Multi-stage Collaborative Microphone Array Beamforming in Presence of Nonstationary Interfering Signals

Danilo Comminiello[1], *Michele Scarpiniti*[1], *Raffaele Parisi*[1],
Albenzio Cirillo[2], *Mauro Falcone*[2] *and Aurelio Uncini*[1]

[1]Department of Information Engineering, Electronics and Telecommunications,
"Sapienza" University of Rome, Via Eudossiana 18, Rome, Italy
[2] Fondazione Ugo Bordoni, Viale del Policlinico 147, Rome, Italy
danilo.comminiello@uniroma1.it

Abstract

This paper describes a novel adaptive beamforming technique, for speech enhancement applications, designed to be robust to nonstationary interfering sources in noisy and reverberant environments. The proposed beamforming architecture aims at extracting the desired source signal and suppressing interfering signals in a multisource environment with unknown *a priori* conditions. This purpose is realized by means of a *multi-stage collaborative generalized sidelobe canceller*. The trademark of this architecture relies on a two-level convex combination of two *multiple-input single-output* (MISO) adaptive systems, which improves the beamformer capability to track undesired sources, in order to achieve a stronger suppression of interfering signals. The potency of the proposed architecture is proved enhancing the speech quality of the desired source in a hands-free teleconferencing application.

Index Terms: Nonstationary Adaptive Beamforming, Speech Enhancement, Combination of Adaptive Filters

1. Introduction

Machine listening aims at extracting, from audio signals, useful informations for computational or human purpose, such as analysis or synthesis of audio signals. In hands-free speech communications, the audio signals of interest are speech signals, and the audio spatial perception is the desired information to retrieve because it allows to distinguish a certain voice in a noisy environment, simulating the binaural human hearing. In multisource environments, the presence of interfering signals and reverberation may cause the loss of spatial information, thus resulting in a degradation of speech intelligibility. In order to tackle this problem, speech enhancement systems are widely employed in distant talking applications. Microphone array beamforming represents a class of such speech enhancement techniques. Beamforming systems exploit the properties of microphone interfaces which facilitate binaural hearing. However, in order to achieve a quite good recovery of binaural perception, beamforming techniques need to control some aspects of the multisource communication: the spatial realism of sound rendering, the high-quality of acquired speech signals, and the nonstationary of sources which can talk without tethered microphones while moving in the environment [1].

Among beamforming techniques, the *generalized sidelobe canceller* (GSC) [2] is highly effective in acquiring a desired source and adaptively reducing interfering signals. The potency of an adaptive beamformer depends on the choice of the adaptive algorithm. The most popular adaptive algorithms in time-domain are based on the gradient rule, such as the *least mean squares* (LMS) algorithm. The advantage of this family of algorithms is the cheaper computational cost. However, LMS shows poor convergence performance when the filter length is quite long [3], that is the rule in acoustic applications. A faster convergence rate can be yield using Hessian-based adaptive filtering; a typical algorithm is the *recursive least squares* (RLS). However, RLS entails an high computational complexity; therefore, adaptation can become prohibitively expensive, compromising real-time implementations. A good compromise can be obtained by using the family of *affine projection algorithm* (APA), which is widely used in adaptive beamforming [4], [5], showing better convergence rates and manageable computational complexity. The proposed beamforming technique uses a *fast affine projection* (FAP) algorithm [6] to adapt filter coefficients.

Moreover, in order to make the beamformer robust to nonstationary sources, we propose a collaborative adaptive structure, in which, for each channel, we perform the convex combination of two adaptive filters of different families in order to obtain an algorithm with superior tracking capability [7]. Furthermore, in order to use the best parameter setting for each filter we introduce a preventive convex combination between filters of the same family, thus achieving optimum conditions for each combination [8]. The proposed approach is realized in a *multi-stage* collaborative architecture, since the filtering is carried out in more steps.

This paper is organized as follows: the microphone array beamforming technique is described in Section 2; the proposed multi-stage collaborative filtering is detailed in Section 3 and in Section 4 the effectiveness of the proposed beamforming system is assessed. Finally, in Section 5 our conclusions are drawn.

2. Microphone Array Beamforming Technique

The beamforming architecture used in this paper, depicted in Fig. 1, is a typical GSC configuration composed of a microphone array interface, a fixed *delay-and-sum beamformer* (DSB), and an *adaptive noise cancelling* (ANC) path.

