Geometrically Constrained Permutation-free Source Separation in an Undercomplete Speech Unmixing Scenario

Erik Visser

SoftMax
4150 Executive Drive Suite 201
San Diego CA 92121
evisser@softmax.com

Abstract

Frequency domain blind source separation (BSS) problems are typically solved in each frequency bin independently and therefore require additional measures to resolve the resulting permutation problem. In this paper, a frequency domain methodology is presented based on a recently introduced extension of Independent Component Analysis (ICA) to multi-variate components which uses a multi-variate activation function to model dependencies between frequency bins and therefore inherently manages to align most of the permutations. Since the latter approach shows slow convergence behaviour and is prone to converging to local optima, additional geometric constraints are used here to force the BSS algorithm to separate sources with a consistent direction of arrival (DOA) over all frequencies into a minimum number of output channels. DOA information is obtained from a priori knowledge or from subband analysis of partially separated source signals. The methodology is illustrated in an undercomplete acoustic source separation scenario with 3 speakers and 4 microphones.

Index Terms: blind separation, geometric constraints, local optima, convergence

1. Introduction

Blind source separation (BSS) algorithms for convolutive mixtures have experienced many developments in the past and both time and frequency domain methods are available. Time domain algorithms can achieve better separation performance [1] but end up with more computations for the same filter length than equivalent frequency domain nethods. Also, since in time domain algorithms, every filter tap adaptation depends on all other taps, convergence may be slow, prone to local minima and may therefore heavily depend on good initalization [1].

To take advantage of the efficiency of performing long convolutions and rapid convergence of frequency domain based BSS algorithms, several approaches have tried to minimize their shortcomings, most importantly the permutation problem [5, 6]. Permutation can be remediated by time-frequency operations to enforce filter smoothness properties and/or using explict look and null direction steering constraints [6]. Finally direction of arrival (DOA) information as well as neighboring frequency bin correlations and speech harmonicity can be used to align frequency permutations [4, 5].

Recently, a simple solution to the permutation problem has emerged by using multivariate activation functions [2, 3]. While the traditional ICA problem uses a uni-variate activation function, a multi-variate activation function results from the assumption of higher-order dependencies within source vectors as opposed to ICA where mutual independence between source vector elements are assumed. The multi-variate activation function introduces an explicit dependency between individual filter weights during the filter learning process, thereby reducing the degrees of freedom leading to random frequency permutation in conventional frequency domain BSS. While most of the permutations can be realigned in this manner in theory, the procedure introduces some dependency on initial conditions similar to what has been observed in the time domain, resulting in slow convergence prone to local minima. Therefore a regularization technique based on a priori known or iteratively learned geometric information is proposed to overcome convergence to local optima.

2. Independent Vector Analysis (IVA)

In the frequency domain, complex ICA is concerned with finding an unmixing matrix $\mathbf{W}(\omega)$ for each frequency ω such that the demixed outputs $\mathbf{Y}(\omega, l) = \mathbf{W}(\omega)\, \mathbf{X}(\omega, l)$, where $\mathbf{X}(\omega, l) = [X_1(\omega, l), \cdots, X_M(\omega, l)]^T$ (time window l, number of mixtures M), is the DFT of time domain mixtures $\mathbf{x}(t)$, are mutually independent. The update rule for $\mathbf{W}(\omega)$ is given by [2]

$$\Delta \mathbf{W}(\omega) \quad = \quad \mu \left[\mathbf{I} - < \Phi(\mathbf{Y}(\omega, l)\, \mathbf{Y}(\omega, l)^H > \right] \; \mathbf{W}(\omega) \quad (1)$$

where $\mathbf{Y}(\omega, l) = [Y_1(\omega, l), \cdots, Y_M(\omega, l)]^T$, $<>$ denotes the averaging operator in time $l = 1, \cdots, L$ and μ is the learning rate. The traditional Infomax activation function is given by $\Phi(Y_j(\omega, l)) = tanh(|Y_j(\omega, l)|) \frac{Y_j(\omega, l)}{|Y_j(\omega, l)|}$ which along with the update rule (1), implies that the ICA problem is solved for each frequency bin independently, leading to the permutation problem [5, 4]. In [2], it was however shown that by assuming signals of interest have a certain dependency in the frequency domain that can be modeled by a multi-dimensional prior, the original dependent sources can be extracted as a group using such a prior. As a result, a multi-variate activation function [2, 3]

$$\Phi\left(Y_j(\omega, l)\right) \quad = \quad \frac{Y_j(\omega, l)}{\sqrt{\sum_\omega |Y_j(\omega, l)|^2}} \qquad (2)$$

is obtained where the term in the denominator relates to the separated source spectra power over all frequencies. It is noted the multi-variate activation function used here is a special case of a more general learning rule derived from general statistical distributions [2]. Scaling ambiguity of $\mathbf{W}$ is resolved by a scaling matrix designed with the minimum distortion principle [5].

