# DEMIXING OF LINEAR MIXTURES OF UNIMODAL SUPERGAUSSIAN SIGNALS BASED ON GEOMETRIC PROPERTIES

*B.Prieto, C.G.Puntonet, P. Martín-Smith, A.Prieto*
Departamento de Arquitectura y Tecnología de Computadores
E.T.S. Ingeniería Informática. Universidad de Granada, E-18071,Granada, Spain
aprieto@ugr.es

This paper presents how it would be possible to separate linear mixtures of statistically independent signals with unimodal supergaussian probability distributions, with a simple neural network. This procedure is based on geometric properties, and we will show that the distribution maxima of the mixed density distribution belong to straight lines, whose direction vectors, taken as columns of a matrix comprise a demixing matrix. The results obtained with synthetic mixtures of real speech signals are shown.

## I. INTRODUCTION

We assume that the observations, $\boldsymbol{e}(t) = [e_1(t),...,e_p(t)]'$, are generated as a linear mixture $\boldsymbol{A}=(a_{ij})$ ($a_{ij}\in \Re$) of $p$ sources, $\boldsymbol{s}(t)=[s_1(t),...,s_p(t)]'$, such that:

$$\boldsymbol{e}(t) = \mathbf{A}\boldsymbol{s}(t) \tag{1}$$

The goal is to estimate a matrix $\mathbf{W}^{-1}$ such that:

$$\boldsymbol{y}(t) = \mathbf{DP}\boldsymbol{s}(t) \quad ; \quad \mathbf{W}^{-1}\mathbf{A} = \mathbf{DP} \tag{2}$$

where **P** is a permutation matrix and **D** is a diagonal matrix.

We will consider sources with a unimodal and supergaussian probability density function. These distributions are of considerable interest in engineering and science because many random variables are discribed by this model. Among these are the bilateral gamma, double exponential (or Laplace) and beta (with $\alpha=\beta>1$) density functions; for example, in [10] it is shown that a good approximation to measure the speech amplitude density is a bilateral gamma of the form:

$$f(s)=\left(\frac{\sqrt{3}}{8\cdot\pi\cdot\sigma\cdot|s-\mu|}\right)^{1/2}\cdot e^{-\left(\frac{\sqrt{3}\cdot|s-\mu|}{2\sigma}\right)} \qquad -\infty\leq s<\infty \tag{3}$$

where $\sigma^2$ is the variance and $\mu$ the mean. A simpler approximation of the Laplacian density is frequently used:

$$f(s)=\frac{1}{\sqrt{2}\sigma}\cdot e^{-\sqrt{2}\cdot\frac{|s-\mu|}{\sigma}} \tag{4}$$

We consider a discretised time: $t=nT_s$, with $n=1,2,...,N$, and $T_s$ the sampling period. We suppose a memoriless model, such that at every time instant, n, the sources generate a vector or point in the source space, $\boldsymbol{s}(n)$, and at the same time instant a vector or point $\boldsymbol{e}(t)$ is produced in the *mixing (observation) space* and detected by the sensors.

We have seen with signals with a supergaussian probability density function, the frequency of the points located in the mixing space along straight lines (*distribution axes*), in which director vectors taken as columns could define a demixing matrix ***W*.** Considering this idea, in previous papers [6, 7], we proposed an adaptive algorithm for $p=2$ sources based on the consideration of taking the centre of the mixing distribution as its origin and then dividing the space ($e_1$, $e_2$) into sectors (clusters). The algorithm counts the incidence frequency of the observation vectors in each cluster, and identifies the clusters containing the relative frequency maxima as directions of the distribution axes.

