin for improvement in many areas.

When facing a two (or more) competing talker situations the activity monitor would need a more robust implementation to be able to detect speech-over-speech situations to effectively prevent the adaptive filtering to diverge. Joint audio-video source localization works quite well, especially obtaining DOA estimations which are enough for the beamforming FB block. However, the WFS block needs to know the distance to the source as well as the angle and the system suffers in some situations. Using better data fusion algorithms between audio and video information could, certainly, alleviate this problem. In the same line, the ability of the face tracking algorithm of detecting and following more than one person in the room should be another interesting feature. Finally, we are also exploring the possibility of introducing some kind of room compensation strategies (following the works in [46]) before the WFS block to achieve a better control over the listening area and reduce the acoustical impairments between the emitting and receiving rooms.

ACKNOWLEDGMENTS

This work was supported by project PCT-350100-04 and by Spanish Science and Technology Department through projects TIC 2003-09061-c03-01 and "Ramon y Cajal." The authors would also like to thank Mariano García for his valuable comments.

REFERENCES

[1] A. Blumlein, "Improvements in and relating to sound transmission, sound-recording and sound reproduction systems," patent no. 394325 December 1931.

[2] W. B. Snow, "Basic principle of stereophonic sound," *Journal of SMPTE*, vol. 61, pp. 567–589, 1953.

[3] W. B. Snow, "Auditory perspective," *Bell Laboratories Record*, vol. 12, pp. 194–198, 1934.

[4] A. Härmä, "Coding principles for virtual acoustic openings," in *Proccedings of the Audio Engineering Society 22nd Conference on Virtual, Synthetic and Entertainment Audio (AES22 '02)*, pp. 159–165, Espoo, Finland, June 2002.

[5] S. Torres, J. A. Beracoechea, I. Pérez-García, et al., "Coding strategies and quality measure for multichannel audio," in *Proceedings of the 116th Audio Engineering Society Convention*, Berlin, Germany, May 2004.

[6] H. Teutsch, S. Spors, W. Herbordt, W. Kellermann, and R. Rabesnstein, "An integrated real-time system for immersive audio aplications," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '03)*, New Paltz, NY, USA, October 2003.

[7] W. Kellermann, "Acoustic signal processing for next generation human/machine interfaces," in *Proceedings of the 8th International Conference on Digital Audio Effects (DAFx '05)*, Madrid, Spain, September 2005.

[8] A. J. Berkhout, "Holographic approach to acoustic control," *Journal of the Audio Engineering Society*, vol. 36, no. 12, pp. 977–995, 1988.

[9] M. M. Boone and W. P. J. Bruijn, "Improving speech intelligibility in teleconferencing by using Wave Field Synthesis," in *Proceedings of the 114th Audio Engineering Society Convention*, Amsterdam, The Netherlands, March 2003.

[10] W. P. J. Bruijn and M. M. Boone, "Application of Wave Field Synthesis in life-size videoconferencing," in *Proceedings of the 114th Audio Engineering Society Convention*, Amsterdam, The Netherlands, March 2003.

[11] B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[12] Bell Labs's Varecoic Chamber http://www.bell-labs.com/org/1133/Research/Acoustics/VarechoicChamber.html.

[13] L. J. Griffiths and C. W. Jim, "Alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, 1982.

[14] B. Widrow and J. M. McCool, "Comparison of adaptive algorithms based on the methods of steepest descent and random search," *IEEE Transactions on Antennas and Propagation*, vol. 24, no. 5, pp. 615–637, 1976.

[15] Y. Liu, Q. Zou, and Z. Lin, "Generalized sidelobe cancellers with leakage constraints," in *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS '05)*, Kobe, Japan, May 2005.

[16] O. Hoshuyama, A. Sugiyama, and A. Hirano, "Robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Transactions on Signal Processing*, vol. 47, no. 10, pp. 2677–2684, 1999.

[17] A. Abad and J. Hernando, "Integrated adaptive beamforming and Wiener filtering for a robust microphone array," in *IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM '04)*, pp. 367–371, Barcelona, Spain, July 2004.

