1

Separation of non-stationary natural signals

Lucas C. Parra, Sarnoff Corporation

Clay D. Spence, Sarnoff Corporation

Abstract

Most approaches to the problem of source separation use the assumption of statistical independence. To capture statistical independence higher order statistics are required. In this chapter we will demonstrate how higher order criteria, such as maximum kurtosis, arise naturally from the property of non-stationarity. We will also show that source separation of non-stationary signals can be based entirely on second order statistics of the signals. Natural signals, be it images or time sequences, are for the most part non-stationary. For natural signals therefore we argue that non-stationarity is the fundamental property, from which specific second or higher order separation criteria can be derived. We contrast the linear bases obtained using second order non-stationarity and ICA for the cases of natural images and speech powers. Based on these results we argue that speech powers can in fact be understood as a linear superposition of non-stationary spectro-temporal independent components, while this is not so evident for a spatial basis of images intensities. Finally we demonstrate the practical utility of the second order non-stationarity concept with a separation algorithm for the problem of convolutive source separation. We show its effectiveness on acoustic mixtures in real reverberant environments.

1.1 Second and higher order separation criteria in the context of non-stationary signals

Most approaches to source separation have been based on the condition of statistical independence of the constituent signals. Conventionally, higher order statistics are required to capture statistical independence.

In fact if the source signals are identically and independently distributed (i.i.d.) samples of a stationary distribution second order statistics are not sufficient for separation. Fortunately, natural signals are often not stationary, nor independently distributed.

Usually natural signals are sampled on a regular lattice, e.g., one-dimensional arrays for time sequences, two-dimensional arrays for images, three-dimensional arrays for image sequences and so on. Neighboring samples on such a lattice are often correlated. Furthermore the statistics of the samples are often not stationary on the lattice, that is, the signals are non-stationary in time or space. This rich spatio-temporal diversity, a result of the spatio-temporal ordering of samples, rather than being a problem, can actually simplify the problem of source separation. In this chapter we try to highlight the statistical properties that arise from non-stationarity that allow us to define useful separation criteria.

First, we will argue that non-stationarity justifies higher order criteria, in particular maximum kurtosis. The kurtosis of a signal has been used for separation of natural signals, such as speech, or to find independent linear bases for natural images. Thereby the assumption is made that the sources of interest have a sparse distribution or high kurtosis. In section 1.2 we show that non-stationarity in fact leads to high kurtosis signals validating the high kurtosis assumption for source separation for natural signals.

It is well known that for temporally correlated sources signal separation can be based entirely on second order statistics (Bradwood, 1978; Bar-Ness et al., 1982; Fety and Van Ulffelen, 1988; Tong and Liu, 1990; Belouchrani et al., 1993; Molgedey and Schuster, 1994; Van Gerven and Van Compernolle, 1995). Less well-known is the fact that non-stationary signals can be separated using decorrelation as well (Souloumiac, 1995; Matsuoka et al., 1995; Parra and Spence, 2000a). Almost identical algorithms can be used in both cases, as we will discuss in section 1.3.

Furthermore, for correlated *and* non-stationary signals the more difficult problem of convolutive source separation can be solved using second order statistics as first indicated by Weinstein et al., 1993 and shown in Kawamoto et al., 1998; Parra and Spence, 2000a. We will present a specific algorithm and examples of recovering speech in a reverberant environment in section 1.5.

In section 1.4 we verify our arguments by applying the criterion of non-stationary second order decorrelation to find linear bases for natural images and speech signals. We contrast the results with a standard

ICA algorithm. For images, in the past, linear bases for small spatial areas of the visual field have been compared to visual receptive fields (Olhausen and Field, 1996). Receptive fields in the auditory domain are found to be spectro-temporal patterns (Kowalski et al., 1996; deCharms and Merzenich, 1998; Theunissen and Doupe, 1998). Different linear bases for small patches of natural images and small spectro-temporal patches of speech powers will be presented in section 1.4. The merits and problems of a linear superposition model of non-stationary independent components will be discussed for both these domains. In essence we argue that acoustic signal powers are well described by such a linear superposition model. We question, however, the concept of linear superposition of image intensities.

1.2 Kurtosis of non-stationary signals

To facilitate the analysis we will first introduce a rather general class of stochastic processes that expresses the main property of non-stationarity we would like to address. Then we will show that this class of non-stationary signals is heavy tailed as measured by their kurtosis.

Assume that at any given instance the signal is specified by a probability density function with zero mean and arbitrary scale or power. Furthermore assume that the signal is non-stationary in the sense that its power varies from one time instance to the next.† A closely related class of signals is the so-called spherical invariant random process (SIR-P). If the signals are short time Gaussian and the powers vary slowly the class of signals we have just described are approximately SIRPs. SIRPs have been shown to cover a large range of different stochastic processes with very different higher order properties depending on the distribution of powers. They have been used in a variety of signal processing applications (Goldman, 1976; Rangaswamy et al., 1993; Rupp, 1993). Band-limited speech in particular has been shown to be well described by SIRPs (Brehm and Stammler, 1987). Natural images have also been modeled by what in essence is closely related to SIRPs — a finite (Spence and Parra, 2000) or infinite (Wainwright and Simoncelli, 2000) mixture of linear Gaussian features.

Consider a stochastic process with samples $z(t)$ drawn from a zero mean distribution $p_z(z)$. Assume we observe a scaled version of this

† Throughout this chapter we will refer to signals that are sampled in time. Note that all the arguments apply equally well to a spatial rather than temporal sampling, that is, images rather than time series.

process, where the magnitude or scale changes over time. If the scale at any instant is given by $s(t) > 0$ sampled from $p_s(s)$ the resulting observable process,

$$x(t) = s(t)z(t)\,, \tag{1.1}$$

is distributed according to

$$p_x(x) = \int_0^\infty ds\, p_s(s)\, p_x(x|s) = \int_0^\infty ds\, p_s(s)\, s^{-1}\, p_z(\frac{x}{s})\,. \tag{1.2}$$

We refer to $p_x(x)$ as the long term distribution and $p_z(z)$ as the instantaneous distribution. In essence $p_x(x)$ is a mixture distribution with infinitely many kernels $s^{-1}p_z(\frac{x}{s})$. We would like to relate the sparseness of $p_z(z)$, as measured by the kurtosis, to the sparseness of the observable distribution $p_x(x)$. Kurtosis is defined as the ratio between the fourth and second cumulant of a distribution (Kendal and Stuart, 1969). As such it measures the length of the distribution's tails, or the sharpness of its mode. For a zero mean random variable x this reduces (up to a constant†) to

$$K[x] = \frac{\left\langle x^4 \right\rangle_x}{\left\langle x^2 \right\rangle_x^2}\,. \tag{1.3}$$

The expectation over $p_x(x)$ is denoted, $\left\langle f(x) \right\rangle_x = \int dx f(x) p_x(x)$.

We will show now that the kurtosis of the long term distribution is always larger than the kurtosis of the instantaneous distribution unless the scale is stationary, i.e. $K[x] \geq K[z]$, where the equality holds for $p_s(s) = \delta(s - c)$ for any arbitrary constant c.