Considering a microphone array interface composed of M sensors, the m-th microphone signal, with $m = 0, \ldots, M-1$, is a delayed replica of the target signal $s[n]$ convolved with the acoustic impulse response (AIR) $\mathbf{a}_m$ between the m-th microphone and the desired source, with the addition of background

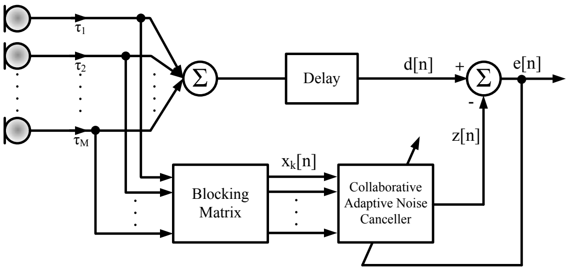

Figure 1: Microphone array beamforming architecture.

noise $v_m[n]$. The DSB spatially aligns the microphone signals with reference to the desired source direction, yielding the speech reference signal $d[n]$:

$$d[n] = \sum_{m=0}^{M-1} \sum_{l=0}^{L-1} a_m[l]\, s[n-l] + v_m[n] \qquad (1)$$

where we suppose that each AIR between the desired source and the m-th microphone has the same length denoted with L.

In the adaptive path of the beamformer, the *blocking matrix* (BM) generates the noise references $x_k[n]$, with $k = 0, \ldots, K-1$, being $K = M-1$. The blocking matrix is implemented by pairwise differences between microphone signals [9]. The noise reference signals are then processed by the *collaborative ANC*, whose structure will be described in the next section. The task of the collaborative ANC is to remove the residual noise components in the speech reference signal, minimizing the output power and yielding the beamformer output signal $e[n]$.

3. Collaborative Adaptive Noise Canceller

The trademark of the proposed beamforming technique is represented by the structure of the collaborative ANC. Generally, an ANC is composed of an adaptive filter bank forming a MISO system. The adopted architecture, depicted in Fig. 2, is a multi-stage convex combination of adaptive filters. In particular, the structure is composed of four different MISO systems, each bringing different filtering capabilities to the whole beamformer. Each MISO system receives the same input signals, which are the noise reference signals coming from the BM. The j-th MISO system can represent the input signals in an $L \times P^{(j)}$ reference noise matrix $\mathbf{X}_{n,k}^{(j)}$:

$$\mathbf{X}_{n,k}^{(j)} = \begin{bmatrix} \mathbf{x}_{n,k} & \mathbf{x}_{n-1,k} & \cdots & \mathbf{x}_{n-P^{(j)}+1,k} \end{bmatrix} \qquad (2)$$

$$= \begin{bmatrix} x_k[n] & \cdots & x_k\left[n-P^{(j)}+1\right] \\ x_k[n-1] & \cdots & x_k\left[n-P^{(j)}\right] \\ \vdots & \ddots & \vdots \\ x_k[n-L+1] & \cdots & x_k\left[n-P^{(j)}-L+2\right] \end{bmatrix}$$

where $P^{(j)}$ represents the projection order for all the filters of the j-th MISO system. We denote with $\mathbf{w}_{n,k}^{(j)} = \left[w_k^{(j)}[n], w_k^{(j)}[n-1], \ldots, w_k^{(j)}[n-L+1]\right]^T$ the $L \times 1$ coefficient vector of the k-th filter belonging to the j-th MISO system, with $j = 1, \ldots, 4$, at n-th time instant. Each filter of the ANC is adapted according to the *fast affine projection* (FAP) algorithm [6], which is summarized in Table 1, omitting the MISO system index for a better comprehension.