[18] Z. M. Saric and S. T. Jovicic, "Adaptive microphone array based on pause detection," *Acoustic Research Letters Online*, vol. 5, no. 2, pp. 68–74, 2004.

[19] Y. Zheng and R. Goubran, "Adaptive beamforming using Affine Projection Algorithms," in *Proceedings of 5th International Conference on Signal Processing (ICSP '00)*, Beijing, China, August 2000.

[20] J. A. Apolinário Jr., M. L. R. De Campos, and C. P. O. Bernal, "Constrained conjugate gradient algorithm," *IEEE Signal Processing Letters*, vol. 7, no. 12, pp. 351–354, 2000.

[21] J. A. Beracoechea, S. Torres, L. García, et al., "Source separation for microphone arrays using conjugate gradient techniques," in *Proceedings of the 8th International Conference on Digital Audio Effects (DAFx '05)*, Madrid, Spain, September 2005.

[22] H. Buchner, S. Spors, and W. Kellermann, "Wave-domain adaptive filtering: acoustic echo cancellation for full-duplex systems based on wave-field synthesis," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '04)*, vol. 4, pp. 117–120, Montreal, Quebec, Canada, May 2004.

[23] S. Low, S. Nordholm, and N. Grbic, "Subband generalized Sidelobe approach—a constrained region approach," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '03)*, New Paltz, NY, USA, October 2003.

[24] G. Glentis, "Implementation of adaptive generalized sidelobe cancellers using complex valued arithmetic," *International Journal of Applied Mathematics and Computer Science*, vol. 13, no. 4, pp. 549–566, 2003.

[25] W. Herbordt and W. Kellermann, "Efficient frequency-domain realization of robust generalized sidelobe cancellers," in *Proceedings of IEEE 4th Workshop on Multimedia Signal Processing*, pp. 377–382, Cannes, France, October 2001.

[26] Z. L. Yu and M. H. Er, "An extended generalized sidelobe canceller in time and frequency domain," in *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS '05)*, vol. 3, pp. 629–632, Vancouver, BC, Canada, May 2004.

[27] J. M. Páez Borrallo and M. García Otero, "On the implementation of a partitioned block frequency domain adaptive filter (PBFDAF) for long acoustic echo cancellation," *Signal Processing*, vol. 27, no. 3, pp. 301–315, 1992.

[28] L. García, S. Torres, J. A. Beracoechea, et al., "Conjugate Gradient techniques for Multichannel acoustic echo cancellation," in *Proceedings of the 8th International Conference on Digital Audio Effects (DAFx '05)*, Madrid, Spain, September 2005.

[29] O. Hoshuyama, B. Begasse, A. Sugiyama, and A. Hirano, "Realtime robust adaptive microphone array controlled by an SNR estimate," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '98)*, vol. 6, pp. 3605–3608, Seattler, Wash, USA, May 1998.

[30] N. Strobel, T. Meier, and R. Rabenstein, "Speaker localization using a steered filter-and-sum beamformer," in *Erlangen Workshop '99: Vision, Modeling and Visualization*, Erlangen, Germany, November 1999.

[31] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, Englewood Cliffs, NJ, USA, 1991.

[32] H. Teutsch and W. Kellermann, "EB-ESPRIT: 2D localization of multiple wideband acoustic sources using eigen-beams,"

in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, Philadelphia, Pa, USA, March 2005.

[33] C. H. Knapp and G. C. Carter, "Generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[34] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '97)*, vol. 1, pp. 187–190, Munich, Germany, April 1997.

[35] J. DiBiase, H. Silverman, and M. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays; Signal Processing Techniques and Applications*, pp. 157–180, Springer, Berlin, Germany, 2001.

[36] O. Faugeras, *Three-Dimensional Computer Vision. A Geometric Viewpoint*, MIT press, Cambridge, Mass, USA, 1993.

[37] N. Strobel, S. Spors, and R. Rabenstein, "Joint audio-video signal processing for object localization and tracking," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds., pp. 197–219, Springer, Berlin, Germany, 2001.

