GRADIENT FLOW CONVOLUTIVE BLIND SOURCE SEPARATION

Michael Syskind Pedersen*
Informatics and Mathematical Modelling, Building 321
Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark
Phone: +45 4525 3904
E-mail: msp@imm.dtu.dk
Web: imm.dtu.dk/~msp

Chlinton Møller Nielsen
Technology & Innovation, Bang & Olufsen
Peter Bangs Vej 15, DK-7600 Struer, Denmark
Phone: +45 9684 4058.
Email: chn@bang-olufsen.dk

Abstract. Experiments have shown that the performance of instantaneous gradient flow beamforming by Cauwenberghs et al. is reduced significantly in reverberant conditions. By expanding the gradient flow principle to convolutive mixtures, separation in a reverberant environment is possible. By use of a circular four-microphone array with a radius of 5 mm, and applying convolutive gradient flow instead of just applying instantaneous gradient flow, experimental results show an improvement of up to around 14 dB can be achieved for simulated impulse responses and up to around 10 dB for a hearing aid application with real impulse responses.

INTRODUCTION

The *gradient flow* blind source separation technique proposed by Cauwenberghs et al. [5] uses a four microphone array to separate 3 sound signals. The gradient flow can be regarded as a preprocessing step in order to enhance the difference between the signals before a blind separation algorithm is applied. The gradient flow technique requires small array sizes. Small array sizes occur in some source separation applications such as hearing aids. Here the physical dimensions of the microphone array may limit the separation performance due to the very small difference between the recorded signals. In the literature, some attempts exist to separate sound signals by use of a microphone arrays with a dimension of about 1 cm [2, 6, 7]. These techniques

*This work was supported by the Oticon Foundation

are either based on beamforming, blind source separation [3], or a combination of these techniques. The gradient flow method is able to estimate delayed versions of the source signals, as well as the source arrival angles. As shown in the simulations, the model may fail in reverberant environments, i.e. when each of the source signals is convolved in time. Here, a model is proposed that extends the instantaneous gradient flow model to a convolutive gradient flow model. Simulations show that the convolutive model is able to cope with reverberant situations, in which the instantaneous model fails.

INSTANTANEOUS GRADIENT FLOW MODEL

The gradient flow model is described into details in [5, 8, 10, 11]. Each signal x_{pq} is received by a sensor placed at location (p, q), which is shown in Figure 1. At a point in the coordinate system $\mathbf{r}$, there is a delay, $\tau(\mathbf{r})$, between an incoming wavefront and the origin. The delay with respect to the n'th source signal, s_n is denoted as $\tau^n(\mathbf{r})$. It is assumed that the sources are located in the far-field. Hence the wavefront of the incoming waves is linear. Using that assumption the delay can be described the following way [5]:

$$\tau(\mathbf{r}) \approx \frac{1}{c}\mathbf{r}\cdot\mathbf{u}, \tag{1}$$

where $\mathbf{u}$ is a unit vector pointing in the direction of the source and c is the velocity of the wave.
Now consider a sensor placed at the coordinates (p, q) as in Figure 1. The time delay from the source can be expressed as

$$\tau_{pq}^n = p\tau_1^n + q\tau_2^n, \tag{2}$$

where $\tau_1^n = \mathbf{r}_1 \cdot \mathbf{u}_n/c$ and $\tau_2^n = \mathbf{r}_2 \cdot \mathbf{u}_n/c$. τ_1^n and τ_2^n are the time differences in the directions of the two orthogonal vectors $\mathbf{r}_1$ and $\mathbf{r}_2$ as shown in Figure 1. The point $\mathbf{r}_{pq}$ can be described as $\mathbf{r}_{pq} = p\mathbf{r}_1 + q\mathbf{r}_2$.

