face, the user might find a helpful and information-using interface to be more intelligent. Therefore, "intelligence" does not actually mean cognition in this context, but it means using information in an appropriate manner [7, 8]. Interfaces can be intelligent about the user. The interface can also be sensitive to requirements and needs of the user. This ties closely with the user-based model, but it deals more with the interface adaptability than with the outright use of models. The *Association for Computing Machinery* defines *human-machine interaction* as "a discipline concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them" [9].

In this context, an IAI plays an important role for human-machine interaction since it translates acoustic information from user to computer, and *vice versa*, in order to allow a homogeneous interaction between parties. Through the use of a user-based model, an IAI can tailor audio services (both in acquisition and reproduction) to the user. From the user point of view, an IAI should be as invisible and intuitive as possible: working with and understanding an IAI should not be a task so that the user should be able to concentrate on the task which he is going to perform. In many cases, an IAI must learn user behaviour, mood and personality in order to yield an answer being as compliant as possible to user needs. Moreover, an IAI could also exploit feedbacks from user to improve its processing [10].

2.2. Immersive Audio Applications Exploiting IAIs

IAIs can be widely used in several fields of application, most of which focused on speech and acoustic audio processing [1]. It is possible to think to applications such as: speech/audio real-time interaction, automatic speech analysis, automatic music composition and transcription, automatic genre and context recognition in broadcast programs, high-interactivity entertainment, development of "intelligent rooms" in which both speakers and speech commands must be recognized and the perception of acoustic impulse responses can be controlled.

Immersive audio offers great opportunities for acoustic and speech signal processing and implies the use of IAIs. An IAI for immersive audio aims at extracting useful informations for computational or human purpose, such as analysis or synthesis of audio signals. This feature is also known as *machine listening*. At the same time, an IAI has to reproduce desired acoustic information taking into account that the listener would hear the sound exactly as in the original sound field. This feature indeed is known as *spatial sound reproduction*. To these ends, an IAI needs to replicate four attributes of face-to-face information exchange [11, 2]: full-duplex exchange, freedom of movement without tethered microphones, high-quality speech signals captured from a distance, spatial realism of sound rendering. These requirements imply that multiple microphones and loudspeakers would be used and the entire audio infrastructure might need to be renovated.

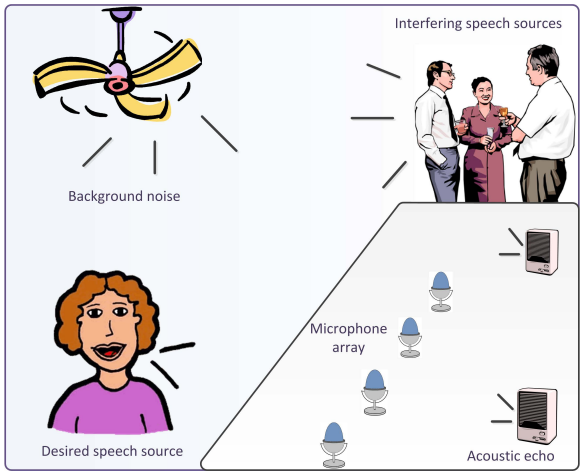

Fig. 2: Immersive speech communications in multisource environment.

In this work we mainly focus on immersive audio from the machine listening point of view, but the same problems can be addressed in the immersive audio reproduction. In this context, immersive audio services take often place in multisource environments where interfering signals may degrade quality and intelligibility of the desired acoustic source. Therefore, acquisition of desired signals with high quality is far more difficult and challenging for immersive audio than in the classical case where the microphone is close to the user. One of the main significant immersive audio services is the *hands-free speech communications*, where multiple parties may be involved and the microphone array is not neces-

$$= \begin{bmatrix} x_p[n] & \cdots & x_p[n-M+1] \\ x_p[n-1] & \cdots & x_p[n-M-2] \\ \vdots & \ddots & \vdots \\ x_p[n-K_j+1] & \cdots & x_p[n-M-K_j-2] \end{bmatrix}$$

where K_j represents the projection order for all the filters of the j-th MISO system. We denote the coefficient vector of the p-th filter belonging to the j-th MISO system at n-th time instant as:

$$\mathbf{w}_{n,p}^{(j)} \in \mathbb{R}^M = \begin{bmatrix} w_p^{(j)}[n] & w_p^{(j)}[n] & \dots & w_p^{(j)}[n] \end{bmatrix}^T. \tag{2}$$

