OPTIMAL BLIND SEPARATION OF CONVOLUTIVE AUDIO MIXTURES WITHOUT TEMPORAL CONSTRAINTS

Kostas Kokkinakis and Asoke K. Nandi

Signal Processing and Communications Group, Department of Electrical Engineering and Electronics,
The University of Liverpool, Brownlow Hill, Liverpool, L69 3GJ, U.K.
e-mail:{kokkinak,a.nandi}@liv.ac.uk

ABSTRACT

This paper addresses the blind separation of convolutive and temporally correlated speech mixtures, through the use of a multichannel blind deconvolution (MBD) method. In the proposed method (NGA-LP) spatio-temporal separation is achieved by entropy maximization using the natural gradient algorithm (NGA), while a temporal prewhitening stage, based on linear prediction (LP), preserves the original spectral characteristics of each source contribution. It is further shown that a parameterized optimal nonlinearity derived from the generalized Gaussian density (GGD) model, increases the overall separation performance. Experiments with convolutive mixtures illustrate the merits of the proposed method.

1. INTRODUCTION

This paper concentrates on the problem of blind signal separation (BSS), in the general scenario where any m observed signals $\mathbf{x}(t) = [x_1(t), \ldots, x_m(t)]^T \in \mathbb{R}^m$, are considered to be linear and convolutive mixtures of n unknown, yet statistically independent (at each time instant) sources $\mathbf{s}(t) = [s_1(t), \ldots, s_n(t)]^T \in \mathbb{R}^n$. This is often the case in typical acoustic environments, where each microphone (sensor) not only captures the direct contributions from each sound source, but also several reflected copies of the original signals at totally different propagation delays. In this context, the signal observed at the output of the ith sensor is given by:

$$x_i(t) = \sum_{j=1}^{n} \sum_{k=0}^{l-1} h_{ij}(k)\, s_j(t-k), \quad i = 1, 2, \ldots, m. \qquad (1)$$

with t the discrete-time index, $[h_{ij}(k)]$ the room impulse response characterizing the path between the jth source and the ith sensor and $(l-1)$ the order of the FIR filters that model the room acoustic effects. Substituting for the acoustic transfer function $H_{ij}(z)$, the BSS model in the z-domain reads:

$$X_i(z) = \sum_{j=1}^{n} H_{ij}(z)\, S_j(z), \quad i = 1, 2, \ldots, m. \qquad (2)$$

where the convolution operation in (1) reduces to a simple multiplication. The goal of BSS consists of recovering the independent sources from the recorded mixtures without making use of any a priori knowledge. In acoustic separation tasks, resorting to frequency or z-domain is computationally faster than working in time domain. Exploiting this fact, various authors have proposed speech separation techniques, which solely operate in the discrete Fourier domain [11]–[14]. Most tend to perform well even in highly reverberant conditions, nonetheless they all suffer from scaling and permutation indeterminacies. Scaling problems arise due to variations in scaling among different frequency bands, whilst permutation disparities appear as misaligned re-orderings between neighbouring bins. To alleviate these effects, a number of method dependent rules have been reported [11, 12] but a rather general solution is still awaited. MBD methods have also been widely applied in the area of convolutive BSS, operating both in time [15, 16] and the frequency domain [6]–[9]. A typical assumption made in this case — apart from the spatial independence of the sources — is that each source is also an i.i.d. (independent and identically distributed) sequence. In general, the objective of MBD is fulfilled, provided that the recovered estimates are permuted and arbitrarily filtered versions of the sources. This however, results in the output estimates having rather flat spectral characteristics due to the temporal constraints imposed.

The question raised in this paper is whether it is possible to resort to MBD, while at the same time retain the original spectral characteristics in the recovered sources. We propose a BSS method based on MBD and show that it is particularly suited for spatially independent, yet temporally correlated sources. Due to the entropy maximization criterion [2], the efficacy of the optimization process is closely related to the nonlinearity used to model the estimated sources and a mismatch is unavoidable, especially when modelling under the assumption of a fixed distribution shape. To cope with this problem, we introduce a new modelling parameter concerning the shape of the source distributions and aim to improve performance by estimating its optimal value. The validity of the suggested approach is verified through experimental results and performance comparisons.