Since z and s are independent it is easy to show that,

$$\left\langle x^n \right\rangle_x = \left\langle s^n \right\rangle_s \left\langle z^n \right\rangle_z\,. \tag{1.4}$$

With this we can write the kurtosis $K[x]$ of the long term distribution in terms of the kurtosis $K[z]$ of the instantaneous distribution,

$$K[x] = K[z]\frac{\left\langle s^4 \right\rangle_s}{\left\langle s^2 \right\rangle_s^2}\,. \tag{1.5}$$

Since for any density of a positive random variable, $p_s(s) \geq 0$ and $p_s(s) = 0$ for $s < 0$, we know that

$$\int_0^\infty ds\, p_s(s)\, (s^2 - c^2)^2 \geq 0 \tag{1.6}$$

† The conventional definition is $K[x] = \left\langle x^4 \right\rangle_x / \left\langle x^2 \right\rangle_x^2 - 3$. We neglect for convenience the constant -3 in our definition.

for arbitrary c. Equality holds only when the integrand vanishes everywhere, i.e. if and only if $p_s(s)$ vanishes except for $s^2 = c^2$. The only distribution that vanishes everywhere except in one point is the Dirac δ-distribution. Therefore equality holds if and only if $p_s(s) = \delta(s - c)$. Rewriting 1.6 we obtain

$$\int ds\, p_s(s)\,(s^4 - 2s^2c^2 + c^4) = \left\langle s^4 \right\rangle_s - 2\left\langle s^2 \right\rangle_s c^2 + c^4 \geq 0 \qquad (1.7)$$

The minimum with respect to c occurs at $\left\langle s^2 \right\rangle_s^{1/2}$. Inserting this gives

$$\frac{\left\langle s^4 \right\rangle_s}{\left\langle s^2 \right\rangle_s^2} \geq 1 \qquad (1.8)$$

We have therefore,

$$K[x] \geq K[z] \qquad (1.9)$$

The equality holds if and only if $p_s(s) = \delta(s - c)$. This result has been first published by Beale and Mallows, 1959 for symmetric $p_z(z)$.

This result means that if the scale $s(t)$ is fixed, i.e. the magnitude of the signal is stationary, the kurtosis will be minimal. Inversely, non-stationary signals, defined as a variable scaling of an otherwise stationary process, will have increased kurtosis.

In the discussion above the time t did not play a particular role other than to indicate that we sample the random variables over time and that a variable scale translates to scale non-stationarity. We also did not demand that samples be drawn independently. That is, we are implicitly allowing signals that are correlated over time.

In summary we can say that signals with varying power will tend to have high kurtosis.

In our definition a stationary Gaussian signal has kurtosis 3. Non-stationary Gaussian signals will be leptokurtic ($K > 3$). All SIRPs are therefore leptokurtic. The assumption that a distribution is sparse, frequently used in source separation of natural signals, is therefore justified. However we will not go into the specific approaches of source separation based on higher order statistics as they have been extensively studied and are described elsewhere in this volume. A good overview of higher order contrast functions is given by Cardoso, 1999. In the following section we will demonstrate how non-stationarity can be used more explicitly and in fact simplifies the problem of source separation by allowing us to use statistics of only second order.

1.3 Separation based on non-stationary second order statistic

In this section we will concentrate on instantaneous mixtures. The more complex case of convolutive mixtures will be presented in the next section. To clarify the notation let us restate the basic problem. Assume d_s statistically independent sources $\mathbf{s}(t) = [s_1(t), ..., s_{d_s}(t)]^T$. These sources are mixed in a linear medium leading to d_x sensor signals $\mathbf{x}(t) = [x_1(t), ..., x_{d_x}(t)]^T$ that may include additional sensor noise $\mathbf{n}(t)$,

$$\mathbf{x}(t) = A\mathbf{s}(t) + \mathbf{n}(t)\,. \tag{1.10}$$

How can one identify the $d_x d_s$ coefficients of the mixture A and how can one find an estimate $\hat{\mathbf{s}}(t)$ for the unknown sources?

Early work in the signal processing community proposed that the linearly mixed sources could be recovered by finding the linear transformation W that decorrelates the measured signals, or, more specifically, that diagonalizes the measured auto-correlations at multiple time delays (Bradwood, 1978; Bar-Ness et al., 1982; Fety and Van Ulffelen, 1988; Tong and Liu, 1990). For an instantaneous mix of non-white signals this is in fact sufficient as discussed by Molgedey and Schuster, 1994; Van Gerven and Van Compernolle, 1995. Almost the same approach can be taken if the signals are non-stationary rather than non-white. In this case we can use the covariance estimated during different time intervals instead of the covariance for different delays, as we will now show. This was indicated by Weinstein et al., 1993 and has been explicitly used by Souloumiac, 1995; Kawamoto et al., 1998.

Note that strictly speaking we are not searching for independent components. We are merely searching for decorrelated model signals that explain a static and instantaneous linear mixture. Certainly statistically independent signals are uncorrelated, but the inverse is not always true.

1.3.1 Forward model estimation

We can formulate the instantaneous covariance $R_x(t)$ of the measured signals at time t with the assumption of independent noise as

$$\begin{aligned} R_x(t) \equiv& \left\langle \mathbf{x}(t)\mathbf{x}^T(t) \right\rangle = A \left\langle \mathbf{s}(t)\mathbf{s}^T(t) \right\rangle A^T + \left\langle \mathbf{n}(t)\mathbf{n}^T(t) \right\rangle \\ \equiv& A\Lambda_s(t)A^T + \Lambda_n(t)\,. \end{aligned} \tag{1.11}$$

Since we assume uncorrelated sources at all times, we postulate diagonal covariance matrixes $\Lambda_s(t)$. We also assume uncorrelated noise at each sensor, i.e. diagonal $\Lambda_n(t)$. Any reasonable definition of average $\langle f(t) \rangle$

that satisfies these diagonality criteria is applicable, such as the average over an ensemble of independent realizations of the signal $\mathbf{s}(t)$ and noise $\mathbf{n}(\mathbf{t})$, or the average over a time window surrounding the time instance t, $\langle f(t) \rangle = \sum_{\tau=-T/2}^{T/2} f(t+\tau)$. We prefer to reserve the latter definition however for the empirical sample average $\hat{R}(t)$, which is a noisy metric of a presumed underlying instantaneous correlation $R(t)$ that gave rise to the signals.

Note that any scaling and permutation of the coordinates of $\Lambda_s(t)$ can be absorbed by A. It is well known that the solution is therefore only specified up to an inherently arbitrary permutation and scaling. Thus we are free to choose the scaling of the coordinates in $\mathbf{s}$. For now we choose $A_{ii} = 1\,, i = 1, ..., d_s$, which places d_s conditions on our solutions.

For non-stationary signals, a set of K equations (1.11) for different times $t_1, ...t_K$ and the d_s scaling conditions give a total of $Kd_x(d_x + 1)/2 + d_s$ constraints on $d_s d_x + d_s K + d_x K$ unknown parameters A, $\Lambda_s(t_1), ..., \Lambda_s(t_K)$, $\Lambda_n(t_1), ..., \Lambda_n(t_K)$.† Assuming all conditions are linearly independent‡ we have sufficient conditions if

$$Kd_x(d_x+1)/2 + d_s \geq d_s d_x + d_s K + d_x K\,. \tag{1.12}$$

It is interesting to note that in the square case, $d_s = d_x$, there are not sufficient constraints to determine the additional noise parameters unless $d_x \geq 4$, no matter how many more times one considers.§ If we assume zero additive noise, in principle $K = 2$ is sufficient to specify the solution up to arbitrary permutations.

In the square case, $d_s = d_x$, with zero noise in principle $K = 2$ is sufficient to specify the solution up to arbitrary permutations. In that case the problem can be solved as a non-symmetric eigenvalue problem as outlined by Molgedey and Schuster, 1994. The covariances at times t_1 and t_2 satisfy,

$$\begin{aligned} R_x(t_1) &= A\Lambda_s(t_1)A^T && (1.13) \\ R_x(t_2)^{-1} &= A^{-T}\Lambda_s(t_2)^{-1}A^{-1} && (1.14) \end{aligned}$$

which can be combined to

$$R_x(t_1)R_x(t_2)^{-1}A = A\Lambda_s(t_1)\Lambda_s(t_2)^{-1} \tag{1.15}$$

† We will abbreviate the notation in the reminder of the paper by writing $\Lambda_s(..)$ when we refer to all $\Lambda_s(t_1), ..., \Lambda_s(t_K)$. We use this notation also for $\Lambda_n(t)$.