All the filters of each MISO system, represented by (2), contain the same number M of coefficients and are adapted according to the *affine projection algorithm* (APA) [16], which was already used in adaptive beamforming [17]. The updating rule of the APA is:

$$\begin{aligned} \mathbf{w}_{n,p}^{(j)} =& \mathbf{w}_{n-1,p}^{(j)} + \mu_p^{(j)}[n]\, \mathbf{X}_{n,p}^{(j),T} \\ & \cdot \left(\delta_j \mathbf{I} + \mathbf{X}_{n,p}^{(j)} \mathbf{X}_{n,p}^{(j),T}\right)^{-1} \mathbf{e}_n^{(j)} \end{aligned} \tag{3}$$

where $\mathbf{e}_n^{(j)} \in \mathbb{R}^{K_j}$ is the error vector of the j-th MISO system containing the last K_j samples of the j-th error signal, which results from:

$$\mathbf{e}_n^{(j)} = \mathbf{d}_n^{(j)} - \sum_{p=0}^{P-1} \mathbf{y}_{n,p}^{(j)}. \tag{4}$$

In (4), $\mathbf{d}_n^{(j)} \in \mathbb{R}^{K_j} = \begin{bmatrix} d[n] & d[n-1] & \dots & d[n-K_j+1] \end{bmatrix}^T$ is the vector containing the last K_j samples of the desired signal and $\mathbf{y}_{n,p}^{(j)} \in \mathbb{R}^{K_j} = \mathbf{X}_{n,p}^{(j)} \mathbf{w}_{n-1,p}^{(j)}$ is the vector containing the K_j projections of the output signal relative to the p-th filter of the j-th MISO system. The parameter δ_j in (3) is the *regularization factor*, which is the same for all the filters of the j-th MISO system. Also in (3), the parameter $\mu_p^{(j)}[n]$ represents the *variable step size* (VSS) related to the p-th filter of the j-th MISO system. The use of a VSS allows to achieve a good trade-off between convergence rate at transient state, i.e. a larger value of the step size, and performance precision at steady state, i.e. a smaller value of the step size. The VSS is computed at each iteration and it derives from a minimization process of the mean square deviation [18]:

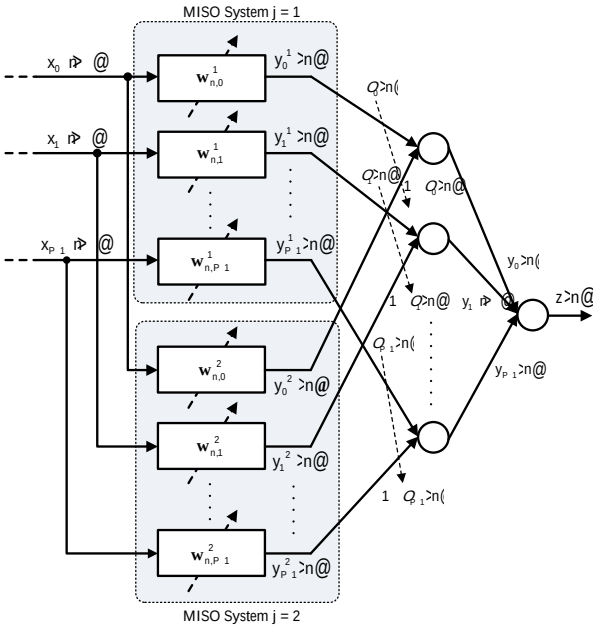

Fig. 4: Combined adaptive noise canceller scheme.

$$\mu_j^{(p)}[n] = \left| 1 - \frac{\sqrt{\left|\widehat{\sigma}_{x_p}^2[n] - \widehat{\sigma}_y^2[n]\right|}}{\widehat{\sigma}_e^2 + \zeta} \right| \tag{5}$$

where ζ is a small positive value avoiding divisions by zero. The generic parameter $\widehat{\sigma}_\alpha^2[n]$, with $\alpha = \{x_p, y, e\}$, represents the power estimate of the corresponding generic sequence $\alpha_p[n]$, and it can be computed as:

$$\widehat{\sigma}_\alpha^2[n] = \beta \widehat{\sigma}_\alpha^2[n-1] + (1-\beta)\, \alpha_p^2[n] \tag{6}$$

where β is a smoothing factor.

Using the update equation (2) it is possible to differentiate the two MISO systems simply by choosing two different values for the projection order K_j. In particular, we set $K_1 = 1$ and $K_2 = 4$ since it has

been proved that combining a first-order gradient-based filter (that can be achieved by setting $K_j = 1$) and a second-order Hessian-based filter ($K_j > 1$), an improvement of the tracking performance in nonstationary conditions can be obtained [19].