The use of a multi-variate activation function as in eq. (2) avoids the permutation problem in theory by introducing an explicit dependency between individual frequency bin filter weights during the filter learning process. Practically, this simultaneous connected adaptation of filter weights introduces increased convergence dependency on initial filter conditions similar to what has been observed in time domain algorithms [1]. Therefore geometric constraints are used here to overcome these practical limitations.

3. IVA with Linear Geometric Constraints

Geometric constraints can be used to constrain the spatial response of a particular output channel to a particular orientation and placing null beams in others. This a common concept underlying linearly constrained adaptive beamforming, in particular GSC [7]. The idea put forward here is to add a regularization term to the IVA cost function that supports its objective of focusing on a particular source direction by placing spatial nulls in interfering source directions. The following regularization term is proposed

$$J(\omega) \quad = \quad \alpha(\omega)\,||\mathbf{W}(\omega) * \mathbf{D}(\omega,\hat{\theta}) - \mathbf{C}(\omega)||^2 \qquad (3)$$

where $\mathbf{C}(\omega) = \begin{bmatrix} c_1(\omega) & 0 & 0 & \cdots \\ 0 & c_2(\omega) & 0 & \cdots \\ 0 & 0 & \cdots & \end{bmatrix}$. The columns of the directivity matrix $\mathbf{D}(\omega,\hat{\theta})$ are composed of the vectors d_j

$$d_j \quad = \quad exp\left(-i * cos(\hat{\theta}_j) * pos * \omega/c\right) \qquad (4)$$

with $pos = [p_1\ p_2\ \cdots p_M]^T$ being the sensor positions. The $\hat{\theta}_j$s are source direction of arrival (DOA) estimates which are available either from a priori knowledge or need to be determined iteratively in the following manner. It has been shown previously [5] that using the inverse of the unmixing matrix $\mathbf{W}$, the DOA of the separated outputs Y_j can be estimated with

$$\theta_{j,mn}(\omega) \quad = \quad \textbf{arc cos}\,\frac{c\ *\ arg(\frac{[\mathbf{W}^{-1}]_{nj}(\omega)}{[\mathbf{W}^{-1}]_{mj}(\omega)})}{\omega\ *\ ||(p_m - p_n)||} \qquad (5)$$

where $\theta_{j,mn}(\omega)$ is the DOA of source j relative to sensor pair m and n, p_m and p_n the positions of mic m and n respectively and $c = 340\frac{m}{s}$ the sound propagation velocity. When several microphone pairs are used, the DOA $\hat{\theta}_j$ for a specific IVA output Y_j can be computed by plotting a histogram of the $\theta_{j,mn}(\omega)$ from eq. (5) over all microphone pairs and frequencies in selected subbands (see example, Figure 3 and 4). The average $\hat{\theta}_j$ is then the maximum or center of gravity $\sum_{\theta_j=0}^{180} \frac{N(\theta_j)*\theta_j}{\sum_{\theta_j=0}^{180} N(\theta_j)}$ of the resulting histogram (θ_j,$N(\theta_j)$), where $N(\theta_j)$ is the number of DOA estimates at angle θ_j. Reliable DOA estimates from such histograms may only become available in later learning stages when average source directions emerge after a number of iterations. The estimates in eq. 5 are based on a far field model valid for source distances from the microphone array beyond $(2 \sim 4) * D^2/\lambda$, with D the largest array dimension and λ the shortest wavelength considered [7].

