BLIND SEPARATION OF AUDIO SOURCES CONVOLUTIVE MIXTURES USING PARAMETRIC DECOMPOSITION

A. Aïssa-El-Bey, K. Abed-Meraim and Y. Grenier

{elbey, abed, grenier}@tsi.enst.fr
ENST-Paris, 46 rue Barrault 75634, Paris Cedex 13, France

ABSTRACT

This paper introduces new algorithms for the blind separation of audio sources of instantaneous and convolutive mixtures using modal decomposition. Indeed, audio signals and, in particular, musical signals can be well approximated by a sum of damped sinusoidal (modal) components. Based on this representation, we propose a two steps approach consisting of a signal analysis (extraction of the modal components) followed by a signal synthesis (pairing of the components belonging to the same source). For the signal analysis, we consider a parametric estimation algorithm using ESPRIT technique. A major advantage of the proposed method resides in its ability to separate more sources than sensors in the instantaneous mixture case. Simulation results are given to assess the performance of the proposed algorithm.

1. INTRODUCTION

The problem of blind source separation consists of finding independent source signals from their observed mixtures without a priori knowledge on the actual mixing matrix.
The source separation problem is of interest in various applications [1] such as the localization and tracking of targets using radars and sonars, separation of speakers (problem known as "cocktail party"), detection and separation in multiple access communication systems, independent components analysis of biomedical signals (EEG or ECG), separation of multispectral astronomical images etc.
In the case of non-stationary signals (including the audio signals), certain solutions using time-frequency analysis of the observations exist for the underdetermined case [4,5]. In this paper, we propose an approach using modal decomposition of the received signals [3]. More precisely we propose to decompose the signal into its various modes. The audio signals and more particularly the musical signals can be modeled by a sum of damped sinusoids [6] and hence are well suited for our separation approach. We propose here to exploit this last property for the separation of audio sources by means of modal decomposition. To start, we consider first the case of instantaneous mixtures, then we treat the more challenging problem of convolutive mixtures in the overdetermined case.

2. INSTANTANEOUS MIXTURE CASE

2.1. Data model and assumptions

The blind source separation model assumes the existence of N independent signals $s_1(t), \ldots, s_N(t)$ and M observations $x_1(t), \ldots, x_M(t)$ that represent the mixtures. These mixtures are supposed linear and instantaneous, i.e.

$$x_i(t) = \sum_{j=1}^{N} a_{ij} s_j(t) \qquad i = 1, \ldots, M. \tag{1}$$

This can be represented compactly by the mixing equation

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \tag{2}$$

where $\mathbf{s}(t) \stackrel{\text{def}}{=} [s_1(t), \ldots, s_N(t)]^T$ is a $N \times 1$ column vector collecting the source signals, vector $\mathbf{x}(t)$ similarly collects the M observed signals, and the $M \times N$ mixing matrix $\mathbf{A} \stackrel{\text{def}}{=} [\mathbf{a}_1, \ldots, \mathbf{a}_N]$ with $\mathbf{a}_i = [a_{1i}, \ldots, a_{Mi}]^T$ contains the mixture coefficients. We will suppose that for any pair (i, j) with $i \neq j$, the vectors $\mathbf{a}_i$ and $\mathbf{a}_j$ are linearly independent.
The source signals are supposed to be decomposable in a sum of modal components $c_i^j(t)$, i.e:

$$s_i(t) = \sum_{j=1}^{l_i} c_i^j(t) \qquad t = 0, \ldots, T-1. \tag{3}$$

The usual source independence assumption is replaced here by a quasi-orthogonality assumption of the modal components, i.e.

$$\frac{\langle c_i^j | c_{i'}^{j'} \rangle}{\|c_i^j\| \|c_{i'}^{j'}\|} \approx 0 \quad \text{for} \quad (i,j) \neq (i',j') \tag{4}$$

where

$$\langle c_i^j | c_{i'}^{j'} \rangle \stackrel{\text{def}}{=} \sum_{t=0}^{T-1} c_i^j(t) c_{i'}^{j'}(t)^* \tag{5}$$

and

$$\|c_i^j\|^2 = \langle c_i^j | c_i^j \rangle. \tag{6}$$

In this work, the modal components are in fact damped sinusoids and hence:

$$c_i^j(t) = \Re \left\{ \alpha_i^j {z_i^j}^t \right\} \tag{7}$$

where α_i^j represents the complex amplitude and $z_i^j = e^{d_i^j + \mathrm{i}\omega_i^j}$ is the pole where d_i^j is the negative damping factor and ω_i^j is the angular-frequency. $\Re(\cdot)$ represents the real part of a complex entity. For the extraction of the modal components, we propose to use the ESPRIT-like (Estimation of Signal Parameters via Rotation Invariance Technique) technique that estimates the poles of the signals by exploiting the row-shifting invariance property of the data Hankel matrix. We use Kung's algorithm given in [3][1].