Description of field

The field is described by the incoming waves. At the center of the coordinate system, the contribution to the field from the n'th source is given by $s_n(t)$. By using the Taylor series expansion, the field from the n'th source at the point $\mathbf{r}$ in the coordinate system is given by $s_n(t + \tau^n(\mathbf{r}))$, where [11]

$$s_n(t + \tau^n(\mathbf{r})) = s_n(t) + \frac{1}{1!}\tau^n(\mathbf{r})\dot{s}_n(t) + \frac{1}{2!}(\tau^n(\mathbf{r}))^2\ddot{s}_n(t) + \dots. \tag{3}$$

Here, $\dot{}$ and $\ddot{}$ denote the 1'st and 2'nd order derivative, respectively. Hence, the received signal at $\mathbf{r}_{pq}$ can be written as

$$x_{pq}(t) = \sum_{n=1}^{N} s_n(t + \tau^n(\mathbf{r})) = \sum_{n=1}^{N} s_n(t) + \frac{1}{1!}\tau^n(\mathbf{r})\dot{s}_n(t) + \frac{1}{2!}(\tau^n(\mathbf{r}))^2\ddot{s}_n(t) + \dots. \tag{4}$$

Additionally, a noise term $\varepsilon_{pq}(t) \propto N(0, \sigma)$ can be added [5]. The received signal can be approximated by using only the first two terms of (4):

$$x_{pq}(t) = \sum_{n=1}^{N} s_n(t + \tau^n(\mathbf{r})) \approx \sum_{n=1}^{N} s_n(t) + \tau^n(\mathbf{r})\dot{s}_n(t). \tag{5}$$

Notice, the Taylor approximation only holds, if the dimension of the array is not too large (see [5] for details).

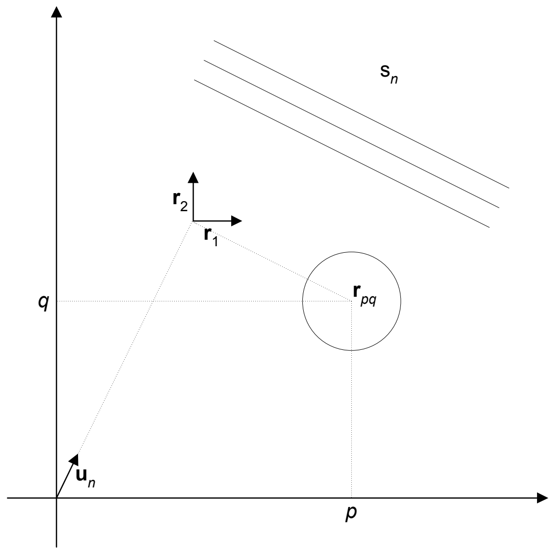

Figure 1: Sensor placed at the point $\mathbf{r}$ with the position coordinates (p, q) so that the point is described the following way: $\mathbf{r}_{pq} = p\mathbf{r}_1 + q\mathbf{r}_2$, where $\mathbf{r}_1$ and $\mathbf{r}_2$ are orthogonal vectors. The time delay between (p, q) and the origin with respect to the n'th source signal is denoted as τ^n_{pq}.

Gradient Flow

The spatial derivatives along the position coordinates (p, q) around the origin in the coordinate system are found of various orders (i, j) [11].

$$\xi_{ij}(t) \equiv \frac{\partial^{i+j}}{\partial^i p \partial^j q} x_{pq}(t) \mid_{p=q=0} \tag{6}$$

$$= \sum_{n=1}^{N} (\tau^n_1)(\tau^n_2) \frac{d^{i+j}}{d^{i+j}t} s_n(t) \tag{7}$$

Additionally, the derivative of the sensor noise $\nu_{ij}(t)$ may be added. Corresponding to (5), the 0'th and 1'st order terms yield:

$$\xi_{00}(t) = \sum_n s_n(t) \tag{8}$$

$$\xi_{10}(t) = \sum_n \tau_1^n \frac{ds^n(t)}{dt} = \sum_n \tau_1^n \dot{s}_n(t) \tag{9}$$

$$\xi_{01}(t) = \sum_n \tau_2^n \frac{ds^n(t)}{dt} = \sum_n \tau_2^n \dot{s}_n(t) \tag{10}$$