This paper addresses the following questions:

- Considering analytic geometry, the procedure principle is formalised. In particular, we show that, for signals with a unimodal supergaussian probability density function, the maxima of the distribution density obtained in the mixing space are along axis, from which a demixing matrix ***W*** can be defined, by using a vector from each axis as column of that matrix.
- A simple neural network is proposed to implement the above concept, and can be used for two or more sources. The network weights are taken as elements of the demixing matrix and are changed adaptively and in an unsupervised way following one of the easiest rules of competitive learning, but without using different

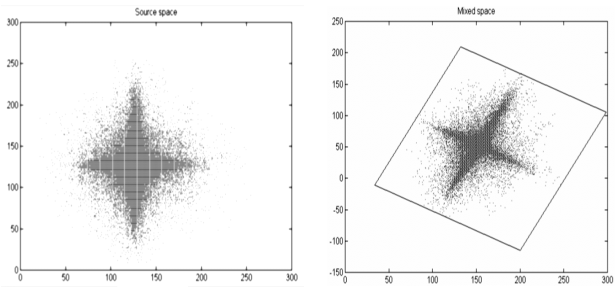
phases of learning and recall. This procedure is more efficient than the previous one [6, 7], because it can be used for more than 2 signals, it is more precise, and its convergence is faster.

- The simulation results show, with synthetic mixtures of real speech signals, that the process achieves an adequate demixing matrix.
- It also noteworthy that compared to previous geometric procedures [5, 8, 9], the hypothesis of bounded sources is not necessary.

## II. THEORETICAL CONSIDERATIONS

Geometric source separation methods are based on considerations about spacial distribution of points in the source and mixing spaces. Thus, we have observed [5, 8, 9] that if the sources have a uniform bounded probability density function, the whole of the source space points form a rectangle if $p=2$, a rectangular parallelepiped if p=3, otherwise a rectangular hyperparallelepiped. The points belonging to this rectangle are mapped into a parallelepiped (or an hyperparallelepiped in general) of the mixing space. It is easy to prove that a vertex of the hyperparallelepiped of the source space is mapped on a vertex of the mixing space, edges map on edges, the centre on the centre, etc.

We have also shown in previous papers [5, 9] that the matrix $\boldsymbol{W}=(w_{ij})$, obtained from p vectors (considered as columns) located at the p edges that are incident upon on any one of the vertices of the mixed hyperparallelepiped is a demixing matrix.




(a) (b)

**Figure 1**. *Source space (a) and mixing space (b) of two speech signals.*

This latter property can be used efficiently to demix bounded signals with a subgaussian probability density function, otherwise it is impossible to obtain a point on an edge (that is, the probability of obtaining these signals is, in fact, very low). In the case of speech signals the point distribution in the source and mixing spaces presents the pattern shown in Figure 1 (if $p=2$) and the point density at the edges is seen to be very low.

When there is a signal, $s_i$, with a distribution function that is unimodal (that is, whose probability fnction only has a local maximum) and symmetrical, the mean ($\mu_i$) the median and the mode coincide [11] and thus the maximum distribution of points is obtained for $s_i=\mu_i$, which, in the source space corresponds to an axis (for $p=2$), to a plane (for $p=3$) or to a hyperplane (for any $p$). When $p=2$, if the two signals ($s_1$, $s_2$) present a distribution function of the above type, the distribution maxima are found at the straight lines $s_1=\mu_1$ and $s_2=\mu_2$, which are parallel to the axes $s_1$ and $s_2$, respectively. The greatest density of points is obtained at the point $C=(\mu_1,\mu_2)$, which we term the distribution centre.

The above statement can be shown analytically. Consider, for example, two signals, $(s_1,\mu_1,\sigma_1)$ and $(s_2,\mu_2,\sigma_2)$, with a bilateral gamma probability function, i.e. (3). The following is then verified:

$$f(s_1)=\left(\frac{\sqrt{3}}{8\cdot\pi\cdot\sigma_1\cdot|s_1-\mu_1|}\right)^{1/2}\cdot e^{-\left(\frac{\sqrt{3}}{2}\cdot\frac{|s_1-\mu_1|}{\sigma_1}\right)} \qquad -\infty\le s_1<\infty \tag{5}$$
$$f(s_2)=\left(\frac{\sqrt{3}}{8\cdot\pi\cdot\sigma_2\cdot|s_2-\mu_2|}\right)^{1/2}\cdot e^{-\left(\frac{\sqrt{3}}{2}\cdot\frac{|s_2-\mu_2|}{\sigma_2}\right)} \qquad -\infty\le s_2<\infty$$