[1]Note that fast and efficient implementation of this algorithm exists in [7].

For the synthesis of the source signals one observes that, thanks to the quasi-orthogonality assumption, one has:

$$\frac{\langle \mathbf{x}|c_i^j\rangle}{\|c_i^j\|^2} \stackrel{\text{def}}{=} \frac{1}{\|c_i^j\|^2}\begin{bmatrix}\langle x_1|c_i^j\rangle\\ \vdots \\ \langle x_M|c_i^j\rangle\end{bmatrix} \approx \mathbf{a}_i$$

where $\mathbf{a}_i$ represents the i^{th} column vector of $\mathbf{A}$. We can then associate each estimated component $\widehat{c}_i^j$ to a space direction (vector column of $\mathbf{A}$) that is estimated by

$$\widehat{\mathbf{a}}_i^j = \frac{\langle \mathbf{x}|\widehat{c}_i^j\rangle}{\|\widehat{c}_i^j\|^2}$$

Two components of a same source signal are associated to the same column vector of $\mathbf{A}$. Therefore, we propose to gather these components by clustering the vectors $\widehat{\mathbf{a}}_i^j$ into N classes [2]. One will be able to rebuild the initial sources up to a constant by adding the various components within a same class.

2.2. Parametric signal analysis

In this section we present an alternative solution for signal analysis. For that, we represent the source signal and hence the observations as sum of damped sinusoids:

$$x_k(t) = \Re\left\{\sum_{l=1}^{L}\alpha_{l,k}z_l^t\right\} \tag{8}$$

where $\alpha_{l,k}$ represents the complex amplitude and $z_l = e^{d_l+i\omega_l}$ is the l^{th} pole where d_l is the negative damping factor and ω_l is the angular-frequency. $\Re(\cdot)$ represents the real part of a complex entity.

For the extraction of the modal components, we propose to use the ESPRIT-like (Estimation of Signal Parameters via Rotation Invariance Technique) technique that estimates the poles of the signals by exploiting the row-shifting invariance property of the $D\times(T-D)$ data Hankel matrix $[\boldsymbol{\mathcal{H}}(x_k)]_{n_1n_2} \stackrel{\text{def}}{=} x_k(n_1+n_2)$, D being a window parameter chosen in the range $T/3 \leq D \leq 2T/3$.

We use of Kung's algorithm given in [3] that can be summarized in the following steps:

1. Form the data Hankel matrix $\boldsymbol{\mathcal{H}}(x_k)$.
2. Estimate the $2L$-dimensional signal subspace $\mathbf{U}^{(L)} = [\mathbf{u}_1 \dots \mathbf{u}_{2L}]$ of $\boldsymbol{\mathcal{H}}(x_k)$ by means of the SVD ($\mathbf{u}_1 \dots \mathbf{u}_{2L}$ are the principal left singular vectors of $\mathcal{H}(x_k)$).
3. Solve (in the least squares sense) the shift invariance equation
$$\mathbf{U}_{\downarrow}^{(L)}\Psi = \mathbf{U}_{\uparrow}^{(L)} \Leftrightarrow \Psi = \mathbf{U}_{\downarrow}^{(L)\#}\mathbf{U}_{\uparrow}^{(L)} \tag{9}$$
where $\Psi = \Phi\Delta\Phi^{-1}$, Φ being a non-singular $2L\times 2L$ matrix and $\Delta = \text{diag}(z_1, z_1^*, \dots, z_L, z_L^*)$. $()^{\#}$ denotes the pseudo-inversion operation and arrows $\downarrow$ and $\uparrow$ denote respectively the last and the first row-deleting operator.
4. Estimate the poles as the eigenvalues of matrix Ψ.
5. Estimate the complex amplitudes by solving the least squares fitting criterion
$$\min_{\boldsymbol{\alpha}}\|\mathbf{x}_k - \mathbf{Z}\boldsymbol{\alpha}\|^2 \Leftrightarrow \boldsymbol{\alpha} = \mathbf{Z}^{\#}\mathbf{x}_k \tag{10}$$
where $\mathbf{x}_k = [x_k(0)\dots x_k(T-1)]^T$ is the observation vector, $\mathbf{Z}$ is a Vandermonde matrix constructed from the estimated poles and $\boldsymbol{\alpha}$ is the vector of complex amplitudes.