The filter outputs $y_p^{(j)}[n]$ of each MISO system are adaptively combined according to a *filter-by-filter* combination scheme [4]. In particular, the p-th filter output of the first MISO system is convexly combined with the correspondent p-th filter output of the second MISO system, thus generating $P-1$ outputs, each relative to a noise reference (see Fig. 4):

$$y_p[n] = \lambda_p[n]\, y_p^{(1)}[n] + (1 - \lambda_p[n])\, y_p^{(2)}[n] \qquad (7)$$

where $\lambda_p[n]$ is the p-th *mixing parameter*, adapted by using the p-th auxiliary parameter, $a_p[n]$, which is related $\lambda_p[n]$ by means of the following sigmoid function:

$$\lambda_p[n] = \frac{1}{1 + e^{-a_p[n]}} \qquad (8)$$

which aims at constraining the values of $\lambda_p[n]$ in the range $[0, 1]$ [15]. The updating rule of $a_p[n]$ can be written as [20]:

$$a_p[n+1] = a_p[n] + \frac{\mu_{\mathrm{a}}}{r_p[n]} e[n] \left(y_p^{(1)}[n] - y_p^{(2)}[n]\right) \cdot \lambda_p[n]\,(1 - \lambda_p[n]) \qquad (9)$$

where $\mu_{\mathrm{a}}/r_p[n]$ is a normalized step size [20] for the adaptation of mixing parameters which may be chosen as the same for all combinations.

Once computed the adaptive combinations between filters, it is possible to achieve the CANC output signal $z[n]$ by summing the individual output contributions deriving from the combinations (see Fig. 4):

$$z[n] = \sum_{p=0}^{P-1} y_p[n]. \qquad (10)$$

From equation (10) we derive the overall beamformer output signal $e[n] = d[n] - z[n]$.

4. EXPERIMENTAL RESULTS

In this section we prove the effectiveness of the proposed IAI for an immersive acoustic environment.

4.1. Experimental Set-up

The application context is that of an immersive speech teleconference, in which a user speaks and an IAI must acquire the desired speech and send it enhanced to a far-end user. In this environment the IAI can be located far from the desired source. Moreover, several sources may be present and may interfere with the desired source, thus degrading the quality of the desired information. The goal is that of reducing interfering sounds in order to send only the desired information (in high quality) to the far-end user. It is considered only the environment of one end of the communication, but the same IAI can be also applied into the far-end environment.

Experiments are conducted in a large room ($7 \times 5 \times 4$ m about) fit out with varied furniture, such as tables, wall curtains, and others, in order to recreate a typical work environment. The experimental set-up is composed of a desired speech source located in the near-field of an IAI, two interfering sources located sideways and both facing the IAI, and also a noisy source in the far-field which reproduces environment noise in order to recreate a realistic scenario. The desired source and the far-field noisy source are represented by professional loudspeakers, while the two interfering sources are two people which are free to move within the room. The IAI is composed of a microphone array consisting of 8 microphones located 4 cm far each other.

Regarding the audio reproduction equipment, both the desired source signal and the background noise signal are sent out each by an active loudspeaker Event Tuned Reference 6 (TR6). On the other hand, for the acquisition equipment 8 omnidirectional condenser microphones AKG C563CM are used, which are connected to a preamp Behringer ADA8000 Ultragain Pro-8 Digital interfaced with an RME MADI ADI-648.

Reproduced desired speech signals are chosen from the CLIPS database, released in 2004. In particular, the used signals were realized in an anechoic room from professional speakers. Pink noise is used as background noise and it is chosen from NOISE-ROM-0 database. The power of the output signals