	FAP Algorithm
1.	Initialization: $\mathbf{R}_{0,k} = \delta \mathbf{I}_P$, $\hat{\varepsilon}_{0,k} = \mathbf{0}$, $\hat{\mathbf{E}}_{0,k} = \mathbf{0}$
2.	$\mathbf{R}_{n,k} = \mathbf{R}_{n-1,k} + \mathbf{X}_{n,k}^{\lceil P \rceil^T} \mathbf{X}_{n,k}^{\lceil P \rceil} - \mathbf{X}_{n-1,k}^{\lfloor P \rfloor^T} \mathbf{X}_{n-1,k}^{\lfloor P \rfloor}$
3.	$\mathbf{r}_{n,k} = [\mathbf{R}_{n,k}[n-1], \ldots, \mathbf{R}_{n,k}[n-P+1]]$
4.	$y_k[n] = \mathbf{w}_{n,k}^T \mathbf{x}_{n,k} + \mathbf{r}_{n,k} \hat{\varepsilon}_{n,k}$
5.	$e_k[n] = d[n] - y_k[n]$
6.	$\mathbf{E}_{n,k} = \begin{bmatrix} \mu e_k[n] \\ (1-\mu)\hat{\mathbf{E}}_{n-1,k} \end{bmatrix}$
7.	$\hat{\mathbf{E}}_{n,k} = [E_k[n], \ldots, E_k[n-P]]^T$
8.	$\mathbf{g}_{n,k} = \mathbf{R}_{n,k}^{-1} \mathbf{E}_{n,k}$
9.	$\varepsilon_{n,k} = \begin{bmatrix} 0 \\ \hat{\varepsilon}_{n-1,k} \end{bmatrix} + \mathbf{g}_{n,k}$
10.	$\hat{\varepsilon}_{n,k} = [\hat{\varepsilon}_k[n], \ldots, \hat{\varepsilon}_k[n-P]]^T$
11.	$\mathbf{w}_{n,k} = \mathbf{w}_{n-1,k} + \varepsilon_k[n-P+1]\, \mathbf{x}_{n-P+1,k}$

Table 1: Summary of FAP algorithm.

It is well known that the combination of filters of different families of algorithms can improve the tracking capabilities of the whole system [7]. In particular, important results can be achieved combining a family of gradient-based algorithms and a family of Hessian-based algorithms [7]. Taking into account this point, a first distinction between the four MISO systems can be made choosing different values for the projection order. In fact, for $P^{(j)} = 1$ the FAP algorithm turns into the NLMS algorithm yielding gradient-based properties, while for $P^{(j)} > 1$ the FAP algorithm preserves its Hessian nature. Therefore, we set $P^{(j)} = P_1 = 1$ for $j = 1, 2$, and $P^{(j)} = P_2 > 1$ for $j = 3, 4$. This choice will affect the second-stage convex combination. The second stage combination is a system-by-system combination scheme. On the other hand, the convex combination of the first-stage will involve the MISO systems having the same projection order. In particular, the first stage involves two different convex combinations, one for systems $j = 1, 2$ and another one for systems $j = 3, 4$. In this case we differentiate the systems according to the step size value $\mu^{(j)}$: we choose a small step size $\mu^{(j)} = \mu_1$ for $j = 1, 3$ and a large step size $\mu^{(j)} = \mu_2$ for $j = 2, 4$. In this way we further improve the mean-square performance of the adaptive filtering [8]. The kind of combination scheme performed in the first stage is the filter-by-filter scheme.

Let $i = 1, 2$ the index which refers to the convex combination of the first stage. As it is possible to see in Fig. 2, considering the i-th combination, the k-th filter output of the first MISO system, is convex combined with the correspondent k-th filter output of the second MISO system, yielding K outputs, denoted as $z_k^{(i)}[n]$, each related to a noise reference:

$$z_k^{(i)}[n] = \lambda_k^{(i)}[n]\, y_k^{(j)}[n] + \left(1 - \lambda_k^{(i)}[n]\right) y_k^{(j+1)}[n] \qquad (3)$$