If the two signals are statistically independent, the joint probability density at any point in the source space is given by the product of the marginal prbability densities, i.e.:

$$f(s_1,s_2)=f(s_1)\cdot f(s_2)=\frac{\sqrt{3}}{8\pi\cdot\sqrt{\sigma_1\cdot\sigma_2}\cdot\sqrt{|s_1-\mu_1|\cdot|s_2-\mu_2|}}\cdot e^{-\frac{\sqrt{3}}{2}\left(\frac{|s_1-\mu_1|}{\sigma_1}+\frac{|s_2-\mu_2|}{\sigma_2}\right)} \tag{6}$$

It can be seen from (6) that the maximum probability density is produced when the exponent is zero and/or the denominator is minimum; that is, when:

$$s_1=\mu_1 \quad and \quad s_2=\mu_2 \tag{7}$$

Expression (7) implies that the maximum is produced at point $C=(\mu_1, \mu_2)$, for any value of $s_2$, when $s_1=\mu_1$, and for any value of $s_1$ when $s_2=\mu_2$. Thus, the contribution to the maxima densities are as follows:

$$
\begin{aligned}
&\text{Contribution of } s_1 \rightarrow s_1 = \mu_1 \\
&\text{Contribution of } s_2 \rightarrow s_2 = \mu_2 \\
&\dots\dots\dots\dots\dots\dots\dots\dots \\
&\text{Contribution of } s_p \rightarrow s_p = \mu_p
\end{aligned}
\qquad (8)
$$

The equations in (8) represent hyperplanes (axes parallel to the axes $s_1$, $s_2$ when $p=2$).

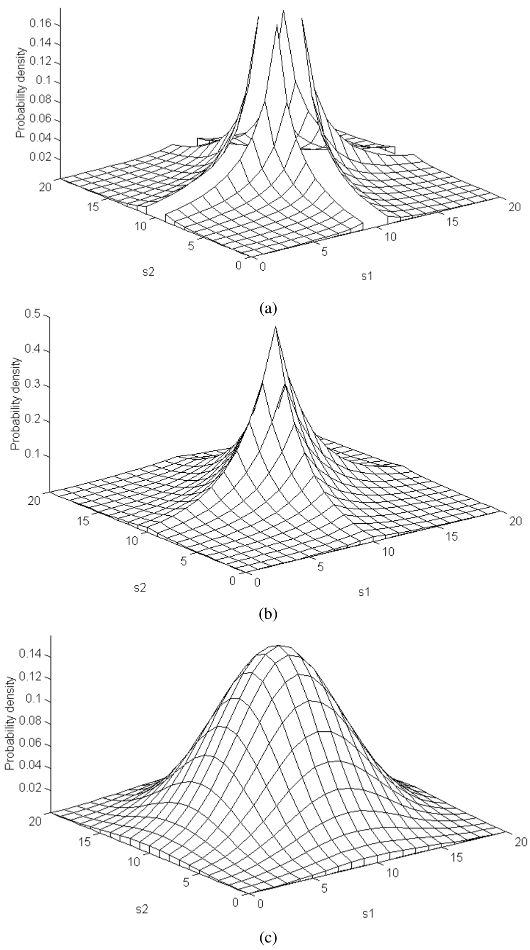



**Figure 2.** *Joint probability density for two signals: (a) Bilateral gamma, (b) Laplacian, (c) Gaussian.*

Figures 2a and 2b show the joint probability density for gamma bilateral and laplacian functions, in the plane ($s_1$,$s_2$). It can clearly be seen that, from the distribution centre the directions with the highest point frequency are $s_1=\mu_1$ and $s_2=\mu_2$.