Objective (3) can be minimized by using the update rule

$$\Delta\mathbf{W}_{constr}(\omega) \simeq \frac{dJ}{dW(\omega)}$$
$$= \mu\ *\ \alpha(\omega)\ *\ 2\quad *\quad \big(\mathbf{W}(\omega)\,\mathbf{D}(\omega,\theta) - \mathbf{C}(\omega)\big)\mathbf{D}(\omega,\theta)^H \qquad (6)$$

where α is a tuning parameter. When update eq. (6) is added to IVA update eq. (1) to determine the constrained IVA weight update $\Delta\mathbf{W}(\omega)$, tuning α allows to suitably enforce the regularization constraint (3) depending on the spatial separability of the acoustic scenario and other considerations (see example).

If the number of sources R is equal to the number of mixtures M, the choice of the desired beam pattern is set to $\mathbf{C}(\omega) = diag\,(\mathbf{W}(\omega) * \mathbf{D}(\omega,\theta))$, thus nulling out sources from interfering orientations while preserving the beam strength into the desired orientation determined by the constrained IVA algorithm at each iteration. If $R < M$, the kth row of $\mathbf{W}$ for which no DOA has been identified will require a corresponding row of zero entries in $\mathbf{C}(\omega)$, hence all sources are nulled out in this output channel and only background noise remains. Alternatively, if $R < M$, a dimension reduction can be performed first using PCA and then performing IVA on the reduced dimension subspace [5]. The reduced dimension constraint gradient reads

$$\Delta\mathbf{W}_{constr}(\omega) \quad = \quad \mu * \alpha(\omega) * 2 * \big(\mathbf{W}(\omega)\,\mathbf{W_r}(\omega)\,\mathbf{D}(\omega,\theta) - \mathbf{C}(\omega)\big)$$
$$*\big(\mathbf{W_r}(\omega) * \mathbf{D}(\omega,\theta)\big)^H \qquad (7)$$

with $\mathbf{C}(\omega) = diag\big(\mathbf{W}(\omega)\ \mathbf{W_r}(\omega)\ \mathbf{D}(\omega,\theta)\big)$ and where $\mathbf{W_r}$ denotes the $R \times M$ PCA dimension reduction matrix.

Since beamforming techniques are employed and speech is a broadband signal, it must be ensured that good performance is obtained for critical frequency ranges. If the far field model underlying eq. (5) is invalid, near field corrections to the beam pattern need to be made [7]. Also the mic distance must be chosen small enough (less than half the wavelength of the highest frequency) so spatial aliasing is avoided. In this case, it is not possible to enforce sharp beams in the very low frequencies. Figure 1 summarizes the proposed scheme.

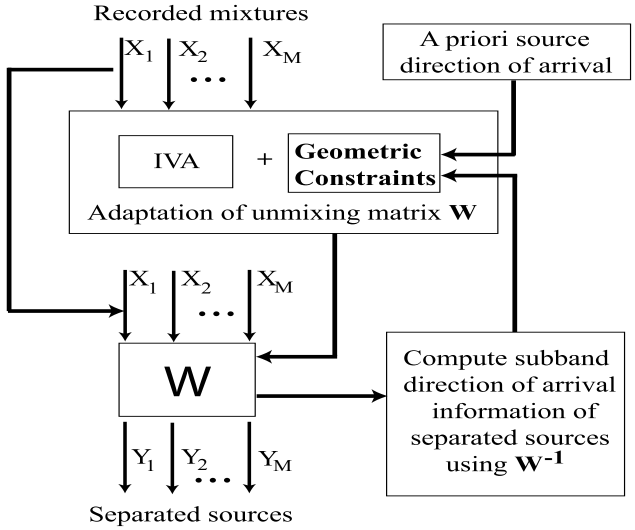
Recorded mixtures
X1 X2 ... XM
A priori source direction of arrival
IVA
+
Geometric Constraints
Adaptation of unmixing matrix W
X1 X2 ... XM
W
Compute subband direction of arrival information of separated sources using W-1
Y1 Y2 ... YM
Separated sources

Figure 1: Overview of proposed scheme based on IVA combined with geometric contraints to avoid local minima and speed up convergence: source direction of arrival information enforced via constraints can either come from a priori knowledge or be computed iteratively on frequency subbands of partially separated sources using the inverse of unmixing matrix $\mathbf{W}(\omega)$.