The estimates of the 0th order term $\xi_{00}(t)$, i.e. the estimate of the field in the origin, can be obtained from the sensors as the average of the signals since the sensors are symmetrically distributed around the origin at the four coordinates (0,1), (1,0), (0,-1) and (-1,0):

$$\xi_{00}(t) \approx \frac{1}{4}(x_{-1,0} + x_{1,0} + x_{0,-1} + x_{0,1}). \tag{11}$$

The estimates of the two 1'st order derivatives can as well be estimated from the sensors:

$$\xi_{10}(t) = \frac{\partial x}{\partial p} \approx \frac{\Delta x}{\Delta p} = \frac{x_{1,0} - x_{-1,0}}{1-(-1)} = \frac{1}{2}(x_{1,0} - x_{-1,0}) \tag{12}$$

$$\xi_{01}(t) = \frac{\partial x}{\partial q} \approx \frac{\Delta x}{\Delta q} = \frac{x_{0,1} - x_{0,-1}}{1-(-1)} = \frac{1}{2}(x_{0,1} - x_{0,-1}) \tag{13}$$

By taking the time derivative of $\xi_{00}(t)$, the following equation can be obtained.

$$\frac{d}{dt}\xi_{00}(t) = \sum_{n=1}^{N} \frac{d}{dt} s_n(t) \tag{14}$$

Thus, the following instantaneous linear mixture can be obtained.

$$\begin{bmatrix} \dot{\xi}_{00}(t) \\ \xi_{10}(t) \\ \xi_{01}(t) \end{bmatrix} \approx \begin{bmatrix} 1 & \cdots & 1 \\ \tau_1^1 & \cdots & \tau_1^N \\ \tau_2^1 & \cdots & \tau_2^N \end{bmatrix} \begin{bmatrix} \dot{s}_1(t) \\ \vdots \\ \dot{s}_N(t) \end{bmatrix} \tag{15}$$

This equation is of the type $\mathbf{x} = \mathbf{A}\mathbf{s}$, where only $\mathbf{x}$ is is known. Assuming that the source signals $\mathbf{s}$ are independent, (15) can be solved by independent component analysis (see e.g. [3]).

EXTENSION TO CONVOLUTIVE MIXTURES

As mentioned in [5], the instantaneous model (15), may be extended to convolutive mixtures. In Figure 2, a situation is shown in which each source signal does not only arrive from a single direction. Here, reflections of each

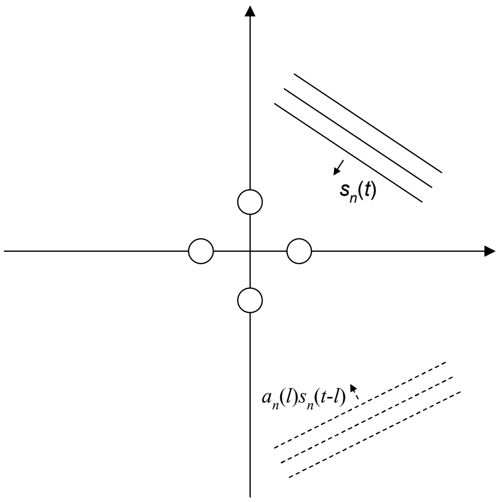

Figure 2: At the time t, a signal $s_n(t)$ originating from source n is arriving at the sensor array. At the same time, reflections from the same source arrive from other directions. These reflections are attenuated by the factor a_n and delayed by the time lag l. Each signal received at the sensor array are therefore convolved mixtures of the original source signals. For simplification, only a single source and a single reflection is shown.

