FREQUENCY DOMAIN PASSIVE BROADBAND SPEAKER LOCALIZATION USING A PERMUTATION-FREE BLIND SOURCE SEPARATION ALGORITHM

Erik Visser

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

ABSTRACT

Traditional passive broadband source localization techniques like maximum likelihood estimation and MUSIC have shown difficulties in situations where multiple correlating source signals are interfering with each other. Blind Source Separation (BSS) algorithms on the other hand have demonstrated good performance in separating correlated mixture signals into independent sources. In this paper it will be shown that the performance of traditional source localization algorithms can be improved by using a permutation-free frequency domain BSS algorithm as a front end. In addition a source localization method based solely on information gained from the separated BSS solution and sensor array architecture is presented. The methodologies are illustrated in an undercomplete acoustic scenario involving 3 speech sources and a 6 element microphone array.

Index Terms— Passive source localization, source separation

1. INTRODUCTION

Blind Source Separation (BSS) or Independent Component Analysis (ICA) algorithms for convolutive mixtures have experienced many developments in the past and recently the performance of frequency domain based methods has become more robust through the solution of the frequency permutation problem [1]. In many studies the BSS solution unmixing multiple source scenario into independent sources has been used to retrieve direction of arrival (DOA) information for each separated source [6], but little research has been done to use BSS for ranging purposes to allow complete source localization. A traditional technique in source localization includes the maximum likelihood (ML) solution [2], but it did not become popular due to its high computational cost. Concurrently, a variety of suboptimal techniques with reduced computations have dominated the field. Standard techniques include the minimum variance method of Capon [2] and the multiple signal classification (MUSIC) method of Schmidt [3]. However a well known problem with these techniques occurs when two or more sources are highly correlated in time or space. The performance of these traditional techniques can be robustified by using BSS as a front end to decorrelate individual interfering signals before source localization as will be shown in the following.

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 $\mathbf{W}(\omega)$ [1] is given by

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

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 [6]. In [1], it was however shown that by assuming the signal of interest have a certain dependency in the frequency domain that can be modeled by a multi-dimensional prior, the dependent sources can be extracted as a group using such a prior. Such an assumption leads for example to the IVA multi-variate activation function [1]

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

with the L_2 norm of $Y_j(\omega, l)$ over all ω at the denominator. The multi-variate activation function used here is a special case of a more general learning rule derived from general statistical distributions [1]. Scaling ambiguity of $\mathbf{W}$ is resolved by a scaling matrix designed with the minimum distortion principle [6].

3. TRANSFORMATION FROM MULTIPLE SOURCE TO SINGLE SOURCE SCENARIO

The separation with IVA algorithm (2) yields the demixed signals $Y_j(\omega, l)$. To transform the original multiple source data $\mathbf{X}$ into single source recorded data, a corresponding data matrix for each source component Y_j is reconstructed by computing the inverse (or pseudo-inverse) of $\mathbf{W}$ and selectively backprojecting an output of interest Y_j (Figure 1) i.e.

$$\hat{\mathbf{X}}(Y_j)(\omega, l) \quad = \quad \mathbf{W}^{-1}(\omega) \left[\hat{S}_1(\omega, l), \cdots, \hat{S}_j(\omega, l), \cdots, \hat{S}_M(\omega, l) \right]^T,$$

where $\hat{S}_j(\omega, l) = Y_j(\omega, l)$ and $\hat{S}_k(\omega, l) = 0$ for $k \neq j$, k=1,$\cdots M$. If the IVA separation is complete, this backprojecting process becomes equivalent to recording a *single source mixture*, with the reconstructed mixture data $\hat{\mathbf{X}}(Y_j)$ now containing the time delay of arrival (TDOA) or range difference (RD) information for a single source in a certain location with respect to the microphone array. Since the reconstructed recording estimate $\hat{\mathbf{X}}(Y_j)$ only contains one source, single source TDOA estimation techniques and corresponding ranging methods can be applied. In addition, source localization

can be performed by using information from the inverse of $\mathbf{W}$ alone. 4 different source localization methods are presented as illustrated by Figure 1.