[2] There exist techniques that perform both the clustering and the estimation of the number of classes. For simplicity, we assumed here the number of sources known.

2.3. Signal synthesis using vector clustering

For the synthesis of the source signals one observes that thanks to the quasi-orthogonality assumption, one has:

$$\frac{\langle \mathbf{x}|c_i^j\rangle}{\|c_i^j\|^2} \stackrel{\text{def}}{=} \frac{1}{\|c_i^j\|^2}\begin{bmatrix}\langle x_1|c_i^j\rangle\\ \vdots \\ \langle x_M|c_i^j\rangle\end{bmatrix} \approx \mathbf{a}_i$$

where $\mathbf{a}_i$ represents the i^{th} column vector of $\mathbf{A}$. We can then associate each component $\widehat{c}_j^k$ to a space direction (vector column of $\mathbf{A}$) that is estimated by

$$\widehat{\mathbf{a}}_j^k = \frac{\langle \mathbf{x}|\widehat{c}_j^k\rangle}{\|\widehat{c}_j^k\|^2}$$

Two components of a same source signal are associated to the same column vector of $\mathbf{A}$, Therefore, we propose to gather these components by clustering the vectors $\widehat{\mathbf{a}}_j^k$ into N classes. One will be able to rebuild the initial sources up to a constant by adding the various components within a same class.

3. CONVOLUTIVE MIXTURE CASE

3.1. Data model and assumptions

The convolutive mixture case can be represented by:

$$\mathbf{x}(t) = \sum_{l=0}^{L}\mathbf{H}(l)\mathbf{s}(t-l) + \mathbf{w}(t) \tag{11}$$

where $\mathbf{H}(l)$ are $M\times N$ matrices for $l\in[0,L]$ representing the impulse response coefficients of the channel. We considered here only the overdetermined case ($M > N$) and the polynomial matrix $\mathbf{H}(z) = \sum_{l=0}^{L}\mathbf{H}(l)z^{-l}$ is assumed to be irreducible (i.e. $\mathbf{H}(z)$ is of full column rank for all z). The sources are assumed, as in the instantaneous mixture case, to be decomposable in a sum of damped sinusoids satisfying approximately the quasi-orthogonality assumption (4).

Knowing that the convolution preserves the different modes of the signal, we can exploit this property to estimate the different modal components of the source signal using the same approach as in the instantaneous mixture case.

3.2. Signal synthesis step

Once the modal components of all source signals are estimated, one needs to group them in such a way to reconstruct each of the sources. Now this problem is more complex than in the instantaneous mixture case as the correlation of one signal component c_i^j of the i^{th} source signal with the observation leads

to an estimate of the vector $\mathbf{h}_i(c_i^j) \stackrel{\text{def}}{=} \sum_{l=0}^{L} \mathbf{h}_i(l)(c_i^j)^{-l}$ where $\mathbf{H}(l) \stackrel{\text{def}}{=} [\mathbf{h}_1(l) \dots \mathbf{h}_N(l)]$.
Clearly, $\mathbf{h}_i(c_i^j)$ depends on both the i^{th} channel (associated to the i^{th} source signal) and on the pole of the considered modal component c_i^j. Consequently, contrary to the instantaneous mixture case, two components c_i^j and $c_i^{j'}$ of a same source signal do not correspond to vectors of a same spatial direction. For this reason, we propose another synthesis solution that exploits subspace orthogonality together with an appropriate sparsity measure.
Indeed, considering the data model in (11) and for a given 'large enough' window parameter w one has:

$$\mathbf{x}_w(t) \stackrel{\text{def}}{=} [\mathbf{x}^T(t) \dots \mathbf{x}^T(t+w-1)]^T = \boldsymbol{\mathcal{H}} \mathbf{s}_w(t) \qquad (12)$$