source signal may be present too. Each reflection is delayed by a factor l and attenuated by an attenuated by a factor $a_n(l)$. Now, similarly to (4) the received signal x_{pq} at the sensor at position (p, q) is described as

$$x_{pq}(t) = \sum_{n=1}^{N} \sum_{l=0}^{L} a_n(l) s_n(t + \tau^n(\mathbf{r}, l) - l), \tag{16}$$

where L is the assumed maximum time delay. Using the Taylor expansion, each received mixture can be written as

$$x_{pq}(t) = \sum_{n=1}^{N} \sum_{l=0}^{L} a_n(l) \big[s_n(t-l) + \tau^n(\mathbf{r}, l) \dot{s}_n(t-l) + \frac{\tau^n(\mathbf{r}, l)}{2} \ddot{s}_n(t-l) + \dots \big] \tag{17}$$

Using only the two first terms of the Taylor expansion and inserting $\tau^n(\mathbf{r}_{pq}, l) = p\tau_1^n(l) + q\tau_2^n(l)$, (17) can be written as

$$x_{pq}(t) \approx \sum_{n=1}^{N} \sum_{l=0}^{L} a_n(l) \big[s_n(t-l) + (p\tau_1^n + q\tau_2^n) \dot{s}_n(t-l) \big]. \tag{18}$$

Similar to the instantaneous mixture case, the spatial derivatives of the convolutive mixture can be found from (6). The 0'th order and the 1'st order derivatives are then similarly to (8)–(10).

$$\xi_{00}(t) \quad = \quad \sum_{n} \sum_{l} a_n(l) s_n(t-l) \tag{19}$$

$$\xi_{10}(t) = \sum_n \sum_l a_n(l)\tau_1^n(l)\frac{ds^n(t-l)}{dt} = \sum_n \sum_l a_n(l)\tau_1^n(l)\dot{s}_n(t-l) \quad (20)$$

$$\xi_{01}(t) = \sum_n \sum_l a_n(l)\tau_2^n(l)\frac{ds^n(t-l)}{dt} = \sum_n \sum_l a_n(l)\tau_2^n(l)\dot{s}_n(t-l) \quad (21)$$

The time derivative of $\xi_{00}(t)$ is expressed as

$$\dot{\xi}_{00}(t) = \sum_{n=1}^{N} \sum_{l=0}^{L} a_n(l)\dot{s}_n(t-l). \quad (22)$$

By expressing (22), (20) and (21) with matrix notation, the following expression can be obtained:

$$\begin{bmatrix} \dot{\xi}_{00}(t) \\ \xi_{10}(t) \\ \xi_{01}(t) \end{bmatrix} \approx \sum_{l=0}^{L} \begin{bmatrix} a_1(l) & \cdots & a_N(l) \\ a_1(l)\tau_1^1(l) & \cdots & a_N(l)\tau_1^N(l) \\ a_1(l)\tau_2^1(l) & \cdots & a_N(l)\tau_2^N(l) \end{bmatrix} \begin{bmatrix} \dot{s}_1(t-l) \\ \vdots \\ \dot{s}_N(t-l) \end{bmatrix} \quad (23)$$

$$= \begin{bmatrix} a_1(l) & \cdots & a_N(l) \\ a_1(l)\tau_1^1(l) & \cdots & a_N(l)\tau_1^N(l) \\ a_1(l)\tau_2^1(l) & \cdots & a_N(l)\tau_2^N(l) \end{bmatrix} * \begin{bmatrix} \dot{s}_1(t) \\ \vdots \\ \dot{s}_N(t) \end{bmatrix}, \quad (24)$$

where $*$ is the convolution operator. This is a convolutive mixture problem of the well-known type $\mathbf{x} = \mathbf{A} * \mathbf{s}$, where only an estimate of $\mathbf{x}$ is known. These estimates are found similarly to the instantaneous case from (11)–(13).