‡ Conditions on $R_x(t)$ and $\Lambda_s(t)$ for linear independence are outlined in Molgedey and Schuster, 1994.

§ One can see this by re-writing the inequality as $K(d_x^2 - 3d_x) + 2(d_x - d_x^2) \geq 0$. The second term is never positive, and the first is only positive if $d_x \geq 4$.

Equation (1.15) represents a non-symmetric eigenvalue problem. In general its solutions, A, are not orthogonal as expected.

The difficulty with such algebraic solutions, however, is that one does not have perfect estimates of $R_x(t)$. Even if we assume zero noise at best one can assume non-stationary signals and measure the sample estimates $\hat{R}_x(t)$ within some time interval. If we interpret the inaccuracy of that estimation as measurement error

$$E(t) \equiv \hat{R}_x(t) - \Lambda_n(t) - A\Lambda_s(t)A^T \,, \qquad (1.16)$$

it is reasonable to estimate the unknown parameters by minimizing the total measurement error for a sufficiently large K,

$$\hat{A}, \hat{\Lambda}_s(..), \hat{\Lambda}_n(..) = \underset{A,\Lambda_s(..),\Lambda_n(..),A_{ii}=1}{\arg\min} \sum_{k=1}^{K} \|E(t_k)\|^2 \ . \qquad (1.17)$$

The matrix norm here is the sum of the absolute squared values of every coefficient. Note that $\|E(k)\|^2 = \mathrm{Tr}[E(k)E^H(k)]$. This is a least squares (LS) estimation problem.

1.3.2 Simultaneous diagonalization with unitary transformations

In the context of source separation it is common to reduce the problem of finding a general linear transformation A to that of finding only a rotation V by first diagonalizing the long term covariance (Cardoso and Souloumiac, 1993; Comon, 1994).

This approach is feasible only in the the case that the noise is of equal power, σ^2, in all sensors. In addition, for our case this approach requires stationary noise, $\Lambda_n(t) = \sigma^2 I$. Under these constraints the solution to (1.17) can be computed explicitly using an elegant simultaneous diagonalization technique presented in Cardoso and Souloumiac, 1996.

Consider the long term average $\bar{R}_x = \sum_t R_x(t)$. To properly account for additive noise, one has to first obtain an estimate of the long term covariance $\bar{R}_y$ of the signal portion in the mixture, $y(t) = As(t)$. Following (1.11) this covariance is given by $\bar{R}_y = \bar{R}_x - \sigma^2 I$. Conventionally for $d_s < d_x$ a more stable estimate of $\bar{R}_y$ is obtained using an eigenvalue decomposition of the sample average estimate of $\bar{R}_x$, i.e. $\bar{R}_x = U^T DU$, where D represents a diagonal matrix with the eigenvalues of $\bar{R}_x$ and U the corresponding matrix of eigenvectors. The eigenvectors with the largest eigenvalues are conventionally referred to as the principal components. A robust estimate of $\bar{R}_y$ is given by $\bar{R}_y = U^T \max(D - \sigma^2 I, 0)U$.

This is a classic subspace analysis where the d_x-dimensional data is assumed to originate in a d_s-dimensional subspace. The average of the $d_x - d_s$ smallest eigenvalues provide an estimate for σ^2.†

One can show that $\bar{R}_y$ is diagonalized, i.e. $Q\bar{R}_y Q^T = I$, by a whitening operation $Q = (D - \sigma^2 I)^{-1/2} U$.† Furthermore, for any arbitrary rotation V the matrix

$$A = Q^+ V \bar{\Lambda}_s^{-1/2} \tag{1.18}$$

satisfies $\bar{R}_x = A\bar{\Lambda}_s A^T - \bar{\Lambda}_n$, which is precisely the diagonalization condition (1.11) for the long term averages. Q^+ represents the pseudo-inverse of Q. Both of these results can only be derived for equal power noise. To find the correct A one has to determine V using additional constraints. It is important to note, however, that the problem of finding a general linear transformation A has been reduced to finding a rotation V.

Consider now the whitened signal portion of the mixture, $\mathbf{z} = Q\mathbf{y}(t)$. Using (1.18) its instantaneous covariance can be written as

$$R_z(t) = QR_y(t)Q^T = QA\Lambda_s(t)A^TQ^T = V\Lambda_s(t)\Lambda_s^{-1}V^T \ . \tag{1.19}$$

We find therefore that the remaining rotation V has to diagonalize $R_z(t)$ for all times. With the same reasoning on a minimal estimation error, therefore, we can rewrite (1.11) as

$$E_z(t) \equiv QE(t)Q^T = R_z(t) - V\Lambda_s(t)\Lambda_s^{-1}V^T \tag{1.20}$$

$$\hat{V}, \hat{\Lambda}_s(..), = \underset{V=V^T, \Lambda_s(..)}{\arg\min} \sum_{k=1}^{K} \|E_z(t_k)\|^2 \ . \tag{1.21}$$

This is exactly the problem addressed in the approximate joint eigenspace algorithm described in Cardoso and Souloumiac, 1996.

Note that $R_y(t)$ can only be properly estimated if $\Lambda_n(t)$ is known. This is why we requested stationary noise powers that may be estimated with the subspace analysis outlined above. Additionally we had to demand equal noise powers. There might be cases where this conditions are too restrictive. In addition, and as we will see in section 1.5, one may wish to place other constraints on A. In such case direct optimization with respect to A as in (1.17) may be desired.

† If in fact, $\bar{\Lambda}_n = \sigma^2 I$, in principle the smallest eigenvalues should all have the same magnitude σ^2. However, a sample average is never exact and sampling average instabilities require us to use $\max(D - \sigma^2 I, 0)$.

† In case that $d_s < d_x$ only the d_s rows of Q corresponding to non-negative values of $D - \sigma^2 I$ have to be considered.

1.3.3 Gradient based diagonalization

To find the extrema of the LS cost $J = \sum_{k=1}^{K} \|E(t_k)\|^2$ in (1.17) let us compute the gradients with respects to its parameters‡

$$\frac{\partial J}{\partial A} = -4 \sum_{k=1}^{K} E(t_k) A \Lambda_s(t_k) \tag{1.22}$$

$$\frac{\partial J}{\partial \Lambda_s(t_k)} = -2 \operatorname{diag}\left[A^T E(t_k) A\right] \tag{1.23}$$

$$\frac{\partial J}{\partial \Lambda_n(t_k)} = -2 \operatorname{diag}\left[E(t_k)\right] \tag{1.24}$$

We can find the minimum with respect to A, and $\Lambda_s(t_k)$ with a gradient descent algorithm using the gradients (1.22), and (1.23). The optimal $\Lambda_n(t_k)$ for given A and $\Lambda_s(t_k)$ at every gradient step can be computed explicitly by setting the gradient in (1.24) to zero, which yields $\hat{\Lambda}_n(t_k) = \operatorname{diag}\left[\hat{R}_x(t_k) - A\Lambda_s(t_k)A^T\right]$.