where, in this case, the system index is $j = 1$ when $i = 1$,

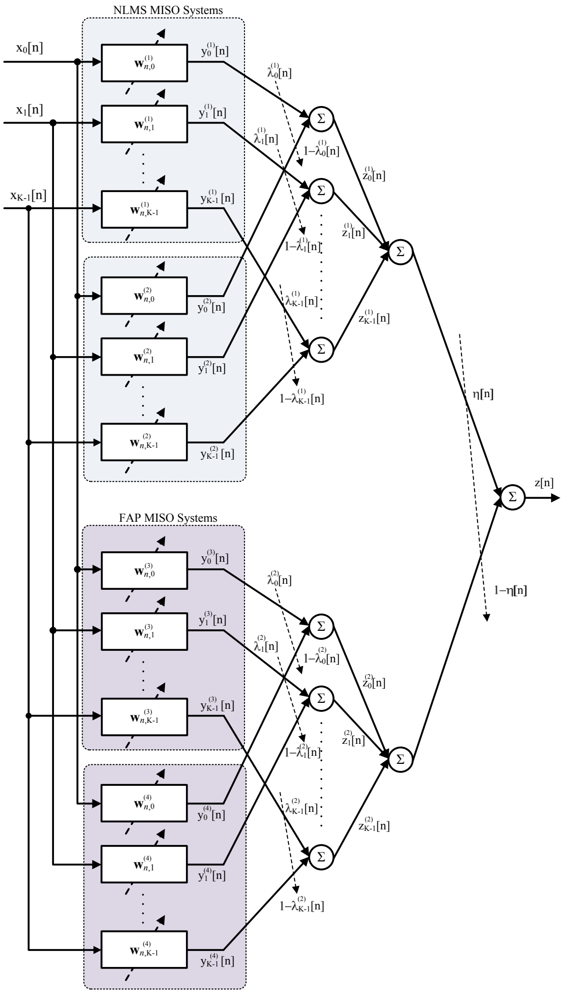

Figure 2: Multi-stage collaborative adaptive noise canceller.

and $j = 3$ when $i = 2$. In (3), $\lambda_k^{(i)}[n]$ represents the k-th mixing parameter of the i-th combination of the first stage, and it is updated using a gradient descent rule through the adaptation of an auxiliary parameter, $a_k^{(i)}[n]$, related to $\lambda_k^{(i)}[n]$ by the expression $\lambda_k^{(i)}[n] = \text{sgm}\left(a_k^{(i)}[n]\right)$, according to [8]:

$$a_k^{(i)}[n+1] = a_k^{(i)}[n] + \qquad (4)$$

$$+\frac{\mu_a}{q_k^{(i)}[n]} e_k^{(j)}[n+1]\,\Delta e_k^{(i)}[n+1]\,\lambda_k^{(i)}[n]\left(1-\lambda_k^{(i)}[n]\right)$$

where $\Delta e_k^{(i)}[n+1] = e_k^{(j+1)}[n+1] - e_k^{(j)}[n+1]$, μ_a is a common step size value for the adaptation of each auxiliary parameter; $q_k^{(i)}[n] = \beta q_k^{(i)}[n-1] + (1-\beta)\left(\Delta e_k^{(i)}[n+1]\right)^2$ is the estimated power of $\Delta e_k^{(i)}[n+1]$, and β is a smoothing factor.

In the second stage a system-by-system convex combination is carried out between the two outputs yielded by the first stage. The second-stage output signal, denoted with $z[n]$, represents the overall ANC output:

$$z[n] = \eta[n]\sum_{k=0}^{K-1} z_k^{(1)}[n] + (1-\eta[n])\sum_{k=0}^{K-1} z_k^{(2)}[n] \qquad (5)$$

where $\eta[n]$ is the mixing parameter of the second stage, adapted using an auxiliary parameter, similarly to (4).

Once computing the second stage convex combination, it is possible to derive the overall beamformer output signal $e[n]$:

$$e[n] = d[n] - z[n]. \qquad (6)$$

The multi-stage collaborative architecture presented above improves the tracking capabilities of the ANC [7], giving robustness to the overall beamforming system in presence of nonstationary interfering signals.

4. Simulation Results

In the this section we carry out two different sets of experiments: the first set aims to assess the effectiveness of the multi-stage collaborative filtering adopted in the proposed beamforming architecture; the second set of experiments is performed to evaluate the proposed beamforming architecture for speech enhancement application in multisource environments. Both the experiments take place in a $10 \times 6, 6 \times 3$ m room with a reverberation time of $T_{60} = 150$ ms.