It should be noted that in the case of gaussian signals , the joint probability density would be:

$$
f(s_1,s_2)=f(s_1)\cdot f(s_2)=\frac{1}{2\cdot\pi\cdot\sigma_1\cdot\sigma_2}\cdot e^{-\left(\frac{(s_1-\mu_1)^2}{2\sigma_1^2}+\frac{(s_2-\mu_2)^2}{2\sigma_2^2}\right)} \qquad (9)
$$

The points of equal density, i.e. $f(s_1,s_2)=f(s_1)\cdot f_2(s_2)=k$ (constant), are as follows:

$$
\frac{(s_1-\mu_1)^2}{2\sigma_1^2}+\frac{(s_2-\mu_2)^2}{2\sigma_2^2}=k \qquad (10)
$$

In other words, the points of equal density are distributed in ellipses (circumferences when $\mu_1=\mu_2$ and $\sigma_1=\sigma_2$) with the centre in C and semiaxes in $\sigma_1\sqrt{2k}$ and $\sigma_2\sqrt{2k}$ . Therefore, there do not exist any suitable distribution axes with which the directions of maximum point frequency could be determined from the central point.

Our goal is to obtain the axes of the mixing space where the distribution maxima are produced. It is well known [4] that if $x_1$ and $x_2$ are statistically independent standard gamma random variables, and $k_1$ and $k_2$ are constants, the random variable:

$$
y=k_1\cdot\ x_1+k_2\cdot\ x_2 \qquad (11)
$$

also has a gamma distribuction function. This property can be applied to each of the mixtures $e_j$, as they are described by expression (11). We thus could obtain, in a similar way to that used for the source space, the points with a maximum joint probability density, as a function of the mixing matrix coefficients, $a_{ij}$. These points can be obtained more easily by using analytic geometry. Then the set of points in the mixing space presenting greatest frequency will correspond to the images of the points in the source space with the greatest frequency. These points correspond to the intersections of the hyperplanes defined in (8). Thus, when $p=2$, the axis $s_1=\mu_1$ is mapped on the following points of the mixing space:

$$
\begin{aligned}
e_1 &= a_{11}\cdot\mu_1+a_{12}\cdot s_2 \\
e_2 &= a_{21}\cdot\mu_1+a_{22}\cdot s_2
\end{aligned}
\qquad (12)
$$

which correspond to the following axis:

$$\frac{e_1 - a_{11} \cdot \mu_1}{a_{12}} = \frac{e_2 - a_{21} \cdot \mu_1}{a_{22}} \quad (13)$$

Similarly, when $p=2$, the other axis of maximum densities ($s_2=\mu_2$) is mapped onto the following axis:

$$\frac{e_1 - a_{12} \cdot \mu_2}{a_{11}} = \frac{e_2 - a_{22} \cdot \mu_2}{a_{21}} \quad (14)$$

When there exist more than 2 signals, the points of maximum density are obtained from the axes at the intersection of the hyperplanes defined by (8). For $p=3$, these axes are at the intersection of the planes ($s_1=\mu_1$, $s_2=\mu_2$), ($s_1=\mu_1$, $s_3=\mu_3$) and ($s_2=\mu_2$, $s_3=\mu_3$), etc. In general, taking point C=($\mu_1$, $\mu_2$,..., $\mu_p$) as the origin of the coordinates, the intersections are mapped onto axes within the mixing space as in the following equations:

$$\begin{aligned}
&\frac{e_1}{a_{11}} = \frac{e_2}{a_{21}} = ...... = \frac{e_p}{a_{p1}} \\
&\frac{e_1}{a_{12}} = \frac{e_2}{a_{22}} = ...... = \frac{e_p}{a_{p2}} \\
&.................................. \\
&\frac{e_1}{a_{1p}} = \frac{e_2}{a_{2p}} = ...... = \frac{e_p}{a_{pp}}
\end{aligned} \quad (15)$$

From expression (15) we deduce that any set of the vectors of the following form:

$$\begin{aligned}
&\mathbf{w}_1 = \left(k_1 a_{11}, k_1 a_{21}, ......., k_1 a_{p1}\right) \\
&\mathbf{w}_2 = \left(k_2 a_{12}, k_2 a_{22}, ......., k_2 a_{p2}\right) \\
&............................................. \\
&\mathbf{w}_p = \left(k_{p1} a_{1p}, k_p a_{2p}, ......., k_p a_{pp}\right)
\end{aligned} \quad (16)$$

(where $k_i$ is an arbitrary constant) can be taken as a column in a demixing matrix **W**, because this matrix will always be related to **A** as in (2). In other words, it is equal to **A** except for one permutation and one scale factor per column.