1.3.4 Estimation of source signals

In the case of a square and invertible mixing matrix $\hat{A}$, the signal estimates are trivially computed to be $\hat{\mathbf{s}} = \hat{A}^{-1}\mathbf{x}$. In the non-square case for $d_s < d_x$ we can compute the LS estimate

$$\hat{\mathbf{s}}_{\mathrm{LS}}(t) = \arg\min_{\mathbf{s}(t)} \|\mathbf{x}(t) - \hat{A}\mathbf{s}(t)\| = (\hat{A}^T\hat{A})^{-1}\hat{A}^T\mathbf{x}(t)\,. \tag{1.25}$$

If we assume the additive noise to be short term Gaussian, but not necessarily white or stationary, we can compute the maximum likelihood (ML) estimate

$$\begin{aligned} \hat{\mathbf{s}}_{\mathrm{ML}}(t) &= \arg\max_{\mathbf{s}(t)} p[\mathbf{x}(t)|\mathbf{s}(t);\hat{A},\hat{\Lambda}_n(t)] \\ &= \left[\hat{A}^T\hat{\Lambda}_n(t)^{-1}\hat{A}\right]^{-1}\hat{A}^T\hat{\Lambda}_n(t)^{-1}\mathbf{x}(t)\,. \end{aligned} \tag{1.26}$$

where $p()$ is the Gaussian probability density given by the noise density. If we further assume the signal to be short term Gaussian, again not necessarily white or stationary, we can compute the maximum a posteriori

‡ The diagonalization operator here zeros the off-diagonal elements, i.e. $\operatorname{diag}(M)_{ij} = \begin{cases} M_{ij}, & i=j \\ 0, & i \neq j \end{cases}$

probability (MAP) estimate. For Gaussian densities the MAP estimate is equal to the conditional expectation $E[\mathbf{s}(t)|\mathbf{x}(t); \hat{A}, \hat{\Lambda}_n(t), \hat{\Lambda}_s(t)]$

$$\begin{aligned}\hat{\mathbf{s}}_{\mathrm{MAP}}(t) &= \arg\max_{\mathbf{s}(t)} p[\mathbf{s}(t)|\mathbf{x}(t); \hat{A}, \hat{\Lambda}_n(t), \hat{\Lambda}_s(t)] \\ &= E[\mathbf{s}(t)|\mathbf{x}(t); \hat{A}, \hat{\Lambda}_n(t), \hat{\Lambda}_s(t)] \\ &= \left[\hat{A}^T \hat{\Lambda}_n(t)^{-1} \hat{A} + \hat{\Lambda}_s(t)^{-1}\right]^{-1} \hat{A}^T \hat{\Lambda}_n(t)^{-1} \mathbf{x}(t)\,.\end{aligned} \tag{1.27}$$

Note however that the resulting estimates may not be uncorrelated. Assuming that the model is correct and that we found the correct estimate $\hat{A} \approx A$,

$$\left\langle \hat{\mathbf{s}}_{\mathrm{LS}} \hat{\mathbf{s}}_{\mathrm{LS}}^T \right\rangle \approx \left\langle \mathbf{s}\mathbf{s}^T \right\rangle + (\hat{A}^T\hat{A})^{-1} \hat{A}^T \Lambda_n \hat{A} (\hat{A}^T\hat{A})^{-1}\,. \tag{1.28}$$

Since the second term may not be diagonal, the resulting estimates can be correlated. However, this is not a problem since the correlation is entirely due to correlated noise and the signal portion of the estimates remains uncorrelated.

Instantaneous source separation based on second order statistics has found applications to image processing (Schießl et al., 2000), magneto-enciphalography (Wübbeler et al., 2000; Tang et al., 2000) and other biomagnetic recordings (Ziehe et al., 2000). The following two sections will demonstrate two applications of source separation. First we will compare the criteria of independence and non-stationary decorrelation in the case of natural images and the time-frequency representation of speech. Then in section 1.5 the strength of the non-stationarity condition is demonstrated on the more complex problem of convolutive source separation in a real acoustic environment.

1.4 Linear basis of images and sounds

In the first two sections we have argued that for non-stationary signals high kurtosis and multiple decorrelation can both be used to find linear combination of independent sources. If the basic model of linearly mixed sources does not strictly apply, however, it is not evident that the two criteria will lead to the same results. In this section we will use this to verify the modeling assumptions for two different domains, natural images and speech signals.

For images, in the past, linear bases for small spatial areas of the visual field have been compared to visual receptive fields (Olhausen and Field, 1996). Receptive fields in the auditory domain are found to be spectro-temporal patterns (Kowalski et al., 1996; deCharms and Merzenich, 1998; Theunissen and Doupe, 1998). We will therefore analyze spatial segments in images and spectro-temporal segments in speech signals.

In the following we find independent components using the simultaneous approximate joint diagonalization of cumulant matrices algorithm (JADE) Cardoso and Souloumiac, 1993. This algorithm assumes that there are non-Gaussian independent sources. On the other hand the multiple decorrelation algorithm described in the previous section assumes that there are non-stationary sources. Both algorithms assume a stationary linear mixture.

1.4.1 Spatial basis of image intensities

ICA has often been used to find a linear basis for images. This may be useful for image coding since independent components are the linear representation with minimal redundancy and maximize therefore coding efficiency (Deco and Obradovic, 1996). It has also been suggested that when applied to natural images the resulting independent bases resemble receptive fields observed in the visual cortex (Bell and Sejnowski, 1997; van Hateren and Ruderman, 1998). The high kurtosis distributions and reduced redundancy (sparseness in space and across stimuli) correspond to the sparseness reported for V1 neurons (Olhausen and Field, 1996). In fact minimum redundancy has been proposed for many years as an organizing principle of visual processing (Barlow, 1961; Atick and Redlich, 1990).

The question arises, however, if a linear basis also expresses something about the generation process of images. It has been argued that independent sources can be considered as underlying causes and that an image patch represents a linear combination of those independent causes (Bell and Sejnowski, 1997). Others have argued that occlusion is the predominant characteristic of image generation. Light in any image regions stems from a single opaque object in contradiction with the concept of a linear superposition (Ruderman, 1998).

We maintain that an intrinsic property of natural signals is non-stationarity. If that is correct, and image patches are in fact a linear combination of independent sources, ICA and our multiple decorrelation algorithm should give the same results. Otherwise one of the assump-

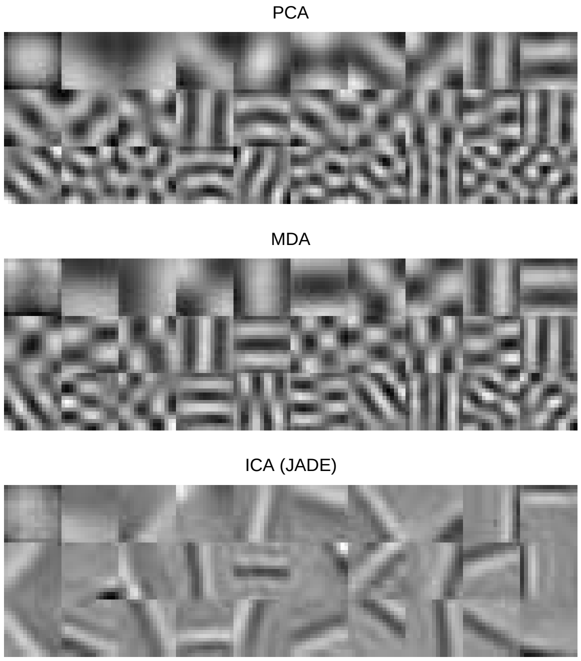

Fig. 1.1. ***Linear bases of natural images:*** *Linear bases in a 30 dimensional subspace for 15x15 image patches of natural images.*

tions has to be dismissed, i.e. images are not a linear combination of independent sources, or the sources are stationary.†

We computed for a set of natural images the linear basis with ICA and our multiple decorrelation algorithm (MDA) as described in section 1.3.2. The results are compared to those of the well-known principal components algorithm.† These components are particularly relevant here since both algorithms (JADE and MDA) use principal component analysis (PCA) to reduce the dimensionality of the problem as a first step of the processing.