2. MBD IN FREQUENCY DOMAIN USING THE NATURAL GRADIENT

In [6], Lambert proposed a natural extension of the scalar matrix algebra to the FIR polynomial matrix algebra. It was shown that any FIR filter mixing matrix can be transformed into an FIR polynomial matrix by performing a Fourier transform on its elements. For a j-source and i-sensor system configuration, the mixing matrix $\underline{\mathbf{H}}(z) \in \mathbb{C}^{m \times \ell \times n}$ can be defined as an FIR polynomial matrix,

This work is supported by the Engineering and Physical Sciences Research Council of the U.K. and the University of Liverpool.

with its elements being complex valued FIR polynomials given by:

$$H_{ij}(z) = \sum_{\ell=0}^{k} h_{ij} z^{-\ell} \tag{3}$$

and with the indices, $i = [1, 2, \ldots, m]$, $j = [1, 2, \ldots, n]$ and $\ell = [0, 1, \ldots, l-1]$, representing the observations, sources and each filter coefficient, respectively. Note the term FIR polynomial used here to denote either the z-domain or the discrete Fourier domain. Exploiting the isomorphism between scalar block-Toeplitz and FIR matrices (See [6] for proof), it was realized that the natural gradient algorithm (NGA) of [1], could be easily extended to employ FIR polynomials in the frequency domain [7]–[9]. Based on the entropy maximization (mutual information minimization) approach derived in [2] and combined with the FIR polynomial matrix algebra, the natural gradient learning rule may be shown to accept the following form:

$$\underline{\mathbf{W}}_{k+1} = \underline{\mathbf{W}}_k + \mu \left[\underline{\mathbf{I}} - \text{FFT}\left[\varphi(\mathbf{u})\right] \underline{\mathbf{u}}^H\right] \underline{\mathbf{W}}_k \tag{4}$$

where $(\cdot)^H$ is the Hermitian operator, μ is the step size and $\underline{\mathbf{W}}$ defines the separating FIR polynomial matrix expressed in the frequency domain. In addition, the identity (unit) FIR polynomial matrix $\underline{\mathbf{I}}$ is given by:

$$\underline{\mathbf{I}} = \begin{bmatrix} \bar{1} & \bar{0} \\ \bar{0} & \bar{1} \end{bmatrix} \tag{5}$$

where $\bar{1}$ and $\bar{0}$ represent a sequence of all ones and all zeros, respectively, while the vector FFT $[\varphi(\mathbf{u})]$ denotes the frequency domain representation of the nonlinear monotonic activation function $\varphi(\mathbf{u}) = [\varphi_1(u_1), \ldots, \varphi_m(u_m)]^T$ which in turn operates in the time domain and is equal to:

$$\varphi_i(u_i) = -\frac{\frac{\partial p_{u_i}(u_i)}{\partial u_i}}{p_{u_i}(u_i)} \tag{6}$$

where $p_{u_i}(u_i)$ defines the pdf of each source estimate u_i for all $i = 1, 2, \ldots, m$. The separating matrix $\underline{\mathbf{W}}(z)$ yields the outputs:

$$\underline{\mathbf{u}}(z) = \underline{\mathbf{W}}(z)\, \underline{\mathbf{x}}(z) \tag{7}$$

with the column vectors of the z–transforms of the estimates and mixtures written as:

$$\underline{\mathbf{u}}(z) = [U_1(z), \ldots, U_m(z)]^T \tag{8}$$

$$\underline{\mathbf{x}}(z) = [X_1(z), \ldots, X_m(z)]^T \tag{9}$$

While the update equation in (4) benefits from the computational speed of adapting the separating (unmixing) filters in the frequency domain, it also manages to avoid any permutation indeterminacies, since the optimization criterion operates exclusively in time domain. However, an analysis of the equilibrium points of (4) and after substituting FFT $[\varphi(\mathbf{u})]$ with the complex vector quantity $\mathbf{\Phi}(\mathbf{U}) = [\Phi_1(U_1), \ldots, \Phi_m(U_m)]^T$, reveals that in order for the stationarity conditions of the on-diagonal terms to hold, the following must be satisfied:

$$E\left[\Phi_i(U_i) U_i^*\right] = \bar{1} \tag{10}$$

where $E[\,\cdot\,]$ represents the expectation operator and $(\cdot)^*$ denotes complex conjugation, clearly proving that the NGA imposes a certain scaling constraint on the spectra of the source estimates. This becomes even more tangible if we rewrite (10) in the time domain as:

$$\sum_{\kappa=0}^{t-1} E\left[\varphi_i(u_i(\kappa)) u_i(t-\kappa)\right] = \delta_t \tag{11}$$

for some non-zero time lag κ with δ_t the Kronecker delta function, equal to 1 for $t = 0$ and 0 otherwise.