where $\boldsymbol{\mathcal{H}}$ is a block-Sylvester matrix of full column rank (see [8] for more details) and $\mathbf{s}_w(t) = [\mathbf{s}^T(t-L) \dots \mathbf{s}^T(t+w-1)]^T$. Hence, the data matrix $\mathbf{X}_w = [\mathbf{x}_w(0) \dots \mathbf{x}_w(T-w)]$ is given by

$$\mathbf{X}_w = \boldsymbol{\mathcal{H}} \mathbf{S}_w \qquad (13)$$

This structure suggests to exploit the signal subspace method [9] to characterize the source signals. More precisely, using the SVD of $\mathbf{X}_w$ one can write (in the noiseless case)

$$\mathbf{X}_w = \mathbf{U} \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{V}_s^H \\ \mathbf{V}_n^H \end{bmatrix} \qquad (14)$$

where $\mathbf{V}_n$ is a basis of the orthogonal subspace to $Range(\mathbf{S}_w^H)$. Hence, $\mathbf{V}_n$ satisfies

$$\mathbf{S}_w \mathbf{V}_n = \mathbf{0} \qquad (15)$$

Using the block Hankel structure of $\mathbf{S}_w$, one can transform the above equality into

$$\boldsymbol{\mathcal{V}}_n \mathbf{S}_T = \mathbf{0} \qquad (16)$$

where $\mathbf{S}_T = [\mathbf{s}(-L) \dots \mathbf{s}(T-1)]^T$ and $\boldsymbol{\mathcal{V}}_n$ is a noise subspace projection matrix constructed from $\mathbf{V}_n$ as shown in [9]. Using the modal decomposition (3) of the sources, one can write

$$\mathbf{S}_T = \mathbf{C}\boldsymbol{\Lambda} \qquad (17)$$

where $\mathbf{C}$ is the matrix whose column vectors correspond to the modal components $\mathbf{c}_i^j = [c_i^j(-L) \dots c_i^j(T-1)]^T$. Therefore equation (16) becomes

$$\boldsymbol{\mathcal{V}}_n \mathbf{C}\boldsymbol{\Lambda} = \mathbf{0} \qquad (18)$$

In other words $\boldsymbol{\Lambda}$ belongs to the Kernel of

$$\mathbf{Q} \stackrel{\text{def}}{=} \mathbf{C}^H \boldsymbol{\mathcal{V}}_n^H \boldsymbol{\mathcal{V}}_n \mathbf{C}$$

Let $\mathbf{P}$ be a matrix whose N column vectors form a basis of Ker$(\mathbf{Q})$. Then

$$\boldsymbol{\Lambda} = \mathbf{P}\widetilde{\boldsymbol{\Lambda}} \qquad (19)$$

where $\widetilde{\boldsymbol{\Lambda}}$ is an unknown $N \times N$ matrix. This means that the subspace solution provides an instantaneous mixture of the sources. To estimate the desired matrix $\widetilde{\boldsymbol{\Lambda}}$, we propose to use a sparsity criterion as shown next.

3.3. Sparsity based criterion

Let us observe that, under the data model assumption, each source signal $\mathbf{s}_i$ is a linear combination of a reduced number l_i of the modal components. In other words, the column vectors of $\boldsymbol{\Lambda}$ should be sparse in the sense that each of them is zero except for a reduced number of entries.
Now, to get the appropriate matrix $\boldsymbol{\Lambda}$, we exploit this property and search for a non-singular $N \times N$ matrix $\widetilde{\boldsymbol{\Lambda}}$ such that $\boldsymbol{\Lambda}$ is sparse. To guarantee that all sources are extracted (i.e the non-singularity of $\widetilde{\boldsymbol{\Lambda}}$), we proceed as follows.
Without loss of generality assume that the sources are uncorrelated and of unit norm so that

$$\frac{1}{T}\mathbf{S}_T^H \mathbf{S}_T \approx \mathbf{I} \qquad (20)$$

Hence,

$$\frac{1}{T}\widetilde{\boldsymbol{\Lambda}}^H \mathbf{P}^H \mathbf{C}^H \mathbf{C}\mathbf{P}\widetilde{\boldsymbol{\Lambda}} \approx \mathbf{I} \qquad (21)$$