Figure 1.1 shows the three basis sets obtained for the natural images used in Bell and Sejnowski, 1997. A total of 15,842 image patches of 15x15 pixels where used as input. The bases were computed in the subspace of the first 30 principal components. Separate correlations for each

† We consider the possibility of non-stationary independent components with long term Gaussian distribution to be unlikely.

† In section 1.3.2 the vectors in U with the largest eigenvalues.

image were computed for MDA. We see that components obtained using non-stationarity are not much different from the principal components and differ quite considerably from the independent components. The independent components reported here vary also from the components obtained in Bell and Sejnowski, 1997. This is because in that work a specific higher order statistics is used to find the linear components, while JADE makes no particular assumption on the statistics of the components other than non-gaussianity.

In summary, we conclude that while independent components for images may be useful for coding and for a compact representation of the data, the assumption that image patches are linear combinations of independent causes is questionable.

1.4.2 Spectro-temporal basis of speech powers

The situation is quite different for sounds. It is well known that the signal *powers* of independent acoustic signals combine additively. The linear superposition assumption for signal powers is well justified. The question is whether there are independent components and if these components are non-stationary.

In section 1.5 we will consider the case in which there are multiple sound sources in a reverberant environment and one makes multiple observations of those sources by using multiple microphones. In that case it is reasonable to assume that the sources are independent and the basic physics of acoustics indicates that *amplitudes* combine additively (barring non-linear phenomena in the microphones and amplifiers). The difficulty there however is that the linear combination is convolutive rather than instantaneous.

In this section we want to analyze the statistical properties of the *powers* of a *single source*, in particular for speech signals. We are interested in the frequency and time properties of signal powers. We will therefore look for a basis that contains spectral as well as temporal information. Guided by what is know of auditory perception we compute the frequency components on a Bark scale for short consecutive time intervals (Pinter, 1996). For computational reasons we must limit the number of bands and neighboring time slices used. We choose to find a basis for a segment of 21 Bark-scale bands and 8 neighboring time slices corresponding to 128 ms of signal between 0 and 4 kHz.† A set of

† We used half overlapping windows of 256 samples such that for a 8 kHz signal neighboring time slices are 16 ms apart.

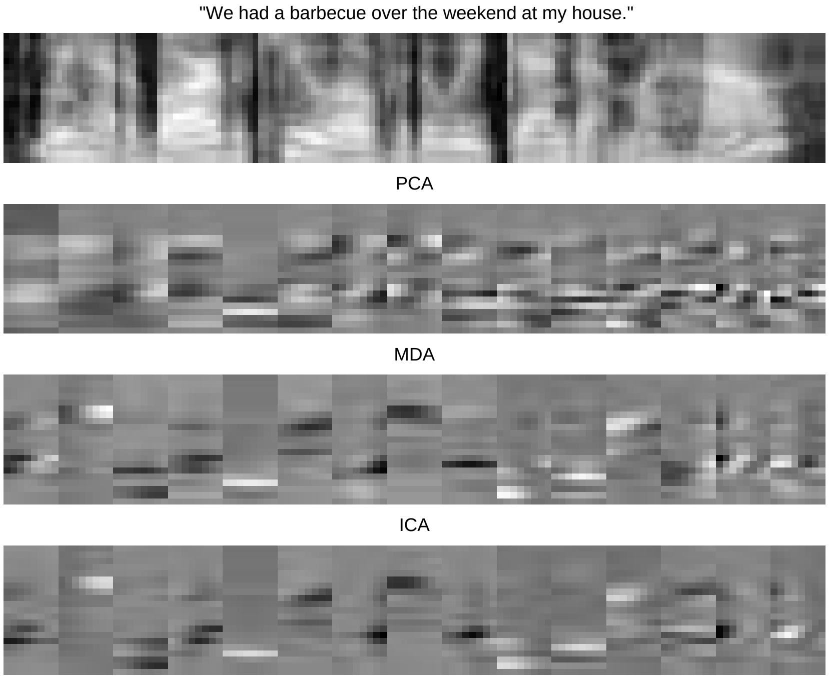

Fig. 1.2. ***Spectro-temporal linear basis representation of speech:*** *One pixel in the horizontal direction corresponds to 16 ms. In the vertical direction 21 Bark scale power bands are displayed. The upper diagram shows the log-powers for a 2.5 s segment of the 200 s recording used to compute the different linear bases. The three lower diagrams show three sets of 15 linear basis components for 21x8 spectro-temporal segments of the speech powers. The sets correspond to PCA, MDA, and ICA respectively. Note that these are not log-powers, hence the smaller contribution of the high frequencies as compared to the log-power plot on top.*

7808 such spectro-temporal segments were sampled from a 200 second recording of clean speech of a female speaker with signal to noise ratios of at least 30 dB. Figure 1.2 shows the results obtained for a subspace of 15 components. One can see that the components obtained with MDA are quite similar to the result of ICA and differ considerably from the principal components.

From this we conclude that speech powers can in fact be thought of as a linear combination of non-stationary independent components. The relevance of this result for auditory receptive fields should not be overemphasized, however. The linear superposition model applies to speech powers, while it is generally believed that auditory sensitivity scales with the logarithm of power. In that case a linear superposition is no longer correct.

These results were interesting from a theoretical point of view. The following section concentrates on an actual application in a realistic environment. The purpose for presenting this here is to demonstrate the strength of these second order methods in the case that the non-stationarity, independence, and linear superposition assumptions are strictly met. In a real acoustic environment however the mixing problem is more complicated as we have to consider convolutive rather than instantaneous mixtures.

1.5 Convolutive source separation of non-stationary signals

In a real environment, where the signals travel slowly compared to their correlation time, the instantaneous mix is not a good description of the linear superposition. The signals arrive at the different sensors with different time delays. In fact, the signals may be reflected at boundaries and arrive with multiple delays to a particular sensor. This scenario is referred to as a multi-path environment and can be described as a finite impulse response (FIR) convolutive mixture,

$$\mathbf{x}(t) = \sum_{\tau=0}^{P} A(\tau)\mathbf{s}(t-\tau) \tag{1.29}$$

How can one identify the $d_x d_s P$ coefficients of the channels A and how can one find an estimate $\hat{\mathbf{s}}(t)$ for the unknown sources? This situation is considerably more complicated than in the previous sections as one has now a matrix of filters rather than a matrix of scalars. Even once the channel has been identified, inverting it is a more difficult task as in principle the inverse should be a recursive, and therefore potentially an unstable, infinite impulse response (IIR) filter.

Alternatively one may formulate an FIR inverse model W,

$$\hat{\mathbf{s}}(t) = \sum_{\tau=0}^{Q} W(\tau)\mathbf{x}(t-\tau) \tag{1.30}$$

and try to estimate W such that the model sources $\hat{\mathbf{s}}(t) = [\hat{s}_1(t), ..., \hat{s}_{d_y}(t)]^T$ are statistically independent. Since any convolution of the individual model sources will keep the sources statistically independent this criteria specifies $\hat{\mathbf{s}}$ only up to arbitrary convolutions.

In order to simplify the notation and concentrate the discussion on the difficulties stemming from the convolution we have ignored additive noise in this section. A complete treatment including additive noise can be found in Parra and Spence, 2000a.

1.5.1 Cross-correlations, circular and linear convolution

First consider the cross-correlations $R_x(t, t+\tau) = \left\langle \mathbf{x}(t)\mathbf{x}(t+\tau)^T \right\rangle$. For stationary signals the absolute time does not matter and the correlations depend on the relative time, i.e. $R_x(t, t+\tau) = R_x(\tau)$. Denote with $R_x(z)$ the z-transform of $R_x(\tau)$. We can then write

$$R_x(z) = A(z)\Lambda_s(z)A(z)^H \tag{1.31}$$

where $A(z)$ represents the matrix of z-transforms of the FIR filters $A(\tau)$, and $\Lambda_s(z)$ are the z-transform of the auto-correlation of the sources, which again is diagonal due to the independence assumptions.