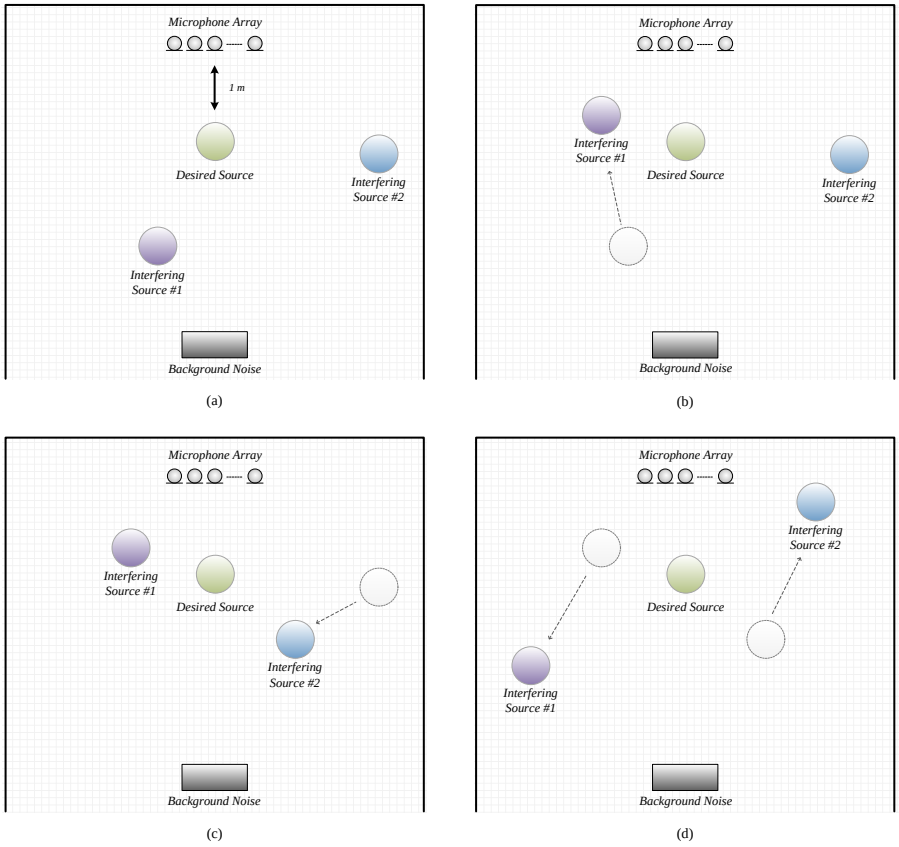

Fig. 5: Second configuration. (a) 0-5 seconds: initial positions (same as first configuration); (b) 5-10 seconds: position change of the first interfering source; (c) 10-15 seconds: position change of the second interfering source; (d) 15-20 seconds: simultaneous position change of both the interfering sources.

and the microphone array sensitivity are properly tuned.

Using the described experimental set-up, we consider two different scenarios. In the *first configuration* all the sources keep the same position for the entire length of the experiment. In particular, the desired source is located in front of the microphone array at a distance of 1 m from the center of the microphone array. The two interfering sources are placed respectively at $1,9$ m and $2,8$ m from the center of the array: both are male speakers, located respectively one on the right and one on the left of the desired source, as it is possible to see in Fig. 5 (a). In the *second configuration* the desired source and the noisy source do not go through any position change, while the two interfering sources do. Indeed, both the interfering sources change their position for three times during the experiment. In the first 5 seconds the position of all sources are the same of the first configuration (see Fig. 5 (a)). At second 5 the first position change occurs, as depicted in Fig. 5 (b): the interfering source #1 moves from position #1 to to position #2, while the interfering source #2 remains in its position #1. At second 10 the interfering source #2 moves in position #2 while source #1 holds steady (see Fig. 5 (c)). Finally, at second 15 both the interfering sources simultaneously move in their respective position #3, as represented in Fig. 5 (d).

In both the configurations the overall length of the experiment is 20 seconds.

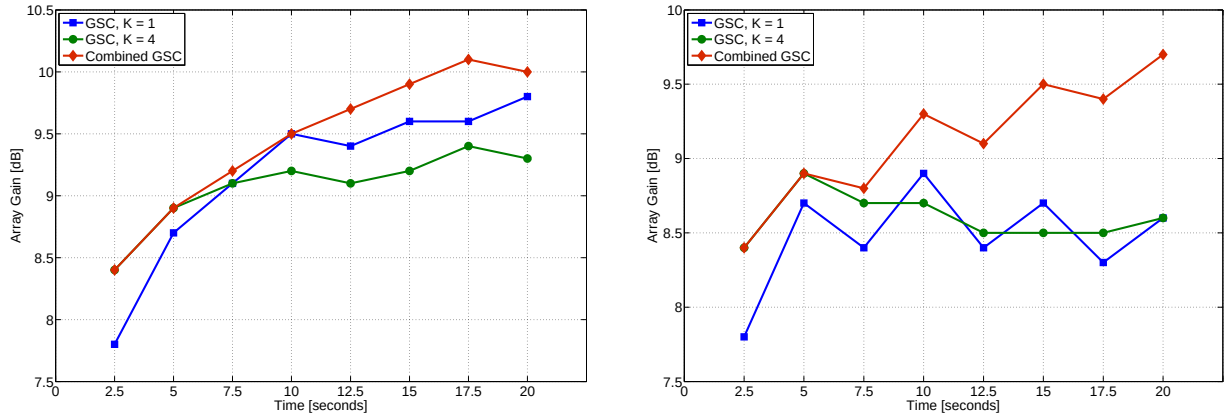

Fig. 6: Performance evaluation in terms of array gain for: (a) the first configuration and (b) the second configuration of the immersive speech scenario.