3. NATURAL GRADIENT ALGORITHM BASED ON LINEAR PREDICTION ANALYSIS

The temporal constraints imposed within each source, are translated into unknown linear filtering operations, which in turn produce signal estimates with equalized (white) spectra. The side-effect of whitening — realized as the flattening of the signal power spectrum, with energy at higher frequency bands being increased at the expense of energy in lower frequencies — is clearly undesired in speech separation applications. To remedy this situation, [15] operated on the assumption that the spectral characteristics of each source are dominant over the rest in each mixture and used a cascaded system of separating and linear prediction (LP) filters to preserve each source colour at the output. In [5], we exploited the temporal model of speech and suggested an alternative system configuration also based on LP, without availing ourselves of the above assumption. In this modification, the spatial separation filters are adapted using temporally independent LP residuals, while the contribution of each source is obtained by applying the estimated filters to the original mixtures without any modification. Hence, we may define the diagonal FIR polynomial matrix $\underline{\mathbf{A}}(z)$:

$$\underline{\mathbf{A}}(z) = \text{diag}\left[A_1(z), \ldots, A_m(z)\right] \tag{12}$$

with each diagonal entry of the matrix simply consisting of the pth-order linear prediction error (LPE) filters with transfer functions subsequently given by:

$$A_i(z) = 1 - \sum_{k=1}^{p} \alpha_i(k) z^{-k} \tag{13}$$

where each vector $[\alpha_i(k)]$ represents the linear prediction coefficients and is defined for $1 \leq k \leq p$ and for all $i = 1, 2, \ldots, m$. From the observed mixtures at the sensor output, the acquired innovation processes, i.e., the prediction error signals (residuals) in the z-domain can be written as:

$$V_i(z) = \sum_{i=1}^{m} A_i(z)\, X_i(z), \quad i = 1, 2, \ldots, m \tag{14}$$

For every ith observation, the coefficients $\alpha_i(k)$ in (13) may be typically estimated in the time domain by minimizing the mean squared prediction error of the mixture $x_i(t)$ with respect to its past samples $x_i(t-k)$, which yields the set of Yule-Walker autocorrelation equations [10]:

$$E\left[\left(x_i(t) - \sum_{k=1}^{p} \alpha_i(k)\, x_i(t-k)\right) x_i(t-\ell)\right] = 0 \tag{15}$$

for $\ell = 1, 2, \ldots, p$ and $i = 1, 2, \ldots, m$, from which the optimal prediction coefficients can be obtained via the Levinson-Durbin recursive method. The estimated innovations can thus be used to

adapt the coefficients of the spatial separation FIR polynomial matrix $\underline{\mathbf{W}}(z)$ according to:

$$\underline{\mathbf{W}}_{k+1} = \underline{\mathbf{W}}_k + \mu \left[\underline{\mathbf{I}} - \text{FFT}\left[\varphi(\mathbf{u})\right] \underline{\mathbf{u}}^H\right] \underline{\mathbf{W}}_k \qquad (16)$$

with the spatially and temporally independent outputs in the frequency domain written as:

$$\underline{\mathbf{u}}(z) = [U_1(z), \ldots, U_m(z)]^T = \underline{\mathbf{W}}(z)\, \underline{\mathbf{v}}(z) \qquad (17)$$

Applying the separating matrix in the original mixtures, produces the spatially separated yet temporally correlated source estimates:

$$\underline{\widehat{\mathbf{s}}}(z) = \left[\widehat{S}_1(z), \ldots, \widehat{S}_m(z)\right]^T = \underline{\mathbf{W}}(z)\, \underline{\mathbf{x}}(z) \qquad (18)$$

Experiments carried out in [5] demonstrate the advantages of the NGA-LP. The removal of inherent speech short-time correlations, was shown to greatly benefit the proposed method, which reduces to a spatial separation process resulting in increased stability, separation performance and speed of convergence. More significantly, the original contribution of each source signal is extracted with its unique power spectral characteristics fully preserved.