FREQUENCY DOMAIN SEPARATION

In [8], the *Jade* algorithm [4] was successfully applied to solve the instantaneous mixing ICA problem (15). The *Jade* algorithm is based on joint diagonalization of 4'th order cumulants. In order to solve the convolutive mixing problem (23), the problem is transformed into the frequency domain [9]. Hereby, the convolution in the time domain can be approximated by multiplications in the frequency domain, i.e. for each frequency bin,

$$\xi(f, m) \approx \mathbf{A}(f)\dot{\mathbf{s}}(f, m), \quad (25)$$

where m denotes the index of the frame of which the short-time Fourier transform STFT is calculated. f denotes the frequency. When solving the ICA problem in the frequency domain, different permutations for each frequency band may occur. In order to solve the frequency permutations, the method suggested in [1] has been used. It is assumed that the mixing matrices in the frequency domain will be smooth. Therefore, the mixing matrix at frequency band k, $\mathbf{A}(f_k)$ is compared to the mixing matrix at band $k-1$, $\mathbf{A}(f_{k-1})$. This is done by calculating the distance between any possible permutations of $\mathbf{A}(f_k)$ and $\mathbf{A}(f_{k-1})$, i.e.

$$D(p) = \sum_{i,j} |a_{ij}^{(p)}(f_k) - a_{ij}(f_{k-1})|, \quad (26)$$

Where p represents the p'th permutation. The permutation which yields the smallest distance is assumed to be the correct permutation. Notice, for an $N \times N$ mixing matrix, there are $N!$ different permutations. Therefore this method becomes slow for large N. For a 3×3 mixing matrix there are only six possible permutations.

EXPERIMENTS

Signals with synthetic impulse responses

Three speech sentences have been artificially mixed – two female speakers and one male speaker. The duration of each speech signal is 10 seconds, and the speech signals have a sampling frequency of 20 kHz. A demonstration of separated sounds is available at `wwm.imm.dtu.dk/~msp`. The microphone array consists of four microphones. These are placed in a horizontal plane. An application for such a microphone array is shown in Figure 3, where the four microphones are placed in a single hearing aid. Here, the distance between the microphones and the center of the array is 5 mm. By use of the gradient flow method, it is possible to separate up to three sources [8]. If there are more than three sources, an enhancement of the signals may be achieved even though full separation of all sources isn't possible. In the first experiment, a convolutive mixture of the three sources is simulated. The arrival angles as well as the attenuation factor of the reverberations have been chosen randomly. The maximum delay in this experiment has been chosen to 25 samples. No sensor noise is added. The differentiator has been chosen to be a 1000 order FIR differentiator estimated with a least squares approach (even though a smaller order could be sufficient). The integrator is implemented as a first order Alaoui IIR filter as in [8]. Here, all 200000 samples have been used to estimate the separated sounds. In order to achieve on-line separation, the separated sounds may be estimated using blocks of shorter duration [8]. The instantaneous Jade performs well if only the direct sounds are present, but if reverberations are present too, the separation performance is significantly reduced. The signal to interference ratio improvement is calculated as

$$\Delta\mathrm{SIR}(i) = 10\log\left(\frac{\langle (y_{i,s_i})^2\rangle}{\langle (\sum_{j\neq i} y_{i,s_j})^2\rangle}\right) - 10\log\left(\frac{\langle (x_{10,s_i})^2\rangle}{\langle (\sum_{j\neq i} x_{10,s_j})^2\rangle}\right), \qquad (27)$$

Here, y_{i,s_j} is the i'th separated signal, where only the the j'th of the original signals has been sent through the mixing and unmixing system. x_{10,s_i} is the recorded signal at the microphone at position (1,0) with only the i'th source signal active. $\langle\cdot\rangle$ denotes the expectation over all samples.