Therefore $\widetilde{\boldsymbol{\Lambda}}$ can be estimated up to a unitary matrix $\mathbf{U}$ as the inverse square root of

$$\boldsymbol{\mathcal{S}} = \frac{1}{T}\mathbf{P}^H \mathbf{C}^H \mathbf{C}\mathbf{P} \qquad (22)$$

$$\widetilde{\boldsymbol{\Lambda}} = \boldsymbol{\mathcal{S}}^{-\frac{1}{2}}\mathbf{U} \qquad (23)$$

Now, to obtain the remaining unitary matrix $\mathbf{U}$, we use the sparsity of $\boldsymbol{\Lambda}$. We estimate $\mathbf{U}$ in such a way to minimize

$$\|\boldsymbol{\Lambda}\|_p = \|\mathbf{P}\boldsymbol{\mathcal{S}}^{-\frac{1}{2}}\mathbf{U}\|_p \qquad (24)$$

where $\|\cdot\|_p$ represents the L_p norm $p < 2$ (in our simulation we used $p = 1$) which is know to be a good measure of sparsity [10].
To minimize (24) under unitary constraint, we decompose $\mathbf{U}$ as product of Givens rotation so that we minimize the criterion iteratively as in [11] where at each iteration only one scalar rotation is optimized using a line search technique.

4. SIMULATION

We present first a simulation example in the instantaneous mixture case that illustrates the performance of our blind separation algorithm. For that, we consider a uniform linear array with $M = 3$ sensors receiving the signals from $N = 4$ audio sources. The angles of arrival of the sources are chosen randomly. The sample size is set to $T = 5000$ samples. The observed signals are corrupted by an additive white noise of covariance $\sigma^2\mathbf{I}$ (σ^2 being the noise power). The separation quality is measured by the normalized mean squares estimation errors (NMSE) of the sources evaluated over 100 Monte-Carlo runs. The plots represent the averaged NMSE over the N sources. In figure 1, we compare the separation performance obtained by our algorithm using the parametric technique with $l_i = 30$ for $i = 1, \dots, N$. As a reference, we plot also the NMSE obtained by pseudo-inversion of matrix $\mathbf{A}$ (assumed exactly known). In figure 2, we present a simulation example in the convolutive mixture case that illustrates the performance of our blind separation algorithm. For that, we consider a uniform linear array with $M = 4$ sensors receiving the signals from $N = 3$ audio sources in noiseless case. The filter coefficients are chosen randomly and the channel order is $L = 3$. The sample size is set

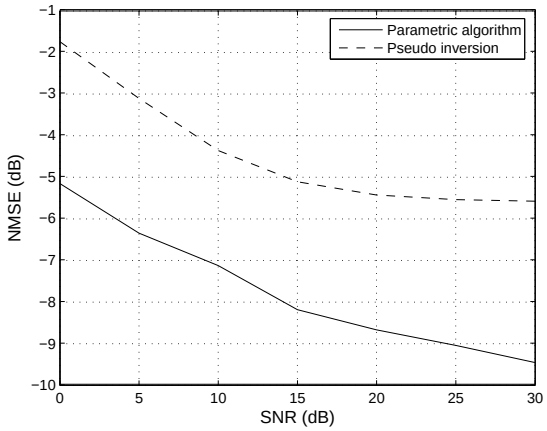
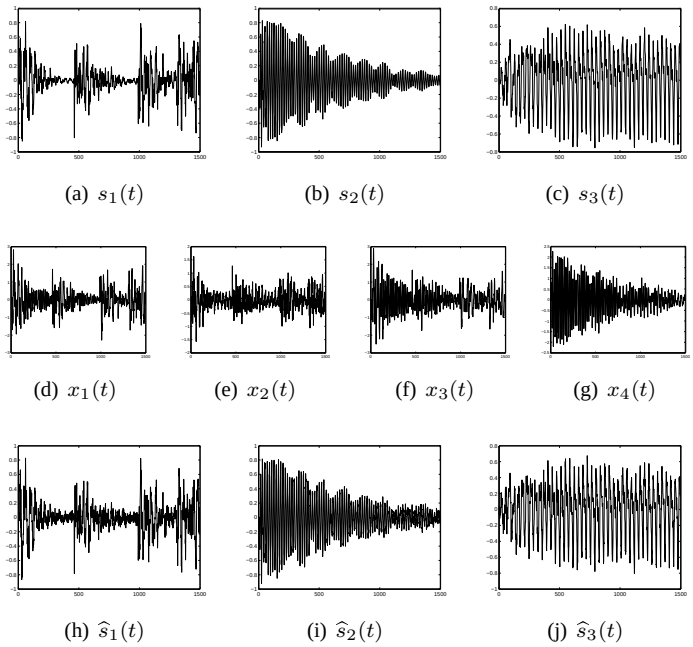