4. OPTIMAL NONLINEARITY BASED ON THE GGD

The separation performance and convergence properties of the approach at hand, highly depend upon the relation of the nonlinear function used in the model and the pdf of the sources to be recovered. Although a certain flexibility can be afforded, an ill matched activation function can result in a model mismatch and furthermore in a non converging solution. In [5], we stipulated that the innovations retain enough information to preserve the optimization criterion and hence the spatial separation filters are capable of separating the coloured observations. However, this is often not the case. In NGA-LP the spatial separation filters are adaptively estimated using the temporally independent versions of the output source estimates. In effect, the assumption of a Laplacian distribution model is merely an approximation of the actual density of the LP residuals and therefore the nonlinearity $\varphi(\mathbf{u}) = \text{sign}(\mathbf{u})$ cannot be regarded as being optimal. A strong corroboration point for this argument comes also from [3], where a similar observation has been made. Since, the innovation processes have a sparse distribution, in general, it is possible to sufficiently approximate their distribution by employing the generalized Gaussian density (GGD) model. For a zero-mean and unit variance speech signal x the GGD is defined as:

$$p_x(x) = \frac{\beta}{2\alpha\Gamma(1/\beta)}\, e^{-(|x|/\alpha)^\beta} \qquad (19)$$

where α and β are positive real parameters and $\Gamma(\cdot)$ denotes the Gamma function defined as $\Gamma(\beta) = \int_0^\infty x^{\beta-1} e^{-x} dx$. Parameter α is a generalized measure of the variance of the distribution and is referred to as the dispersion or scale parameter, while β describes the exponential rate of decay and defines the shape of the distribution [4]. As special cases of the GGD, a Laplacian distribution is defined for $\beta = 1$, a standard Gaussian distribution for $\beta = 2$ and a Gamma distribution for $\beta = 0.5$ as shown in Fig. 1. Substituting (19) into (6) we can deduce the family of nonlinear activation functions based on the GGD:

$$\varphi_i(u_i) = |u_i|^{(\beta-1)}\, \text{sign}(u_i) \qquad (20)$$

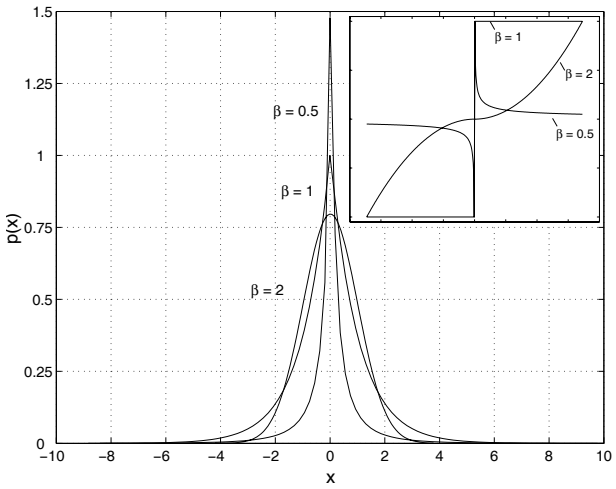

Fig. 1. Pdf of the generalized Gaussian density (GGD) model for different values of the shape parameter $\beta = 0.5, 1, 2$. The corresponding nonlinear functions φ_i for each distribution are also superimposed in the graph.

which by taking into account that $\text{sign}(u_i) = u_i/|u_i|$, may be further reduced to:

$$\varphi_i(u_i) = \frac{u_i}{|u_i|^{(2-\beta)}}, \quad 0 < \beta < 1 \qquad (21)$$

defined for $u_i \neq 0$ with $\varphi_i(u_i)$ acting elementwise on the source estimate components u_i for all $i = 1, 2, \ldots, m$. Note here that (21) depends solely on the shape parameter of each source distribution.