4. Experiments

An example of an acoustic scenario in a reverberant room (3m × 5m × 3m, T60 = 340 ms) is investigated (see Figure 2) where 4 omnidirectional microphones are used to separate 3 speakers playing back prerecorded speech. A 20 second 4 channel recorded mixture was acquired at 8 kHz sampling rate and processed through 20 sweeps with the IVA algorithm using learning rule (1) and activation function (2) with identity matrices for $\mathbf{W}(\omega)$ as initial conditions (filter length 256). After computing the frequency domain filter taps $W(\omega)$, the equivalent time domain filters filters were reconstructed using the IDFT and the mixture signals filtered to obtain the separated source signals. Quantitative SIR results are shown in Table 1 where SIR is defined as the ratio of the signal power of the target signal to the signal power from the interfering signals.

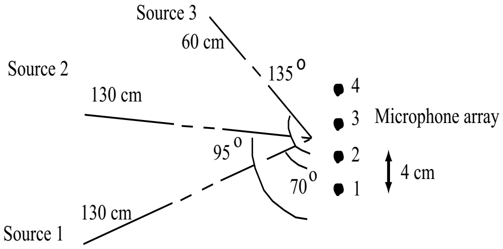

Figure 2: Acoustic scenario with 3 speakers and 4 microphones

To compactly represent the DOAs obtained for each separated IVA output with respect to closely spaced mic pairs (1,2),(2,3) and (3,4), a histogram of DOAs computed for each separated IVA output channel with formula (5) is plotted over all mic pairs and frequencies in the [0 - 4 kHz] band as shown in Figure 3. It can be seen that a defined DOA has become apparent for output channels 1 and 2 which correspond to sources 1 and 2 with respective DOA of 70 and 95 degrees. However IVA outputs 3 and 4 do not exhibit a clearly defined maximum when the whole available frequency range of [0-4 kHz] is considered.

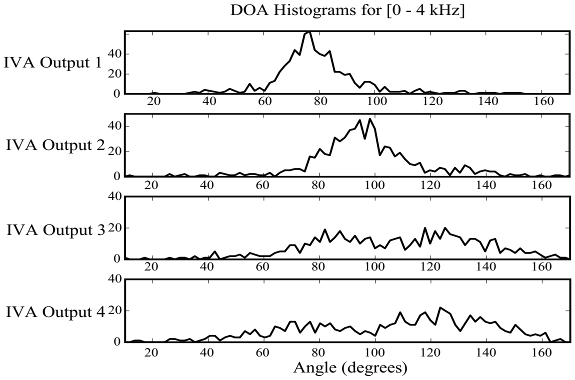

Figure 3: Histogram of estimated DOAs with eq. (5) for mic pairs (1,2), (2,3) and (3,4) over all frequencies in [0-4kHz] band for each IVA separated output using learning rule (1) and activation function (2): IVA outputs 1 and 2 correspond to sources 1 and 2 respectively while no defined DOA is perceived in outputs 3 and 4

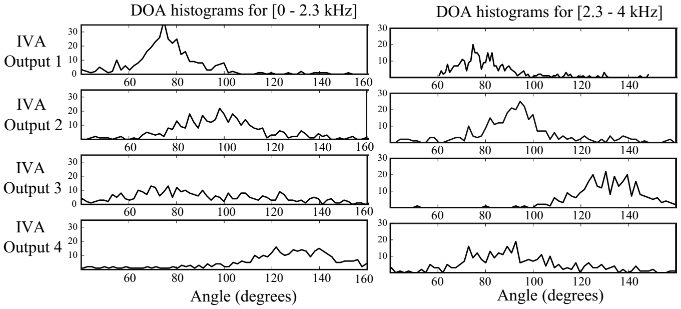

Figure 4: Decomposition of DOA histogram in Figure 3 into low frequency ([0-2.3kHz], left plot) and high frequency ([2.3-4kHz], right plot) bands: output channels 1 and 2 show consistent DOAs around 75 and 95 degrees in both bands while a DOA in the range of [120-140] degrees is apparent in the low frequency band of IVA output channel 4 (left plot) and in the high frequency band of output channel 3 (right plot). The unconstrained IVA has thus separated a continuous high frequency band of source 3 (135 degrees) into output channel 3 while its low frequency band is separated into output channel 4, therefore converging to a local minimum

After splitting the histogram into low ([0-2.3kHz], left plot in Figure 4) and high frequency ([2.3-4 kHz], right plot in Figure 4) components, it can be seen that there exists a third source with a DOA around 135 degrees but that the respective high and low frequencies have been aligned in non permuted bands in different output channels (Figure 4). The unconstrained IVA algorithm has thus converged to a local optimum.