Figure 1: *NMSE versus SNR for 4 audio sources and 3 sensors in instantaneous mixture case: comparison of the performance of our algorithm with the pseudo-inversion of mixing matrix* $\mathbf{A}$ *(assumed exactly known).*

to $T = 1500$ samples. the upper line represents the original source signals, the second line represents the M mixtures and the bottom one represents estimates of sources by our algorithm. We have observed in our simulation that the propose algorithm is very sensitive to noise effect. Obtaining a robust solution is still an open problem under investigation.



(a) $s_1(t)$ (b) $s_2(t)$ (c) $s_3(t)$

(d) $x_1(t)$ (e) $x_2(t)$ (f) $x_3(t)$ (g) $x_4(t)$

(h) $\widehat{s}_1(t)$ (i) $\widehat{s}_2(t)$ (j) $\widehat{s}_3(t)$

Figure 2: *Blind source separation example for 3 audio sources and 4 sensors in convolutive mixture case: the upper line represents the original source signals, the second line represents the* M *mixtures and the bottom one represents estimates of sources by our algorithm.*

5. CONCLUSION

This paper introduces a new blind separation method for audio-type sources using modal decomposition. The proposed method can separate more sources than sensors in the instantaneous mixtures case and provides, in that contest, a better separation quality than the one obtained by pseudo-inversion of the mixture matrix (even if it is known exactly). For the convolutive mixture case we propose to use again modal decomposition but the signal synthesis is more complex and requires the use of subspace projection in conjunction with an appropriate sparsity criterion.

6. REFERENCES

[1] A.K. Nandi (editor), "Blind estimation using higher-order statistics." *Kluwer Acadimic Publishers*, Boston 1999.

[2] I.E. Frank and R. Todeschini, "The data analysis handbook", *Elsevier, Sci. Pub. Co.*, 1994.

[3] S.Y. Kung, K.S. Arun and D.V. Bhaskan Rao, "State-space and singular-value decomposition based approximation methods for the harmonic retrieval problem", *J. Opt. Soc.* Vol. 73, no. 12, 1983.

[4] L. Nguyen, A. Belouchrani, K. Abed-Meraim and B. Boashash, "Separating more sources than sensors using time-frequency distributions." in *Proc. ISSPA*, Vol. II, pp. 583-586, 2001.

[5] A. Jourjine, S. Rickard, O. Yilmaz, "Blind separation of disjoint orthogonal signals: demixing n sources from 2 mixtures," in *ICASSP*, pp. 2985-2988, June 2000.

[6] R. Boyer and K. Abed-Meraim, "Audio modeling based on delayed sinusoids." *IEEE-Tr-SAP*, Vol. 12, N2, pp. 110-120 , March 2004.

[7] R. Badeau, "Méthodes à haute résolution pour l'estimation et le suivi de sinusoides modulées. Application aux signaux de musique." *Phd Thesis, ENST-Paris*, April 2005.

[8] K. Abed-Meraim, Ph. Loubaton and E. Moulines, "A subspace algorithm for certain blind identification problems", *IEEE Tr. on IT.*, vol. 43, pp. 499-511, Mar. 1997.

[9] A. J. Van Der Veen, S. Talwar, A. Paulraj, "A subspace approach to blind space-time signal processing for wireless communication systems", *IEEE Tr. SP*, Jan. 1997.

[10] M. S. O'Brien, A. N. Sinclair and S. M. Kramer,"Recovery of a sparse spike time series by L_1 norm deconvolution", *IEEE Tr. SP*, Dec. 1994.

[11] J. F. Cardoso and A. Souloumiac, "Jacobi angles for simultaneous diagonalization", In SIAM Journal of Matrix Analysis and Applications, Jan. 1996.