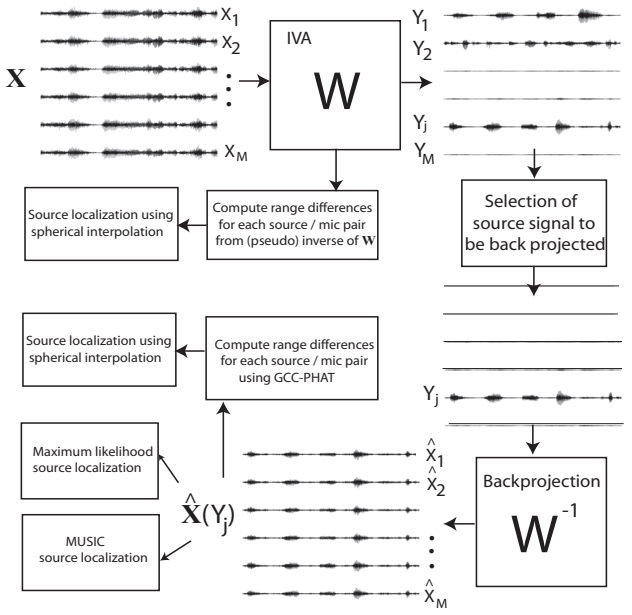

Fig. 1. Overview of blind source separation based source localization strategies

3.1. Source localization using range differences from $\mathbf{W}^{-1}$

Since the recorded mixture signals are separated into estimated source signals $\mathbf{Y}(\omega,l) = \mathbf{W}(\omega)\,\mathbf{X}(\omega,l)$, the inverse of W, W^{-1}, contains information about the mixing scenario $\mathbf{X}(\omega,l) = \mathbf{H}(\omega) * \mathbf{S}(\omega,l)$ ($\mathbf{H}$ denotes the mixing transfer function matrix) as $\mathbf{X}(\omega,l) = \mathbf{W^{-1}}(\omega)\,\mathbf{Y}(\omega,l)$. If the IVA separation is complete and selected outputs in $\mathbf{Y}$ are good approximations of the true sources $\mathbf{S}$, the phase differences between elements $\left[\mathbf{W}^{-1}\right]_{mp}(\omega)$ and $\left[\mathbf{W}^{-1}\right]_{np}(\omega)$ of selected columns p of $\mathbf{W}^{-1}$ are related to the phase differences between the transfer functions $\mathbf{H}_{mp}$ from a source p to mic m and $\mathbf{H}_{np}$ from a source p to mic n [6]. From these, the range difference d^p_{mn} between a source p to mic m distance and a source p to mic n distance defined as
$d^p_{mn} = \sqrt{(x_p-x_m)^2+(y_p-y_m)^2}-\sqrt{(x_p-x_n)^2+(y_p-y_n)^2}$, with (x_p,y_p), (x_m,y_m) and (x_n,y_n) being the coordinates of source p, mic m and n respectively, can be estimated as

$$d^p_{mn}(\omega) \;=\; -\frac{c}{\omega}\,arg\Big(\frac{\left[W^{-1}\right]_{mp}(\omega)}{\left[W^{-1}\right]_{np}(\omega)}\Big) \qquad (3)$$

$c = 340\frac{m}{s}$ being the sound propagation velocity. Therefore the separation process yields a set of range differences (RD) for each source p for each microphone pair over all frequencies. For a given source and microphone pair, the estimated range difference should be consistent over all frequencies. There is however only a limited frequency range $[\omega_{lb},\omega_{ub}]$ over which this quantity can be estimated. The upper bound is given by spatial aliasing that may occur starting around $\omega_{ub} = 0.5\,\frac{c}{\sqrt{(x_m-x_n)^2+(y_m-y_n)^2}}$[7, 2]. A lower bound for this frequency range is motivated by incomplete separation of mixture signals in the very low frequencies due to small microphone spacing and the fact that for a particular band λ to be useful in near field source localization, the source should not exceed a certain distance from the array approximately given by $\frac{2*D^2}{\lambda}$, D being the largest array dimension [7]. Beyond that distance, the source is in the far field of the array for this band and information gained through the use of this band will be unreliable for ranging purposes. Finally, RD estimates in $[\omega_{lb},\omega_{ub}]$ show small random fluctuations (see Figure 3). By modelling these changes by Gaussian noise and averaging the sequence of frequency dependent range RD values for example with a Kalman filter, an average RD $\bar{d}^p_{m,n}$ is obtained for each source p and mic pair (m,n).

The estimated RDs have to be combined to yield a consistent source localization result (x_p,y_p) for each source p over all mic pairs. Here, a least squares spherical interpolator is used by [2]

$$\min_{x_p,y_p} \qquad \sum_{m,n}\left(\bar{d}^p_{m,n}-d(x_p,y_p)_{m,n}\right)^2 \qquad (4)$$

where $d(x_p,y_p)_{m,n} = \sqrt{(x_p-x_m)^2+(y_p-y_m)^2}-\sqrt{(x_p-x_n)^2+(y_p-y_n)^2}$. The unconstrained cost function must be minimized using nonlinear optimization techniques [2].

3.2. Source localization using range difference information gained through weighted cross correlation

Range difference estimates can also be obtained by determining the TDOAs from the cross correlations between reconstructed single source mic recordings $\hat{X}(Y_j)$. For a given mic pair (m,n) and back-projected output Y_j, the cross correlations
$R_{mn}(l,\tau) = \sum_{\omega_{lb}}^{\omega_{ub}}\Psi(\omega)\hat{X}_m(Y_j)(\omega,l)\big(\hat{X}_n(Y_j)\big)^*(\omega,l)\,e^{i\,\omega\,\tau}$

are determined where a GCC-PHAT weighting was chosen [4]:
$\Psi(\omega) = \|\hat{X}_m(Y_j)(\omega)\,\big(\hat{X}_n(Y_j)\big)^*(\omega)\|^{-1}$. The best TDOAs τ are used to compute the average RD
$\bar{d}^p_{m,n} = \sum_{l=1}^{L}\frac{c}{L}*\mathbf{arg\,max}_\tau\, R_{mn}(l,\tau)$ and problem (4) is solved for a source p given the $\bar{d}^p_{m,n}$ over all mic pairs (m,n). A limited range $[\omega_{lb},\omega_{ub}]$ is required as discussed in section 3.1 and because little speech power is contained in the higher frequencies.