From the histograms in Figure 4, three distinct DOAs of 76, 96 and 134 degrees were determined by combining the average values of the histograms' center of gravity (see section 3) in each IVA output in the low and high frequencies bands. These values were used as DOA setpoints in update eq. (6) and the constrained algorithm (update eq. (6) added to (1) with activation function (2)) restarted in the local optimum. Figure 6 and Table 1 illustrate the different effects of choosing parameter α which allows to trade off the regularization constraint against the IVA objective. One way of interpreting the different schemes is to look at the spatial beam patterns of the separated outputs Y_j by plotting the quantity $|W(\omega) * d(\theta, \omega)|$ (d defined as in eq. 4) over all angles θ for each row of the unmixing matrix $\mathbf{W}$. As illustrated for all 3 output beam patterns at frequency 2 kHz in Figure 6, aggressive enforcement of constraints (3) yields unnecessary deep nulls at specified interfering DOAs leading to significantly worse performance (Table 1, conIVA(α=2)) than the one observed in the IVA local minimum (Table 1, IVA). Using a smaller α allows to improve the SIR for separated source 3 substantially (conIVA, α=0.1). The best overall performance is however obtained when the unconstrained IVA algorithm using eq. (1) and (2) is restarted (Table 1, IVAopt) with initial filter values given by the final solution obtained with the constrained IVA using $\alpha = 0.1$. Figure 5 shows the final histograms for the optimal case (IVAopt) with the DOA of source 3 clearly visible in IVA output channel 3 and no directional information apparent in IVA output channel 4. Also, as can be seen in Figure 6, the optimal solution (IVAopt) trades off the depth and exact positioning of the null beams against each other and therefore differs in fine tuning from the constrained IVA solution obtained with α=0.1. This suggests it is advisable to use the geometric constraints to initialize the IVA algorithm in a region close to the optimum using a priori knowledge or to guide it into such a region with small α using iteratively estimated DOAs. Since the latter DOA can never be determined accurately, especially in reverberant environments, small αs are advised throughout the learning process when no a priori spatial information is available.

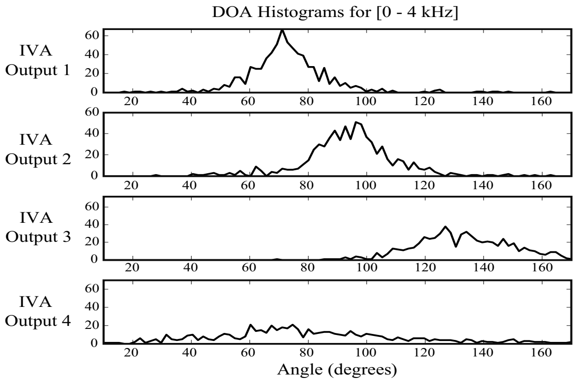

Figure 5: Histogram of DOA estimates from eq. (5) over all mic pairs and frequencies corresponding to IVA umixing solution (update rule (1), activation function(2)) initialized at converged solution of constrained IVA (update eqs. (1) and (6), α=0.1)(see text): as opposed to Figure 3, IVA output 3 clearly emerges as source 3

Since we are dealing with an undercomplete scenario, an alternative route is to apply IVA where a PCA dimension reduction from a 4 dimensional to a 3 dimensional space is applied first followed by IVA using update eqs. (1), (7) and function (2) on the reduced 3 dimensional mixture. Although the constraints were not applied towards the end of the convergence in this case, the overall performance using PCA first (Table 1, PCA-IVA) is inferior to the one observed in the full dimensional case (IVAopt). Thus, by using the regularization term (3), one can force the IVA algorithm to separate all point sources into a minimum of output channels and does not require a dimension reduction first. More degrees of freedom are preserved for the separation in this way through direct use of 4 mic measurements instead of a PCA selected reduced measurement space not necessarily optimal for overall signal separation. By constraining the separated solutions into designated output channels, some channel selection capability is also provided.