It is thus shown that the problem of the blind separation of sources with supergaussian probability density becomes just that of determining an arbitrary set of vectors contained within the distribution axes.

In order to obtain the vectors within the distribution axes more easily, any type of preprocessing can be performed, providing the directions remain unaltered. Specifically, we can translate the axis coordinates to the central distribution point C. Note, too, that the sources are not necessarily centred at their means.

## III NEURAL NETWORK TO OBTAIN THE DISTRIBUTION AXES

The previous section showed that, in order to separate signals using the proposed procedure, it is necessary to design algorithms to identify the directions where the local distribution maxima (modes) are to be found. When a mixing vector is sampled, the coordinates are translated to the distribution centre and normalized. The algorithm must locate the directions where there occur $2 \cdot p$ maxima in the production frequency of the preprocessed vectors , $\boldsymbol{e}_n(t)$, taking into account the fact that the latter are pairwise symmetrical.

To implement the corresponding algorithm, we use an unsupervised neural net with competitive learning. This net contains $2 \cdot p$ neurones, $u_j$ with $\boldsymbol{w}_j$ pairwise symmetrical weight vectors. Initially, the weight vectors are uniformly distributed within the p-dimensional space.

Two variants of this procedure are used for network learning. The simplest one is to calculate the proximity $d_j(t)$ from each input vector $\boldsymbol{e}_n(t)$ to the various weights $\boldsymbol{w}_j$ ($j=1,..., 2 \cdot p$) of the network and then to use the following rule to adapt the weights of the winning neurone, $u_w$:

$$\mathbf{w}_w(n+1) = \mathbf{w}_w(n) + \eta(n) \cdot sign(\mathbf{e}_n(n) - \mathbf{w}_w(n)) \quad (17)$$

where:

(1) The proximity, $d_j(t)$ ($-1 \le d_j \le 1$), from the input vector, $\boldsymbol{e}_n(t)$ to the different neurones, $\boldsymbol{w}_j$, is calculated by the scalar product, as both the mixing vectors and the weights are normalized.

(2) Only the weights of the winning neurone, $u_w$, are updated, as this is the one that produces the greatest scalar product.

(3) Because we wish to locate the points where the distribution density is maximum, the weights are modified, not according to the differences between the vectors $\boldsymbol{e}_n(t)$-$\boldsymbol{w}_w$, but according to their sign. This is because the spatial distribution is asymmetric, i.e. the mode does not coincide with the mean. The use of the difference would create the undesirable effect of

the more distant mixing vectors having a greater influence on weight adaptation than the closer ones, and thus the weight vectors would tend towards the mean rather than towards the local mode.

(4) A frequency, $f_j$, is assigned to count the number of times each neurone, $u_j$, has won. The adaptation gain is modified according to the following expression:

$$\begin{aligned} &\eta(n+1) = 0.1 \cdot e^{-\frac{f_w(n)}{\tau_1}} \quad \text{if } \eta(n+1) < \eta_L \\ &\eta(n+1) = \eta_L \qquad \text{otherwise} \end{aligned} \tag{18}$$

In order to prevent the network from becoming stuck in a metastable state, the learning parameter, $\eta(n)$, is maintained at a low value, $\eta_L$.

The other variant is more complex, but provides better results for $p>2$. The differences are as follows:

- The winning vector is only updated if its proximity from the input vector, $d_w$, is greater than a given value $d_c$, expressed by:

$$\begin{aligned} &d_c(n+1) = \left(1 - 2 \cdot e^{-\frac{n}{\tau_2}}\right) \quad \text{if} \quad d_c(n+1) < d_L \\ &d_c(n+1) = d_L \qquad \text{otherwise} \end{aligned} \tag{19}$$

  where, to ensure that the adaptation does not stop, this proximity is limited to the value $d_L$. The use of this criterion is of great value for locating the local maxima, as the neurones are only influenced by the inputs that lie within a hypersphere the centre of which is in the neurone and which has an exponentially decreasing radius, $1- d_c$.