For practical purposes we have to restrict ourself to a limited number of sampling points of z. Naturally we will take T equidistant samples on the unit circle such that we can use the discrete Fourier transform (DFT). For periodic signals the DFT allows us to express circular convolutions as products such as in (1.31). However, in (1.29) and (1.30) we assumed linear convolutions. A linear convolution can be approximated by a circular convolution if $P \ll T$ and we can write approximately

$$\mathbf{x}(\omega, t) \approx A(\omega)\mathbf{s}(\omega, t), \text{ for } P \ll T \tag{1.32}$$

where $\mathbf{x}(\omega, t)$ represents the DFT of the frame of size T starting at t, $[\mathbf{x}(t), ..., \mathbf{x}(t+T)]$, and is given by $\mathbf{x}(\omega, t) = \sum_{\tau=0}^{T-1} e^{-\mathrm{i}2\pi\omega\tau}\mathbf{x}(t+\tau)$ and corresponding expressions for $\mathbf{s}(\omega, t)$, $A(\omega)$, and $W(\omega)$. In what follows time and frequency domain are identified by their argument τ or ω.

For non-stationary signals the cross-correlation will be time dependent. Estimating the cross-correlation at the desired resolution of $1/T$ is difficult if the stationarity time of the signal is in the order of magnitude of T or smaller. We are content however with any cross-correlation estimate which gives a diagonal result for the source signals. One such sample average is,

$$\hat{R}_x(\omega, t) = \frac{1}{N}\sum_{n=0}^{N-1} \mathbf{x}(\omega, t+nT)\mathbf{x}^H(\omega, t+nT) \tag{1.33}$$

We can then write for such averages

$$\hat{R}_x(\omega, t) \approx A(\omega)\Lambda_s(\omega, t)A^H(\omega) \tag{1.34}$$

If N is sufficiently large we can assume that $\Lambda_s(\omega, t)$ can be modeled as diagonal again due to the independence assumption. For equations (1.34) to be linearly independent for different times t it will be necessary that $\Lambda_s(\omega, t)$ changes over time for a given frequency, i.e. the signal are non-stationary.

1.5.2 Backward model

Given a forward model A it is not guaranteed that we can find a stable inverse. In the two dimensional square case the inverse channel is easily determined from the forward model (Weinstein et al., 1993; Thi and Jutten, 1995). However it is not apparent how to compute a stable inversion for arbitrary dimensions. Therefore we prefer to directly estimate a stable multi-path backward FIR model such as (1.30). From the condition for statistical independence of the model sources $\hat{\mathbf{s}}$ it follows that their cross-power-spectra is diagonal at all times,

$$\Lambda_s(\omega, t) = W(\omega)\hat{R}_x(\omega, t)W^H(\omega) \tag{1.35}$$

In order to obtain independent conditions for every time t_k we have to choose averaging periods for $\hat{R}_x(\omega, t_k)$ that will lead to sufficiently different second order statistics. If we set, $t_k = kTN$, we obtain non-overlapping averaging periods. Overlapping averaging times could have been chosen if the signals vary sufficiently quickly. We again compute an LS estimate of a multi-path channel W that satisfies these equations for K times simultaneously.

$$\begin{aligned} E(\omega, t_k) &= W(\omega)\hat{R}_x(\omega, t_k)W^H(\omega) - \Lambda_s(\omega, t_k) \\ \hat{W}, \hat{\Lambda}_s(..) &= \underset{\substack{W, \Lambda_s(..) \\ W(\tau) = 0, \tau > Q, \\ W_{ii}(\omega) = 1}}{\arg\min} \sum_{\omega=1}^{T} \sum_{k=1}^{K} \|E(\omega, t_k)\|^2 \end{aligned} \tag{1.36}$$

The convolution ambiguity is resolved here by fixing the diagonal terms to the unit filter. Alternatively one can also place constant time delays in the diagonal filters which is required under some microphone and user configurations (Yen and Zhao, 1999). Note also the additional time domain constraint on the filter size Q relative to the frame size T.

This condition can be satisfied by choosing short filters or alternatively larger frame sizes T. Up to that constraint it would seem the various frequencies $\omega = 1, ..., T$ represent independent problems. However the solutions $W(\omega)$ are restricted to those filters that have zero time response for $\tau > Q \ll T$. Effectively we are parameterizing Td_sd_x filter coefficients in $W(\omega)$ with Qd_sd_x parameters $W(\tau)$. Due to this constraint we are forced to use a gradient algorithm to find the LS solutions and can no longer use analytic solutions such as in the instantaneous mixture case. We will first compute the gradients with respect to the complex valued filter coefficients $W(\omega)$ and discuss their projections onto the subspace of permissible solutions in the following section. The gradients † of the LS cost in (1.36) are

$$\frac{\partial J}{\partial W^*(\omega)} = 2\sum_{k=1}^{K} E(\omega, t_k) W(\omega) \bar{R}_x(\omega, t_k) \tag{1.37}$$

We can find the minimum with respect to $W(\omega)$ with a constrained gradient descent algorithm using the gradients (1.37). The optimal $\Lambda_s(\omega, t_k)$ for given $W(\omega)$ at every gradient step can be computed explicitly by setting the gradient in with respect to $\Lambda_s^*(\omega, t_k)$ to zero, which yields $\hat{\Lambda}_s(\omega, t_k) = \text{diag}\left[W(\omega)\hat{R}_x(\omega, t_k)W^T(\omega)\right]$.

The algorithm described so far uses all of the data to be filtered in order to find the optimal separating filter matrix. Only after that can the data be filtered. In many realistic scenarios an on-line algorithm is required, whereby the filter is immediately applied to the data and relatively little data can be stored. An efficient on-line version of the batch gradient algorithm presented here is given in (Parra and Spence, 2000b). Though a more rigorous derivation can be given it basically amounts to removing the sum over times t_k. This converts the exact gradient into a stochastic gradient with updates $\Delta_{t_k} W(\omega)$ given by

$$\Delta_{t_k} W(\omega) = 2\mu(\omega, t_k) E(\omega, t_k) W(\omega) \hat{R}_x(\omega, t_k)\,. \tag{1.38}$$

The current cross-power spectra $\hat{R}_x(\omega, t_k)$ is estimated as a running average. In order to improve the convergence speed of the on-line algorithm we propose in (Parra and Spence, 2000b) an variable learning rate

† For any real valued function $f(\mathbf{z})$ of a complex valued variable $\mathbf{z}$ the gradients with respect to the real and imaginary part are obtained by taking derivatives formally with respect to the conjugate quantities $\mathbf{z}^*$, ignoring the non-conjugate occurrences of $\mathbf{z}$, i.e., $\frac{\partial f(z)}{\partial \Re(z)} + i\frac{\partial f(z)}{\partial \Im(z)} = 2\frac{\partial f(z)}{\partial z^*}$ (Brandwood, 1983; Jänich, 1977) .

$\mu(\omega, t_k)$, which is motivated by second derivatives of the cost function.

$$\mu(t,\omega)^{-1} = \sum_j \frac{\partial^2 \|E(t,\omega)\|^2}{\partial W^*_{ij}(\omega)\partial W_{ij}(\omega)} = 2\|W(\omega)\hat{R}_x(t,\omega)\|^2 \,. \qquad (1.39)$$

This effectively amounts to an adaptive power normalization in each frequency bin. In our experiments the resulting updates were stable and lead to convergence after processing only a few seconds of data.