SIR (dB)	Source 1	Source 2	Source 3
Recording	-4.72	-9.26	-7.02
IVA	18.98	10.10	5.35
conIVA (α=2)	2.13	-3.78	2.63
conIVA (α=0.1)	16.39	10.04	12.76
IVAopt	19.85	10.73	12.97
PCA-IVA	15.29	9.45	13.15

Table 1: SIR results for different separation algorithms: IVA = IVA with update rule (1) using function (2); conIVA=constrained IVA using update rule composed of (1) added to (6) for different settings of α; IVAopt= IVA initialized at final solution obtained with conIVA(α=0.1); PCA-IVA=constrained IVA using update rule (1) added to (7) on reduced 3 dimensional PCA subspace

5. Conclusions

A frequency domain, permutation-free source separation scheme was proposed based on a combination of an independent component vector analysis algorithm and geometric constraints. Using a priori or iteratively computed DOA information, convergence speed of the BSS algorithm can be enhanced and local optima avoided. Moreover no PCA type dimension reduction is required in undercomplete mixing scenarios as geometric constraints can force source signals into a minimum number of output channels. The limitation of this approach is that a large number of mics may be required to allow sufficient spatial resolution of geometric constraints when sources are located close to each other.

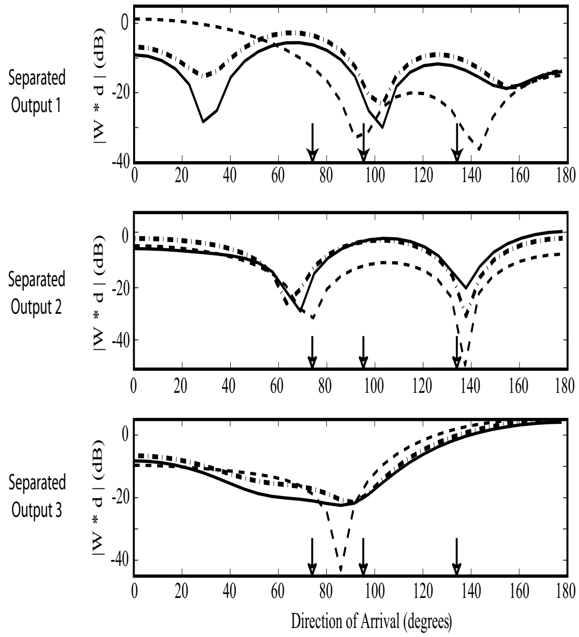

Figure 6: Illustration of separated output beam patterns $|\mathbf{W} * d(\theta)|$ obtained using constrained IVA (update rule (1) added to eq. (6) with activation function (2)) at ω=2 kHz: Null beams are placed at DOA 76, 96, 134 degrees and different aggressiveness of constraint enforcement using α (eq. 6) allows to adjust the depth of resulting null beams (dashed line for α = 2; dashed-dotted line for α=0.1). The solid line corresponds to the optimum solution of unconstrained IVA (update rule (1) with function (2)) initialized with converged filters obtained from constrained IVA using $\alpha = 0.1$

6. References

[1] Ukai, S., Takatani, T., Saruwatari, H., Shikano, K., Mukai, R., Sawada, H., Multistage SIMO-Model-Based Blind Source separation combining frequency domain ica and time domain ica, IECE Trans Fundamentals, E88-A(3), pp.642-650, 2005

[2] Kim, T., Eltoft, T., Lee, T.-W., Independent Vector Analysis: An Extension of ICA to Multivariate Components, Proc. of 6th Conf. on ICA and BSS, pp. 165-172, March 2006

[3] Hiroe, A., Solution of permutation problem in frequency domain ICA using multivariate probability density functions, Proc. of 6th Conf. on ICA and BSS, pp. 601-608, March 2006

[4] Kurita, S., Saruwatari, H., Kajita, S., Takeda, K., Itakura, F., Evaluation of blind signal separation using directivity pattern under reverberant conditions, ICASSP, pp. 3140-3143, 2000

[5] Mukai, R., Sawada, H., Araki, S., Makino, S., Frequency domain blind source separation for many speech signals, Proc. ICA 2004, pp. 461-469, 2004

[6] Parra, L.C., Alvino, C.V., Geometric source separation: Merging convolutive source separation with geometric beamforming, *IEEE Trans. Speech. Audio Proc.*, **10**, pp. 352-362, 2002

[7] Kennedy, R.A., Abhayapala, T.D., Ward, D.B., Broadband nearfield beamforming using a radial transformation, IEEE Transactions on Signal Processing, vol 46, no 8, 1998