The ΔSIR has been found for different DFT lengths as well as the case, where the instantaneous Jade has been been applied to the convolutive mixture. Hamming windows of the same length as the DFT has been used. An STFT overlap of 75% has been used. Table 5.1 shows the separation results of the convolutive mixture. As it can be seen, the length of the DFT should be at

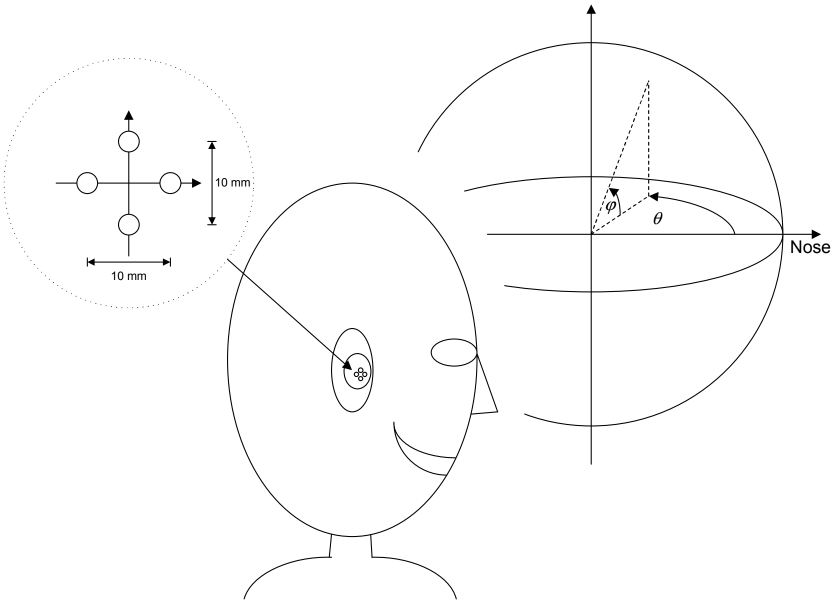

Figure 3: Four microphones are placed in a hearing aid. The distance between the microphones and the center of the array is 5 mm. By using such a configuration, it is possible to separate up to three independent sound sources. The azimuth angle, θ is defined according to the figure so that 0° is the direction of the nose. Likewise, the elevation angle φ is defined according to the figure so that 0° corresponds to the horizontal plane. Both angles increase in the counterclockwise direction.

least 256, in order to separate all three sources. It can as well be seen that, the FIR separating filters have to be significantly longer than the mixing filters in order to ensure separation.

Table 1: Three synthetic, artificially mixed speech signals have been separated. The maximum delay of each convolutive mixture is 25 samples. The arrival elevation angles (φ) and the azimuth (θ) angles of the direct sounds are given. The ΔSIR have been found for the instantaneous case and for different DFT lengths. The best separation is achieved with a DFT length of 256 or 512.

	UK Male	UK female	DK female
θ	0°	-112.5°	-157.5°
φ	0°	-21°	14°
Instantaneous JADE	9.5 dB	2.4 dB	2.5 dB
DFT length=64	10.2 dB	2.4 dB	14.2 dB
DFT length=128	11.0 dB	0.5 dB	11.5 dB
DFT length=256	9.0 dB	9.2 dB	14.6 dB
DFT length=512	8.9 dB	8.5 dB	16.5 dB
DFT length=1024	6.5 dB	8.7 dB	16.2 dB

Table 2: Signals generated from real impulse responses recorded by a four-microphone array placed in the right ear of a head and torso simulator inside an anechoic room. No noise has been added. Here, the "UK female" is the hardest sound to separate. When listening to the sounds, all of them seems to be separated. When the DFT becomes too long, the separation decreases. One explanation could be that the attempt to solve the permutation ambiguity fails.

	UK Male	UK female	DK female
θ	$0°$	$-112.5°$	$-157.5°$
φ	$0°$	$-21°$	$14°$
Instantaneous JADE	2.6 dB	2.2 dB	9.6 dB
DFT length=64	10.6 dB	1.6 dB	8.3 dB
DFT length=128	11.7 dB	-0.4 dB	5.8 dB
DFT length=256	13.1 dB	0.5 dB	6.1 dB
DFT length=512	13.9 dB	-0.2 dB	3.6 dB
DFT length=1024	9.8 dB	0.0 dB	2.6 dB

Real impulse responses

A four-microphone array has been placed in a dummy ear on the right side of a head and torso simulator. In an anechoic room, impulse responses have been estimated from different directions. No sensor noise has been added. Due to the recordings in an anechoic room, the only reflections existing are those from the head and torso simulator. The separation results are shown in Table 5.1. The performance is not as good as in the case of the synthetic impulse responses. In contrast to the synthetic impulse responses, the microphones may have different amplitude and phase responses. This may reduce the performance. The "UK female" seems to be the hardest sound to separate, but from the listening tests, it is easy to determine the separated sound from the two other speech signals.