1.5.3 Permutations and constraints

Note that arbitrary permutations of the coordinates for each frequency ω will lead to the same error $E(\omega, t_k)$. Therefore the total cost will not change if we choose a different permutation of the solutions for each frequency ω. This seems to be a serious problem since only consistent permutations for all frequencies will correctly reconstruct the sources.

Arbitrary permutations, however, will not satisfy the condition on the length of the filter, $W(\tau) = 0$ for $\tau > Q \ll T$. Effectively, requiring zero coefficients for elements with $\tau > Q$ will restrict the solutions to be continuous or "smooth" in the frequency domain, e.g., if $Q/T = 8$ the resulting DFT corresponds to a convolved version of the coefficients with a sinc function 8 times wider than the sampling rate.

We can enforce the filter size constraint by projecting the unconstrained gradients (1.37) to the subspace of permissible solutions. The proper projection is implemented by transforming the gradient into the time domain, zeroing all components with $\tau > Q$, and transforming back to the frequency domain. The unit gain constraint on diagonal filters is simply enforced by keeping the filter coefficient constant to $W_{ii}(\omega) = 1$.

The constraint on the filter size Q versus the frequency resolution $1/T$ links the otherwise independent frequencies, and picks a particular permutation for the frequency permutation problem. In addition, it is a necessary condition for equations (1.35) to hold to a good approximation. Note also that it does not limit the actual filter size, as in principle one can choose an appropriately large frame size T for any given Q.

As we will see in the next section the current continuity condition on the filters gives acceptable performance in a variety of configurations. More recently however we have established that this constraint in fact may not be appropriate for all circumstances (see also work by Ikram and Morgan, 2000). In principle, there is no theoretical argument why smooth filters will give the appropriate separation filters. In fact evidence to the contrary may exist (Liavas and Regalia, 1998). Selecting appropriate permutations remains a subject of current research.

1.5.4 Separation performance in real room environments

We have applied the algorithm presented in the previous section in a variety of situations. The following performance results are given as the ratio between the power of the signals and the power of the remaining cross-talk, which is commonly referred to as the Signal to Interference Ratio (SIR). The separation performance varies depending on the particular configuration. We obtained an improvement of the SIR of anywhere between 0 to 18 dB. We studied the dependence on the type and number of microphones, the number of sources, the user and microphone locations, the size of the room, the size of the filter, and other algorithm parameters.

A first example on signals that are publicly available is shown in figure 1.3. The graph shows the results for varying filter sizes on the separation of two competing speakers recorded with two microphones. The improvement in SIR can be as high as 15 dB for recordings obtained in an office room using uni-directional (cardioid) microphones (upper curve). Separating two speakers from the recordings in a second room with omni-directional microphones seems more challenging (lower curve)†. As expected, the performance initially increases with increasing filter size, as the inverse of the room can be modeled more accurately. However, larger filters may require more training data, and so the performance eventually decreases given the constant amount of data.

We observed in further experiments that separation works better in large conference rooms than in small office rooms with stronger reflecting walls, most likely due to the increased reverberation. This was confirmed with simulated environments of varying size (Parra and Spence, 2000a)

The SIRs in Figs. 1.3 do not change smoothly, which may be explained by the fact that the algorithm is not optimizing the SIR directly but instead multiple decorrelations. Also, the gradient algorithm may be reaching different local minima of the diagonalization criterion.

Another interesting question is how the performance improves if we use additional microphones, given a constant number of sources. Figure 1.4 shows the performance in separating two simultaneous speakers using

† The data for the first example is available from Parra, 1998 and for the second example from Schoebben, 1998

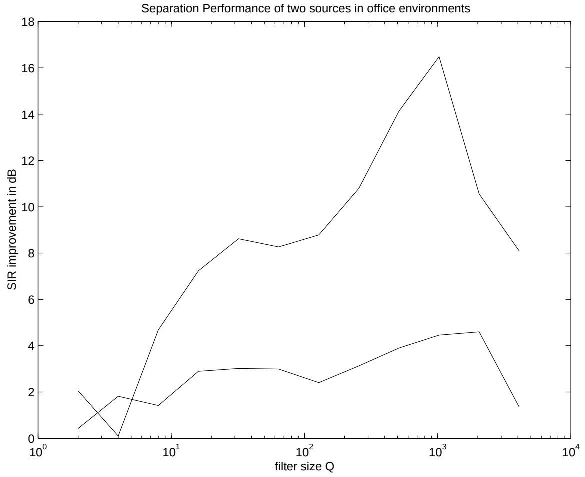

Fig. 1.3. ***Separation in real room:*** *Performance for two speakers recorded with two microphones in two different office environments as a function of separation filter size* Q *and* $T/Q = 8$*. Upper curve: uni-directional microphones in a* $3\,m \times 3.6\,m \times 2.3\,m$ *room, 30 s recordings at 8 KHz, 15 s alternating and 15 s simultaneous speech. Lower curve: 10 s simultaneous speech recorded at 16 KHz in a* $4.2\,m \times 5.5\,m \times 3.1\,m$ *room with omni-directional microphones.*

a variable number of microphones. In this experiment we used 8 cardioid condenser microphones arranged as an equidistant linear array of about 65 cm length. The results are compared to the SIR of simple broadside and end-fire beams constructed using all 8 microphones. While one user is at about 230 cm distance directly at broadside the other user is located at different angles relative to the broad-side and at different distances from the array. The data was sampled at 16 kHz with 10s for training and 10s for testing and we used $K = 5, Q = 1024, T = 4092$.

Finally we want to note that the implementation of this algorithm in C runs in real time on a 155 kHz Intel Pentium processor for a 2 input, 2 sources problem at 8 kHz sampling rate and T=2048.

Bibliography

Atick, J. and Redlich, A. (1990). Towards a theory of early visual processing. *Neural Computation*, 2:308–320.

Bar-Ness, Y., Carlin, J., and Steinberger, M. (1982). Bootstrapping adaptive cross-pol canceller for satellite communications. In *Proc. IEEE Int. Conf. Comunications*, pages 4F.5.1–4F.5.5.

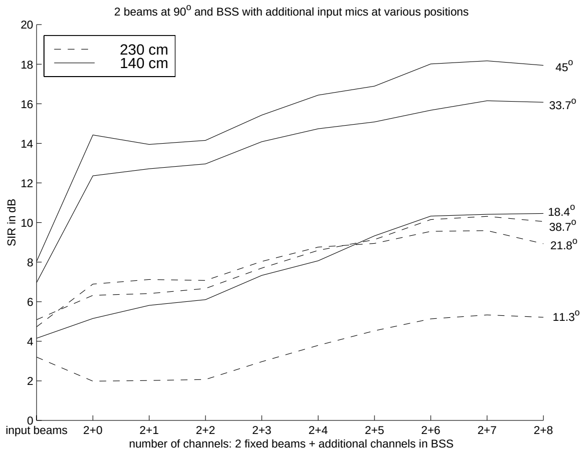

Fig. 1.4. ***Improvement with increased number of microphones:*** *Performance for the separation of two sources as the number of microphones increases. Broadside and end-fire beams were constructed as two input channels. Besides that a variable number of the microphones were used as additional input channels. The recordings where performed in a moderately reverberant room 422 cm x 372 cm.*

Barlow, H. (1961). The coding of sensory messages. In *Current Problems in Animal Behavior*. Cabridge University Press, Cambridge.

Beale, E. and Mallows, C. (1959). Scale mixing of symmetric distributions with zero means. *Annals of Mathematical Statitics*, 30:1145–1151.