4.2. Performance Evaluation

The noise reduction performed by a beamforming system is usually evaluated in terms of signal-to-noise ration (SNR), defined as [14]:

$$\mathrm{SNR} = 10 \log \left[\frac{\mathrm{E}\left\{ s_{\mathrm{in}}^2 [n] \right\}}{\mathrm{E}\left\{ s_{\mathrm{in}}^2 [n] \right\} - \mathrm{E}\left\{ s_{\mathrm{out}}^2 [n] \right\}} \right] \qquad (11)$$

where $s_{\mathrm{in}}[n]$ is the generic input clean signal and $s_{\mathrm{out}}[n]$ is the processed signal. The operator $\mathrm{E}\{\cdot\}$ denotes the mean value. In our case, the input SNR value, denoted as $\mathrm{SNR}_{\mathrm{in}}$, the signal $s_{\mathrm{in}}[n]$ represents the desired source signal, while $s_{\mathrm{out}}[n]$ is the microphone signal $u_i[n]$. Similarly, in order to obtain a measure of the SNR for the output of the beamformer, i.e. $\mathrm{SNR}_{\mathrm{out}}$, the overall signal acquired by the microphone array is represented by the signal $s_{\mathrm{in}}[n]$, while $s_{\mathrm{out}}[n]$ is given by the beamformer error signal $e[n]$. Using input and output SNR values it is possible to achieve the *array gain* G, which is defined in dB as the improvement in signal-to-noise ratio between a reference sensor and the array output:

$$\mathrm{G} = \frac{\mathrm{SNR}_{\mathrm{out}}}{\mathrm{SNR}_{\mathrm{in}}}. \qquad (12)$$

Therefore, we use the array gain to evaluate the noise reduction performance of the proposed IAI. Measurements are performed for both the configuration described in the previous subsection, and results are depicted in Fig. 6. In particular, we compare the proposed beamforming technique, characterized by the combined adaptive noise canceller and denoted in Fig. 6 as "Combined GSC", with two conventional GSC beamformers, one using a unitary projection order $K = 1$ and the other one using $K = 4$. Results show that in the first stationary configuration the noise reduction improvement achieved by the proposed system is slight, as it is possible to see in Fig. 6 (a). This is due to the fact that the considered conventional GSCs, and in particular the one with unitary projection order, turn out to be well-suited for that kind of scenario in which sources keep their positions. However, Fig. 6 (b) shows that conventional GSCs suffer the position changes of interfering sources in the second configuration and their performance belows expectation, most of all at steady state. In this case it is possible to notice that the combined GSC exploits the capabilities of the combination scheme of the two MISO systems and achieves a significant improvement with respect to conventional GSCs, thus resulting reliable against moving interfering sources. Therefore, obtained results show that the combined GSC is an effective system both in simple stationary scenarios and also in the presence of more adverse environment conditions.

Performance results obtained from these experiments are not definitely the best achievable values,

since better results may be achieved by using more sophisticated GSC beamformers, e.g. involving any *voice activity detectors* (VADs), adaptive BMs or post-filters. However, the obtained results are sufficient to show the effectiveness of the proposed IAI compared to conventional methods.

5. CONCLUSIONS

In this work we introduced intelligent acoustic interfaces for immersive audio services. IAIs play a fundamental role in acquiring acoustic information, enhancing desired audio signals and reproducing them under the quality constraints desired by user. The core of IAIs is represented by the audio signal processor which aims at enhancing the quality of audio signals in order to provide user with the perception of sharing a different sound space. We focused on one of the immersive audio services which is the immersive speech communications, where the proposed IAIs must extract desired speech information with the highest quality. To this end, we used as audio processor an adaptive beamforming whose main characteristic lies in the combined adaptive noise canceller, which exploits different capabilities of involved adaptive filters. Experimental results shown the effectiveness of the proposed signal processor in reducing interfering noise even in nonstationary conditions, thus preserving the desired speech quality in immersive speech communications. The proposed system paves the way for new IAIs thus increasingly satisfying quality requirements that allow user to enjoy the immersive audio experience.