- The weight update is carried out in the usual way [2], that is, using the difference $\mathbf{e}_n(t)$-$\mathbf{w}_w$, and not the sign function:

$$\begin{aligned} &\mathit{iff}\ d_w > d_c \\ &\mathbf{w}_w(n+1) = \mathbf{w}_w(n) + \eta(n) \cdot \left(\mathbf{e}(n) - \mathbf{w}_w(n)\right) \end{aligned} \tag{20}$$

Some preliminary results are given in the following section.

## IV. EXPERIMENTAL RESULTS WITH SPEECH SIGNALS

We have tested the algorithms described in Section III with different mixtures of real speech and bilateral gamma signals. Table 1 illustrates some preliminary experiments, including the number of signals mixed, the type of sources, the mixing matrices used, and the criteria used to adapt the parameters of the neural networks. Some of these matrices were randomly chosen by Cichocki [CIC99] and by Hyvärinen [HYV98]. To measure the quality of the separation, the mean of the absolute error, *ea,* between the elements of the original mixing matrix (normalized) and the demixing matrix (sorted and normalized), that is:

$$e_a = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p}\left|a_{ij} - w_{ij}\right|}{p^2} \tag{21}$$

**Table 1**

| EXPERIMENT # | p | SOURCES | MIXING MATRIX | ADAPTATION PARAMETERS |
|---|---|---|---|---|
| 1 | 2 | Speech | $\begin{pmatrix} 0.3 & -0.1 \\ -0.4 & -1 \end{pmatrix}$ | $\tau_1$=1000<br>$\gamma_L$=0.0001 |
| 2 | 2 | Speech | $\begin{pmatrix} 0.8 & -0.88 \\ 0.77 & 0.9 \end{pmatrix}$ | $\tau_1$=1000<br>$\gamma_L$=0.0001 |
| 3 | 3 | Speech | $\begin{pmatrix} 0.90 & -0.03 & -0.09 \\ -0.54 & 0.78 & -0.96 \\ 0.21 & 0.52 & 0.64 \end{pmatrix}$ | $\tau_1$=1000<br>$\gamma_L$=0.00001 |
| 4 | 4 | Bilateral Gamma | $\begin{pmatrix} -0.28 & -0.82 & 0.06 & -0.39 \\ -2.71 & -0.42 & -0.62 & -0.85 \\ 0.03 & 0.26 & -0.48 & -0.37 \\ -0.33 & 0.63 & -0.11 & 0.74 \end{pmatrix}$ | $\tau_1$=1000<br>$\tau_2$=5000<br>$\gamma_L$=0.001<br>$d_L$=0.998 |

Figures 3-6 show the straight lines obtained after adjusting the absolute errors by the squared minima crieteria, and also the intervals for which a confidence rate of 50%. These figures clearly show that in every case the network weights converge towards a mixing matrix that is valid to perform the separation.

Of course, the greater the number of sources, the slower this convergence will be.

## V. CONCLUSIONS

This paper shows that, from concepts of analytic geometry, for sources with a unimodal supergaussian probability density function, the maxima of the vector distribution densities obtained in the mixing space are distributed along straight lines (distribution axes) which can be used to define a demixing matrix ***W***.

A simple unsupervised neural network is also proposed: this network adaptively locates the distribution axes such that the weights of each neurone, taken as the columns of a matrix, define a demixing matrix. The network adapts its weights by adopting a competitive model.

Finally we provide preliminary results that demonstrate how the procedure produces a convergence towards optimal solutions.