4.1. Evaluation of the Multi-stage Collaborative Filtering

In the first set of experiments, in order to prove the effectiveness of the multi-stage collaborative filtering, we analyze a single-channel (i.e. $K = 1$) acoustic echo cancelling application, in which the acoustic environment changes due to a nonstationary source or to an alteration in the environemental conditions. The AIR is simulated by means of *Roomsim*, which is a Matlab tool [10]; the AIR is measured by using an 8 kHz sampling rate and it is truncated after $L = 300$ samples. The length of the experiment is $t = 10$ s. Furthermore, an independent white Gaussian noise with zero mean and unit variance is added as background noise, in order to provide 20 dB of *signal to noise ratio* (SNR). In order to introduce an abrupt change in the environment, we shift the AIR circularly to the right by 50 samples, 5 s after the start of the adaptive process. We choose the following parameter settings: $\mu_1 = 0.1$, $\mu_2 = 0.9$, $P_1 = 1$, $P_2 = 2$, $\delta = 30\sigma_{x_k}^2$, where $\sigma_{x_k}^2$ is the power of the input signal. In order to measure the filtering performance we use the *normalized misalignment* $\mathcal{M}$, expressed in dB, defined as:

$$\mathcal{M} = 20\log_{10}\left(\frac{\left\|\mathbf{h}_n - \hat{\mathbf{h}}_{n,k}^{(j)}\right\|_2}{\|\mathbf{h}_n\|_2}\right) \qquad (7)$$

where $\mathbf{h}_n$ is the AIR column vector, and $\hat{\mathbf{h}}_{\mathbf{n},\mathbf{k}}^{(\mathbf{j})}$ is the estimated filter.

Figure 3 displays the performance results; it is possible to see that the multi-stage collaborative filtering exploits the tracking capabilities of all the four filters, always taking the behaviour of the best performing filtering. Furthermore, in Fig. 4 it is possbile to notice the behaviour of the three mixing parameters, $\lambda^{(1)}[n]$ and $\lambda^{(2)}[n]$ related to the first-stage combination, and $\eta[n]$ related to the second-stage combination. Observing Fig. 4 it's still more easy to comprehend the collaboration between the four different filterings.

4.2. Evaluation of the Collaborative Beamformer

In the second set of experiments we assess the effectiveness of the proposed beamforming architecture in terms of speech enhancement. The scenario is the same of the previous simula-

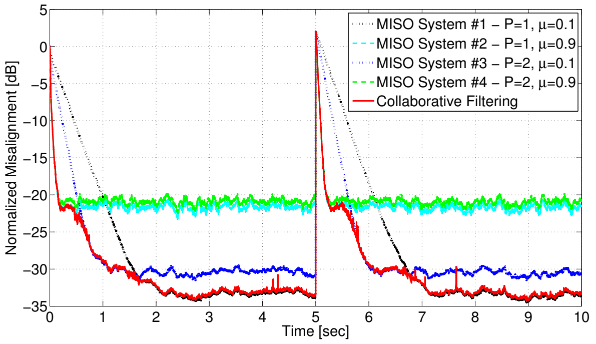

Figure 3: Normalized misalignment comparison.

tions; in this case the source of interest is a female speaker located 50 cm from the center of the microphone array. Two interfering pink noise sources are located respectively 1, 2 m and 2, 2 m from the center of the microphone interface; the first source is on the right of the speaker and the second is on the left. White Gaussian noise is added at microphone signals as diffuse backgroud noise. The overall input SNR level, measured for each microphone signal, is of about 5 dB. The microphone interface is a common *uniform linear array* (ULA) composed of 5 omnidirectional sensors equally spaced with a distance of 4 cm. In order to introduce a change in the acoustic environment, after 5 s from the start of the experiment, we move the two interfering sources 50 cm to the right, keeping unchanged their distance from the center of the array. The enhancement of the speech, provided by the beamformer, and the resulting noise reduction, are usually associated with an SNR improvement, defined as [9]:

$$\mathrm{SNR} = 10 \log \left[\frac{\mathrm{E}\left\{u_{in}^2\left[n\right]\right\}}{\mathrm{E}\left\{u_{in}^2\left[n\right]\right\} - \mathrm{E}\left\{u_{out}^2\left[n\right]\right\}} \right] \qquad (8)$$

where $u_{in}\left[n\right]$ is the generic input clean signal and $u_{out}\left[n\right]$ is the processed signal. The operator $\mathrm{E}\left\{\cdot\right\}$ is the mathematical expectation. We compute the SNR level over the total length of the experiment ($0 - 10$ s) and also in 4 different time subintervals: in the first transient state, from $0 - 2$ s; in the following steady state from $2 - 5$; from $5 - 7$ s to evaluate the new transient state after the path changes; in the following new steady state from $7 - 10$ s. We compare the proposed multi-stage collaborative (MSC) GSC with four simple GSC beamformers, each having one of the MISO system used in the MSC architecture. The results are collected in Table 2 2, in which it is possible to notice the behaviour of the different beamformers and their contribution to the noise reduction in terms of SNR improement. However, it is evident from Table 2 that the best performing architecture is the proposed MSC GSC.