[38] F. Asano, K. Yamamoto, I. Hara, et al., "Detection and separation of speech event using audio and video information fusion and its application to robust speech interface," *EURASIP Journal on Applied Signal Processing*, vol. 2004, no. 11, pp. 1727–1738, 2004.

[39] I. Fasel, B. Fortenberry, and J. Movellan, "A generative framework for real-time object detection and classification," *Computer Vision and Image Understanding*, vol. 98, pp. 182–210, 2005.

[40] S. Bleda, J. J. López, and B. Pueo, "Software for the simulation, performance analysis and real time implementation of Wave Field Synthesis systems for 3D Audio," in *Proceedings the 6th International Conference on Digital Audio Effects (DAFx '03)*, London, UK, September 2003.

[41] D. De Vries, "Sound reinforcement by wavefield synthesis: adaptation of the synthesis operator to the loudspeaker directivity characteristics," *Journal of the Audio Engineering Society*, vol. 44, no. 12, pp. 1120–1131, 1996.

[42] M. M. Boone, "Acoustic rendering with wave field systhesis," in *Proceedings of the ACM Siggraph and Eurographics Campfire on Acoustic Rendering for Virtual Environments*, pp. 37–45, Snowbird, Utah, USA, May 2001.

[43] MUSHRA (MUlti Stimulus test with Hidden Reference and Anchor, ITU-R BS.1534).

[44] ITU-R BS.1116-1, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems".

[45] Albayzin. Spanish Speech Database. Universidad Politécnica de Cataluña. Proyecto TIC91-1488-C06.

[46] S. Spors, H. Buchner, and R. Rabenstein, "A novel approach to active listening room compensation for wave field synthesis using wave-domain adaptive filtering," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '04)*, vol. 4, pp. 29–32, Montreal, Quebec, Canada, May 2004.

J. A. Beracoechea received the Telecom Engineer degree from Universidad Europea de Madrid (UEM) in 2001. He is currently working towards the Ph.D. degree at the Signal Processing Group at the Universidad Politécnica de Madrid (UPM). His research interests include multichannel audio coding, microphone and loudspeaker arrays, beamforming, and source tracking with particular emphasis on the application of the virtual acoustic opening for creating immersive audio systems.

S. Torres-Guijarro received the M.Eng. and Ph.D. degrees in telecommunication engineering from the Universidad Politécnica de Madrid, Spain, in 1992 and 1996, respectively. Dr. Torres worked as a teacher at the Universidad de Valladolid, Universidad Carlos III de Madrid, and Universidad Europea de Madrid. Since 2002 she has been working as a Researcher of the Ramón y Cajal program at the Universidad Politécnica de Madrid, first, and at the Universidad de Vigo, at the moment. Her main research interest includes digital signal processing applied to speech, audio, and acoustics.

L. García received the Automatic Control Engineer degree from Instituto Superior Politécnico JAE., Havana, Cuba, in 1985, the Signals, System, and Radiocommunication Masters and the Ph.D. degrees in technologies and systems of communications from Universidad Politécnica de Madrid, Spain, in 1994 and 2006, respectively. From 1986 to 1990 he was a Solution Developer in the Company of Development of Automated Systems Direction of Cuban Industry Ministry. From 1991 to 1993 he was a Researcher and Professor of multimedia in Superior Art Institute of Havana, Cuba. From 1995 to 1997 he was a Researcher in the Spanish Council for Scientific Research. From 1999 to 2002 he was a Professor in the Universidad Pontificia Comillas of Madrid, Spain. Since 2003 he is a Professor in Universidad Europea de Madrid, Spain. His technical interests are in the areas of signal processing and artificial intelligence.

F. J. Casajús-Quirós received the M.Eng. and Ph.D. degrees in telecommunication engineering from the Universidad Politécnica de Madrid (Technical University of Madrid), Spain, in 1982 and 1988, respectively. He has been an Associate Professor in that University since 1989, where he is Vice-Head of the Signals, System, and Radiocommunications Department. His main research interests are in digital signal processing applied to wideband wireless communications and multimedia. His current research work includes the theory of multi-input multi-output systems as applied to multichannel audio signal processing and wireless communications. In those fields he has authored and coauthored more than 120 publications in journals and conference proceedings.