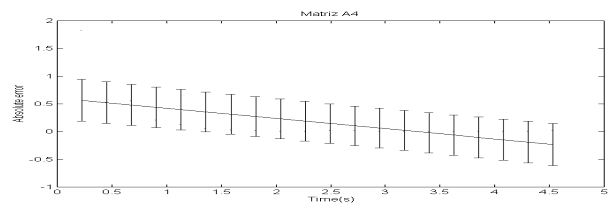



**Figure 3.** *Mean absolute error variation in Experiment 1.*

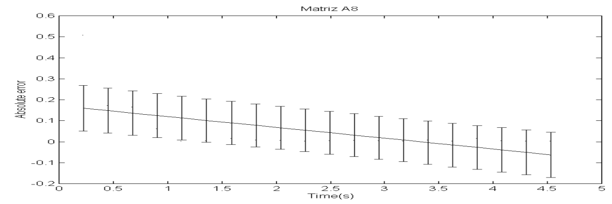



**Figure 4.** *Mean absolute error variation in Experiment 2.*

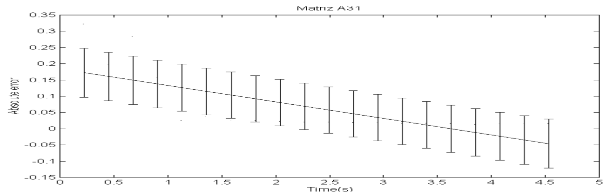



**Figure 5.** *Mean absolute error variation in Experiment 3.*

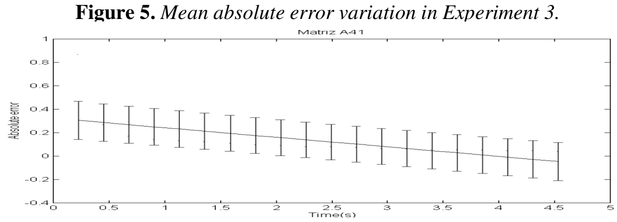



**Figure 6.** *Mean absolute error variation in Experiment 4.*

## VI. REFERENCES

[1] A. Cichocki, R. Thawonmas, *"On-line algorithm for blind signal extraction of arbitrarily distributed, but temporally correlated sources using second order statistics."*. Submitted to Neural Processing Letters, 1999.

[2] S. Haykin, *"Neural networks"*. Prentice Hall, 1999.

[3] A. Hyvärinen, E. Oja, *"Independent component analysis by general nonlinear Hebbian-like learning rules"*. Signal Processing, vol. 64, pp. 301-313, 1998.

[4] N.L.Johnson, S.Kotz, N.Balakrishan, *"Continuous univariate distribuctions"*, Vol.1, 2nd Edt. John Wiley & Sons, 1994.

[5] A. Prieto, C. G. Puntonet, B. Prieto; *"A Neural Learning Algorithm for blind separation of sources based on geometric properties"*. Signal Processing, Vol.64, No. 3, pp. 315-331, 1998.

[6] A.Prieto, B.Prieto, C,G,Puntonet, A.Cañas, P.Martín-Smith, *"Geometric separation of linear mixtures of sources: Application to speech signals"*,International workshop on Independent Component Analysis and Blind Signal Separation (ICA'99) , pp. 295-300, Aussois, France, January 11-15, 1999.

[7] B. Prieto, *"Nuevos algoritmos para separación ciega de fuentes utlizando métodos geométricos"*. Tesis Doctoral, Departamento de Arquitectura y Tecnología de Computadores. Universidad de Granada. October 1999.

[8] C.G.Puntonet, A. Prieto, *"An adaptive geometrical procedure for blind separation of sources"*, Neural Processing Letters, Vol.2, No.5, pp.23-27, Sept. 1995.

[9] C.G.Puntonet, A. Prieto, *"Neural net approach for blind separation of sources based on geometric properties"*, Neurocomputing Vol. 18, No.1-3, pp. 141-164, Jan. 1998.

[10] L.R. Rabiner, and R.W.Schafer, *"Digital Processing Signals"*, Prentice-Hall, 1978.

[11] A. Stuart, J. K. Ord, *"Kendall's advanced theory of statistics"*, Vol. 1: Distribution theory. 6ª Edición. Edward Arnold, 1994.