5. EXPERIMENTAL RESULTS

In this section an empirical approach is undertaken to experimentally define an optimal value for the shape parameter β of the generalized nonlinearity used in the NGA-LP. The data set used, employs two female speech signals and the corresponding algorithm parameters are summarized in Table 1. Convolutive mixtures are generated from a non-minimum phase mixing system consisting of 5-tap filters. The separation performance of the algorithm is measured using the interference-to-signal (ISR) ratio:

$$\text{ISR} = 10 \log_{10} \frac{\left\|\underline{\mathbf{G}}_{ij}\right\|^2}{\left\|\underline{\mathbf{G}}_{ii}\right\|^2}, \quad i \neq j \qquad (22)$$

where the global cascade system is equal to $\underline{\mathbf{G}}(z) = \underline{\mathbf{W}}(z)\, \underline{\mathbf{H}}(z)$. We limit our search for an optimal value for β in the range $[0, 2]$. To investigate the algorithm performance in this range, the separa-

Length of speech signals	5 seconds
Sampling frequency	8 kHz
Blocksize	M = 128 points
Order of LP filters	p = 9
Separating filters	$\underline{\mathbf{W}}$ = 2 × 256 × 2
Step size	μ = 0.001
Number of iterations	N = 30

Table 1. Algorithm parameters.

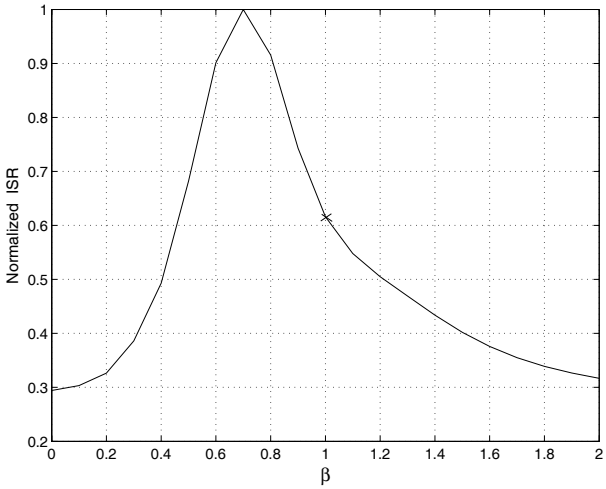

Fig. 2. Plot of separation efficiency versus the evolution of shape parameter $\beta \in [0, 2]$ for GGD-based nonlinearity, with point $\times$ at $\beta = 1$.

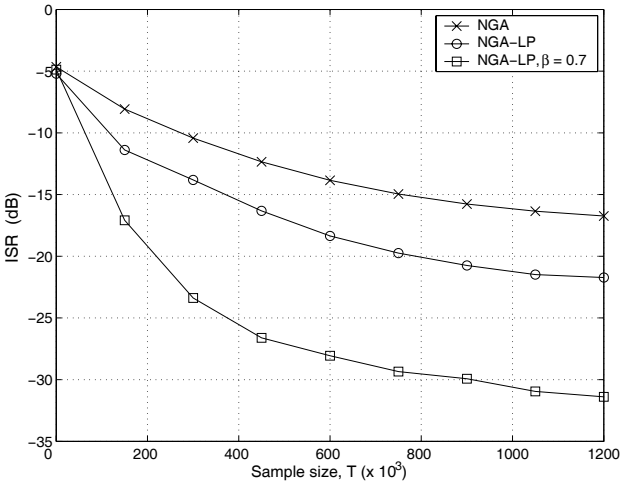

Fig. 3. ISR versus sample size. BSS performance for the NGA, NGA-LP and NGA-LP using the optimal nonlinearity in (21) for $\beta_{\text{opt}} = 0.7$.

tion efficiency expressed as the normalized ISR is plotted against β. As Fig. 2 reveals, there exists a single optimal value $\beta_{\text{opt}} = 0.7$ for the parameter at hand, for which BSS performance is maximized. In hindsight the result is substantiated, due to the fact that the pdf of an innovation process is known to closely resemble a Gamma distribution. Furthermore, a number of experiments with convolutive mixtures of speech carried out in [13], also point to the strong validity of these findings. For the estimated value β_{opt} of the Gaussian exponent, (21) yields the corresponding optimal nonlinear function to be used in the NGA-LP. The performance of the proposed algorithm when compared against the NGA and the NGA-LP, with both operating under the standard threshold nonlinearity, is depicted in Fig. 3. NGA-LP clearly outperforms the NGA (mostly by about 5 dB), while when combined with the optimal nonlinearity, it exhibits an improvement of about 15 dB. It is also apparent from Fig. 2 that the normalized ISR at $\beta = 1$ (the operating point for the unoptimized NGA-LP method) is only about 0.6 of the optimized ($\beta_{\text{opt}} = 0.7$) NGA-LP efficiency, which is essentially reflected in the ISR differences shown in Fig. 3.