Bell, A. and Sejnowski, T. (1997). The independent components fo natural scenes are edge filters. *Vision Research*, 23:3327–3338. images in ftp://ftp.cnl.salk.edu/pub/tony/VRimages/.

Belouchrani, A., Meraim, A., Cardoso, J.-F., and Moulines, E. (1993). Second order blind separation of correlated sources. In *Proc. Int. Conf. on Digital Signal Processing*, pages 346–351, Cyprus.

Bradwood, D. (1978). Cross-coupled cancellation systems for improving cross-polarisation discrimination. In *Proc. IEEE Int. Conf. Antennas and Propagation*, volume I, pages 41–45.

Brandwood, D. (1983). A complex gradient operator and its application in adaptive array theory. *IEE Proc.*, 130(1):11–16.

Brehm, H. and Stammler, W. (1987). Description and generation ofspherically invariant speech-model signals. *Signal Processing*, 12:119–141.

Cardoso, J.-F. (1999). High-order contrasts for independent component analysis. *Neural Computation*, 11:157–192.

Cardoso, J.-F. and Souloumiac, A. (1993). Blind beamforming for non gaussian signals. *IEE Procedings F*, 140(6):362–370.

Cardoso, J.-F. and Souloumiac, A. (1996). Jacobi angles for simultaneous diagonalization. *SIAM J. Mat. Anal. Appl.*, 17(1):161–164.

Comon, P. (1994). Independent comonent analysis, a new concept? *Signal Processing*, 36(3):287–314.

deCharms, Christopher, R. and Merzenich, Miachael, M. (1998). Characteristic neuros in the primary auditory cortex of the awake primate using reverse correlation. In *Advances in Neural Information Processing Systems 10*, pages 124–130.

Deco, G. and Obradovic, D. (1996). *An Information Theoretic Approach to Neural Computing*. Perspective in Neural Computing. Springer.

Fety, L. and Van Ulffelen, J. (1988). New methods for signal separation. In *Proc. of 4th Int. Conf. of HF radio systems and techniques*, pages 226–230, London. IEE.

Goldman, J. (1976). Detection in the presence of spherically symmetric random vectors. *IEEE Transactions on Information Theory*, 22(1):52–59.

Ikram, M. and Morgan, D. (2000). Exploring permutattion inconsistency in blind separation of speech signals in a reverberant environment. In *ICASSP 2000*.

Jänich, K. (1977). *Einführung in die Funktionentheorie*. Springer-Verlag. ch. 2.

Kawamoto, M., Matsuoka, K., and Ohnishi, N. (1998). A method of blind separation for convolved non-stationary signals. *Neurocomputing*, 22:157–171.

Kendal, M. and Stuart, A. (1969). *The Advanced Theory of Statistics*. Charles Griffin & Company Limited, London.

Kowalski, N., Depireux, D., and Shamma, S. (1996). Analysis of dynamic spectra in ferret primary auditory cortex: I. characteristics of single unit responses to moving ripple spectra. *J.Neurophys.*, 76(5):3503–3523.

Liavas, A. P. and Regalia, P. A. (1998). Acoustic echo cancellation: Do iir models offer better modeling capabilities than their fir counterparts? *IEEE Transactions on Signal Processing*, 46:2499–2504.

Matsuoka, K., Ohya, M., and Kawamoto, M. (1995). A neural net for blind separation of nonstationary signals. *Neural Networks*, 8(3):411–419.

Molgedey, L. and Schuster, H. G. (1994). Separation of a mixture of independent signals using time delayed correlations. *Phys. Rev. Lett.*, 72(23):3634–3637.

Olhausen, B. and Field, D. (1996). Emergence of simple-sell receptive field properties by learning sparce code for natural images. *Nature*, 381:607–609.

Parra, L. (1998). Blind source separation based on multiple decorrelations. http://www.sarnoff.com/career_move/tech_papers/BSS.html.

Parra, L. and Spence, C. (2000a). Convolutive blind source separation of non-stationary sources. *IEEE Trans. on Speech and Audio Processing*, pages 320–327.

Parra, L. and Spence, C. (to be printed in summer 2000b). On-line convolutive blind source separation of non-stationary sources. *VLSI Signal Processing Journal*.

Pinter, I. (1996). Perceptual wavelet-representation of speech signals and its application to speech enhancement. *Computer Speech and Language*, 10:1–22.

Rangaswamy, M., Weiner, D., and Oeztuerk, A. (1993). Non-gaussian random vector identification using spherically invariant random processes. *IEEE Transaction on Aerospace and Electronic Systems*, 29(1):111–123.

Ruderman, Daniel, L. (1998). Origins of scaling in natural images. *Vision Research*, 37(23):3385–3398.

Rupp, M. (1993). The behavior of lms and nlms algorithms in the presence of spherically invariant processes. *IEEE Transaction on Signal Processing*, 41(3):1149–1160.

Schießl, I., Stetter, M., Mayhew, J. E. W., McLoughlin, N., Lund, J. S., and Obermayer, K. (May 2000). Blind signal separation from optical imaging recordings with extended spatial decorrelation. *IEEE Trans. Biomed. Engin.*, 47:in press.

Schoebben, D. (1998). Real room recordings and separation results. http://www.esp.ele.tue.nl/onderzoek/daniels/BSS.html.

Souloumiac (1995). Blind source detection and separation using second order non-stationarity. In *International Conference on Acoustics, Speech and Signal Processing*, volume IEEE 0-7803-2431-5/95, pages 1912–1915.

Spence, C. and Parra, L. (2000). Hirarchical image probability (hip) model. In *Advances in Neural Information Processing Systems 12*. MIT Press.

Tang, A. C., Pearlmutter, B., Zibulevsky, M., and Loring, R. (2000). An meg study of responce latency and variability in the human visual system during a visual-motor integration task. In *Advances in Neural Information Processing Systems 12*, pages 185–191. MIT Press.

Theunissen, F. and Doupe, A., J. (1998). Temporal and spectral sensitivity of auditory neurons in the nucleus hvc of male zebra finches. *J. Neuroscience*, 18(10):3786–3802.

Thi, H.-L. N. and Jutten, C. (1995). Blind source separation for convolutive mixtures. *Signal Processing*, 45(2):209–229.

Tong, L. and Liu, R. (1990). Blind estimation of correlated source signals. In *Proc. Asilomar Conference*.

Van Gerven, S. and Van Compernolle, D. (1995). Signal separation by symmetric adaptive decorrelation: Stability, convergence, and uniqueness. *IEEE Trans. Signal Processing*, 43(7):1602–1612.

van Hateren, J. and Ruderman, D. (1998). Independent component analysis of image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proc.R.Soc.Lond. B*, 265:2315–2320.

Wainwright, M. J. and Simoncelli, E. P. (2000). Scale mixtures of gaussians and the statistics of natural images. In *Advances in Neural Information Processing Systems 12*. MIT Press.

Weinstein, E., Feder, M., and Oppenheim, A. (1993). Multi-channel signal separation by decorrelation. *IEEE Trans. Speech Audio Processing*, 1(4):405–413.

Wübbeler, G., Ziehe, A., Mackert, B.-M., Müller, K.-R., Trahms, L., and Curio, G. (2000). Independent component analysis of non-invasively recorded cortical magnetic dc-fields in humans. *IEEE Trans Biomed Eng.*, 47(5). (in press).

Yen, K.-C. and Zhao, Y. (1999). Adaptive co-channel speech separation and recognition. *IEEE Trans. Signal Processing*, 7(2).

Ziehe, A., Müller, K.-R., Nolte, G., Mackert, B.-M., and Curio, G. (2000). Artifact reduction in magnetoneurography based on time-delayed second-order correlations. *IEEE Trans Biomed Eng.*, 47(1):75–87.