5. Conclusions

In this paper we have introduced a new beamforming technique whose trademark relies on the use of a multi-stage collaborative filtering in the ANC block. The multi-stage collaborative structure is composed of four different MISO systems; in the first stage we carry out the convex combinations of MISO systems adapted by the same family of algorithms in order to find the best configuration for each kind of system. Then, in the second stage, the two combination outputs are combined in order to give robustness to the beamformer in nonstationary environments. The proposed architecture is evaluated in terms of convergence performance and SNR improvement in speech enhancement applications, in which the multi-stage collaborative beamformer outperforms standard beamforming techniques.

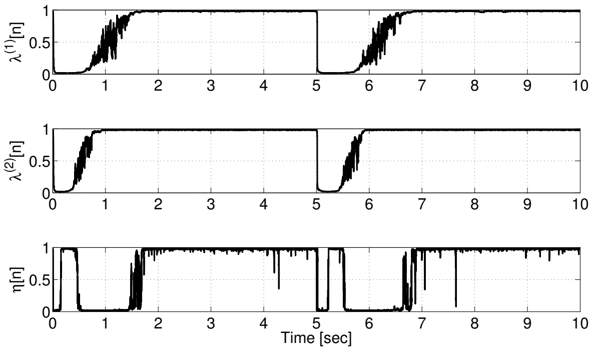

Figure 4: Mixing parameter behaviours.

GSC	0-2 s	2-5 s	5-7 s	7-10 s	0-10 s
NLMS, μ_1	13.2	22.7	12.5	22.5	15.2
NLMS, μ_2	15.1	17.3	15.8	18.2	16.1
FAP, μ_1	15.2	22.6	13.1	23.7	17.9
FAP, μ_2	17.4	18.5	17.1	19.6	17.8
MSC	24.6	31.2	26.7	32.3	28.5

Table 2: SNR comparison in dB.

6. References

[1] Y. Huang, J. Chen, and J. Benesty, “Immersive audio schemes,” *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 20–32, Jan. 2011.

[2] L. Griffiths and C. Jim, “An alternative approach to linearly constrained adaptive beamforming,” *IEEE Trans. on Antennas and Propagation*, vol. 30, no. 1, p. 27 34, Jan. 1982.

[3] A. H. Sayed, *Fundamentals of adaptive filtering*. Hoboken, NJ: John Wiley & Sons, Inc., 2003.

[4] Y. R. Zheng and R. A. Goubran, “Adaptive beamforming using affine projection algorithms,” in *WCCC-ICSP 2000*, vol. 3, Beijing, China, Aug. 2000, pp. 1929–1932.

[5] D. Comminiello, M. Scarpiniti, R. Parisi, and A. Uncini, “A novel affine projection algorithm for superdirective microphone array beamforming,” in *ISCAS 2010*, Paris, France, May 2010, pp. 2127–2130.

[6] S. L. Gay and S. Tavathia, “The fast affine projection algorithm,” in *ICASSP 1995*, vol. 5, Detroit, MI, USA, May 1995, pp. 3023–3026.

[7] M. T. M. Silva and V. H. Nascimento, “Improving the tracking capability of adaptive filters via convex combination,” *IEEE Trans. on Signal Processing*, vol. 56, no. 7, pp. 3137–3149, July 2008.

[8] J. Arenas-García, A. R. Figueiras-Vidal, and A. H. Sayed, “Mean-square performance of a convex combination of two adaptive filters,” *IEEE Trans. on Signal Process.*, vol. 54, no. 3, p. 10781090, Mar. 2006.

[9] M. Bradstein and D. Ward, Eds., *Microphone arrays: Signal processing techniques and applications*. New York: Springer, 2001.

[10] D. R. Campbell, K. J. Palomaki, and G. J. Brown, “Roomsim, a MATLAB simulation of ”shoebox” room acoustics for use in teaching and research,” *Comput. and Inform. Systems*, vol. 9, no. 3, pp. 48–51, 2005.