6. CONCLUSIONS

We have employed the NGA-LP BSS method based on MBD, which combines the natural gradient algorithm and the entropy maximization principle, to separate convolutive mixtures of speech in the frequency domain. Endorsing a temporal prewhitening stage ensures that there are no spectral constraints being imposed on the recovered source estimates. We have also derived an optimal nonlinear function, based on the GGD and have shown its ability to accurately model the underlying distributions of the source innovation processes in the proposed modification. Future work will focus on investigating techniques, towards a continuously adaptive estimate of the generalized Gaussian exponent parameter.

7. REFERENCES

[1] S. Amari, A. Cichocki and H. Yang, "A New Learning Algorithm for Blind Signal Separation" *Advances in Neural Information Processing Systems 8*, MIT Press, Cambridge, 1996, pp. 757–763.

[2] A. Bell and T. Sejnowski, "An Information Maximization Approach to Blind Separation and Blind Deconvolution" *Neural Computation*, Vol. 7, No. 6, July 1995, pp. 1129–1159.

[3] N. Charkani and Y. Deville, "Self-Adaptive Separation of Convolutively Mixed Signals with a Recursive Structure. Part II: Theoretical Extensions and Application to Synthetic and Real Signals" *Signal Processing*, Vol. 75, No. 2, June 1999, pp. 117–140.

[4] S. Gazor and W. Zhang, "Speech Probability Distribution" *IEEE Signal Processing Letters*, Vol. 10, No. 7, July 2003, pp. 204–207.

[5] K. Kokkinakis, V. Zarzoso and A. K. Nandi, "Blind Separation of Acoustic Mixtures based on Linear Prediction Analysis" In *Proc. 4th Int. Symp. on ICA and BSS*, Japan, April 1–4, 2003, pp. 343–348.

[6] R. H. Lambert, *Multichannel Blind Deconvolution: FIR Matrix Algebra and Separation of Multipath Mixtures*. Ph.D. Thesis, University of Southern California, May 1996.

[7] R. H. Lambert and A. J. Bell, "Blind Separation of Multiple Speakers in a Multipath Environment" In *Proc. ICASSP*, Munich, Germany, April 21–24, 1997, pp. 423–426.

[8] T.-W. Lee, A. J. Bell and R. H. Lambert, "Blind Separation of Delayed and Convolved Sources" *Advances in Neural Information Processing Systems 9*, MIT Press, Cambridge, 1997, pp. 758–764.

[9] T.-W. Lee, A. J. Bell, and R. Orglmeister, "Blind Source Separation of Real World Signals" In *Proc. ICNN*, Houston, Texas, June 9–12, 1997, pp. 2129–2135.

[10] J. Makhoul, "Linear Prediction: A Tutorial Review" *Proc. of the IEEE*, Vol. 63, No. 4, April 1975, pp. 561–580.

[11] N. Mitianoudis and M. E. Davies, "Audio Source Separation of Convolutive Mixtures" *IEEE Trans. on Speech and Audio Processing*, Vol. 11, No. 5, September 2003, pp. 489–497.

[12] L. Parra and C. Spence, "Convolutive Blind Separation of Non-Stationary Sources" *IEEE Trans. on Speech and Audio Processing*, Vol. 8, No. 3, May 2000, pp. 320–327.

[13] R. K. Prasad, H. Saruwatari and K. Shikano, "Problems in Blind Separation of Convolutive Speech Mixtures by Negentropy Maximization" In *Proc. 2003 Int. Workshop on Acoustic Echo and Noise Control*, Kyoto, Japan, September 8–11, 2003, pp. 287–290.

[14] P. Smaragdis, "Blind Separation of Convolved Mixtures in the Frequency Domain" *Neurocomputing*, Vol. 22, No. 1–3, November 1998, pp. 21–34.

[15] X. Sun and S. C. Douglas, "A Natural Gradient Convolutive Blind Source Separation Algorithm for Speech Mixtures" In *Proc. 3rd Int. Conf. on ICA and BSS*, San Diego, December 9–13, 2001, pp. 59–64.

[16] K. Torkkola, "Blind Separation of Convolved Sources based on Information Maximization" In *Proc. NNSP*, Kyoto, Japan, September 4–6, 1996, pp. 423–432.