6. REFERENCES

[1] D. Comminiello. *Adaptive Algorithms for Intelligent Acoustic Interfaces*. PhD thesis, 'Sapienza' University of Rome, December 2011.

[2] Y. Huang, J. Chen, and J. Benesty. Immersive audio schemes. *IEEE Signal Processing Magazine*, 28(1):20–32, January 2011.

[3] J. P. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press, Cambridge, MA, revised edition, 1997.

[4] D. Comminiello, M. Scarpiniti, R. Parisi, A. Cirillo, M. Falcone, and A. Uncini. Multistage collaborative microphone array beamforming in presence of nonstationary interfering signals. In *Proc. of the International Workshop on Machine Listening in Multisource Environments (CHiME '11)*, pages 64–67, Florence, Italy, September 1 2011.

[5] L. Griffiths and C. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation*, 30(1):27–34, January 1982.

[6] J. L. Flanagan, J. D. Johnson, R. Zahn, and G. W. Elko. Computer-steered microphone arrays for sound transduction in large rooms. *Journal of Acoustical Society of America*, 78(5):1508–1518, November 1985.

[7] W. E. Hefley and D. Murray. Intelligent user interfaces. In *Proc. of the 1st International Conference on Intelligent User interfaces (IUI '93)*, pages 3–10, Orlando, FL, January 4-7 1993. ACM.

[8] M. Maybury. Intelligent user interfaces: An introduction. In *Proc. of the 4th International Conference Intelligent User Interfaces (IUI '99)*, pages 3–4, Los Angeles,CA, 1999. ACM.

[9] T. T. Hewett, R. Baecker, S. Card, T. Carey, J. Gasen, M. Mantei, G. Perlman, Strong G., and W. Verplank. *ACM SIGCHI Curricula for Human-Computer Interaction*. The Association for Computing Machinery, Inc., New York, NY, 1992.

[10] D. Comminiello, S. Scardapane, M. Scarpiniti, and A. Uncini. User-driven quality enhancement for audio signal processing. In *134th AES Convention*, Rome, Italy, May 2013.

[11] Y. Huang, J. Benesty, and J. Chen. *Acoustic MIMO Signal Processing*. Springer-Verlag, Berlin, 2006.

[12] J. A. Beracoechea, J. Casajus, L. García, L. Ortiz, and S. Torres-Guijarro. Implementation of immersive audio applications using robust adaptive beamforming and wave field synthesis. In *120th AES Convention*, Paris, France, May 2006.

[13] F. Bettarelli, S. Cecchi, L. Palestini, P. Peretti, and F. Piazza. Sub-band adaptive crosstalk cancellation: A novel approach for immersive audio. In *124th AES Convention*, Amsterdam, Netherlands, May 2008.

[14] M. Brandstein and D. Ward, editors. *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, New York, NY, 2001.

[15] J. Arenas-García, A. R. Figueiras-Vidal, and A. H. Sayed. Mean-square performance of a convex combination of two adaptive filters. *IEEE Transactions on Signal Processing*, 54(3):1078–1090, 2006.

[16] K. Ozeki and T. Umeda. An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties. *Electronics and Communications in Japan*, 67-A(5):19–27, 1984.

[17] D. Comminiello, M. Scarpiniti, R. Parisi, and A. Uncini. A novel affine projection algorithm for superdirective microphone array beamforming. In *Proc. of the IEEE International Symposium on Circuits and Systems (ISCAS '10)*, pages 2127–2130, Paris, France, May 30 - June 2 2010.

[18] C. Paleologu, J. Benesty, and S. Ciochină. A variable step-size affine projection algorithm designed for acoustic echo cancellation. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(8):1466–1478, November 2008.

[19] M. T. M. Silva and V. H. Nascimento. Improving the tracking capability of adaptive filters via convex combination. *IEEE Transactions on Signal Processing*, 56(7):3137–3149, July 2008.

[20] L. A. Azpicueta-Ruiz, A. R. Figueiras-Vidal, and J. Arenas-García. A normalized adaptation scheme for the convex combination of two adaptive filters. In *Proc. of the IEEE 13th International Conference on Acoustics, Speech and Signal Processing (ICASSP '08)*, pages 3301–3304, Las Vegas, NV, March 30 - April 4 2008.