CONCLUSION AND FUTURE WORK

The performance by the instantaneous gradient flow beamforming is reduced significantly in reverberant mixtures. By expanding the gradient flow principle to convolutive mixtures, it is possible to separate convolutive mixtures in cases where the instantaneous gradient flow beamforming fails. It has been shown that the extension to convolutive mixtures can be achieved by solving a convolutive ICA problem (23) instead of solving an instantaneous ICA problem (15). A frequency domain Jade algorithm has been used to solve the convolutive mixing problem. In order to cope with a more difficult reverberant environment, other convolutive separation algorithms should be investigated. The mixing coefficients (23) are expected to have certain values. E.g. the first row in the mixing matrices is significantly larger than the two other rows. Prior information on the coefficients of the mixing filters could as well be used in order to improve the separation. The knowledge of the delays in the mixing filters may as well be used in order to determine the arrival angles of the mixed sounds.

ACKNOWLEDGEMENT

The authors would like to thank Jan Larsen and Ulrik Kjems for useful comments and valuable discussions. We also acknowledge the financial support by the Oticon Foundation.

REFERENCES

[1] W. Baumann, B.-U. Köhler, D. Kolossa and R. Orglmeister, "Real Time Separation of Convolutive Mixtures," in **ICA2001**, San Diego, California, USA, December 9–12 2001, pp. 65–69.

[2] W. Baumann, D. Kolossa and R. Orglmeister, "Beamforming-Based Convolutive Source Separation," in **ICASSP2003**, Hong Kong, April 2003, vol. V, pp. 357–360.

[3] J.-F. Cardoso, "Blind Signal Separation: Statistical Principles," **Proceedings of the IEEE**, vol. 9, no. 10, pp. 2009–2025, October 1998.

[4] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non Gaussian signals," **IEE Proceedings-F**, vol. 140, no. 6, pp. 362–370, December 1993.

[5] G. Cauwenberghs, M. Stanacevic and G. Zweig, "Blind Broadband Source Localization and Separation in Miniature Sensor Arrays," in **IEEE Int. Symp. Circuits and Systems (ISCAS'2001)**, May 6–9 2001, vol. 3, pp. 193–196.

[6] G. W. Elko and A.-T. N. Pong, "A simple adaptive first-order differential microphone," in **Proceedings of 1995 Workshop on Applications of Single Processing to Audio and Accoustics**, October 15–18 1995, pp. 169–172.

[7] T. J. Ngo and N. Bhadkamkar, "Adaptive blind separation of audio sources by a physically compact device using second order statistics," in **First International Workshop on ICA and BSS**, Aussois, France, January 1999, pp. 257–260.

[8] C. M. Nielsen, **Gradient Flow Beamforming utilizing Independent Component Analysis**, Master's thesis, Aalborg University, Institute of Electronic Systems, January 5 2004.

[9] V. C. Soon, L. Tong, Y. F. Huang and R. Liu, "A wideband blind identification approach to speech acquisition using a microphone array," in **IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-92)**, San Francisco, California USA: IEEE, March 23–26 1992, vol. 1, pp. 293–296.

[10] M. Stanacevic, G. Cauwenberghs and G. Zweig, "Gradient Flow Broadband Beamforming and Source Separation," in **ICA'2001**, December 2001.

[11] M. Stanacevic, G. Cauwenberghs and G. Zweig, "Gradient Flow Adaptive Beamforming and Signal Separation in a Miniature Microphone Array," in **ICASSP2002**, Florida, USA, May 13–17 2002, vol. IV, pp. 4016–4019.