3.3. Wideband MUSIC source localization

The basic idea in this approach is to find the steering vector orthogonal to the null space of the array recorded data spatial spectral density matrix [3] $P_{xx}(\omega) = \sum_{l=1}^{L}\mathbf{X}(\omega,l)\mathbf{X}(\omega,l)^H/L$. Here, the reconstructed single source matrix $\hat{\mathbf{X}}(Y_j)$ is used instead of the original recorded data $\mathbf{X}$. A singular value decomposition of $P_{xx}(\omega) = [U_s(\omega)U_n(\omega)]\begin{bmatrix}\sum_s(\omega) & 0\\ 0 & \sum_n(\omega)\end{bmatrix}\begin{bmatrix}U_s^H(\omega)\\ U_n^H(\omega)\end{bmatrix}$ is computed to yield the signal subspace $U_s(\omega)$ and noise subspace $U_n(\omega)$ with respective singular values $\sum_s(\omega)$ and $\sum_n(\omega)$ at ω. To find the correct steering vectors
$a(\omega,x,y) = [e^{-i*2*\pi*\omega*t_1}\,e^{-i*2*\pi*\omega*t_2}\cdots e^{-i*2*\pi*\omega*t_M}]^T$
with the time delays t_m, $m=1,\cdots,M$ defined by
$t_m = \sqrt{(x-x_m)^2+(y-y_m)^2}/c$, the indicator

$$I_{MUSIC}(x,y) \;=\; \frac{1}{\sum_{\omega_{lb}}^{\omega_{ub}}\|a(\omega,x,y)^H\,U_n(\omega)\|} \qquad (5)$$

is computed over a frequency range $[\omega_{lb},\omega_{ub}]$ as discussed in sections 3.1 and 3.2. (x_m,y_m) are the location of microphones m, m= $1,\cdots,M$ and (x,y) is the array steering location. The indicator will yield a maximum as the steering vector steers to the correct location.

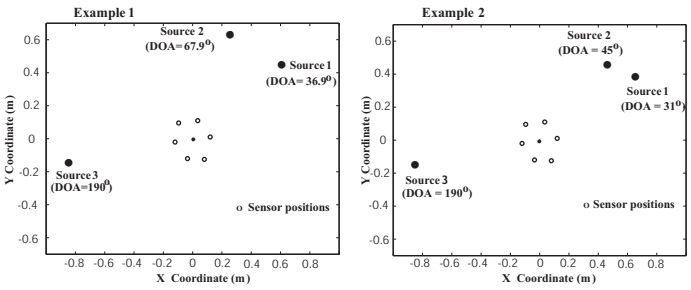

Fig. 2. Acoustic scenario in Example 1 (**left plot**): Male speaker source 1 (0.6,0.45); Male speaker source 2 (0.26,0.64); Female speaker source 3 (-0.85,-0.15). Example 2 (**right plot**): Male speaker source 1 (0.65,0.39); Male speaker source 2 (0.46,0.46); Female speaker source 3 (-0.85,-0.15); DOA = atan($\frac{y}{x}$)

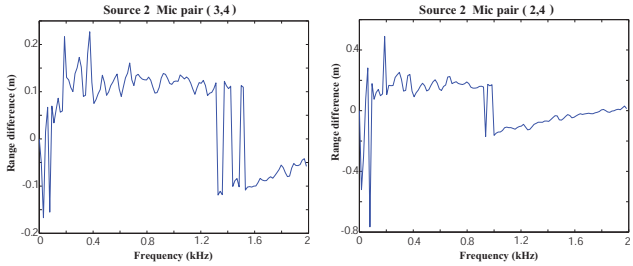

Fig. 3. Example 1: Range difference estimates per frequency bin (eq. 3) for source 2/mic pair (3,4) (**left plot**) and source 2/mic pair (2,4) (**right plot**): spatial aliasing occurs for mic pair (3,4) around 1.4 kHz compared to 1 kHz for wider spaced mic pair (2,4).

3.4. Maximum likelihood source localization

In this approach [2], the array recorded mixture signal is modeled by $\mathbf{X}(\omega) = \mathbf{A}(\omega)\,\mathbf{S}(\omega) + \eta(\omega)$ where A is the steering matrix given by $A(\omega) = \left[a^{(1)}(\omega), \cdots, a^{(P)}(\omega)\right]$ with the steering vectors $a^p(\omega, x_p, y_p) = [g_1^p e^{-i*2*\pi*\omega*t_1^p} \cdots g_M^p e^{-i*2*\pi*\omega*t_M^p}]^T$, $t_m^p = \sqrt{(x_p - x_m)^2 + (y_p - y_m)^2}/c$, (x_p, y_p) and (x_m, y_m) the coordinates of source p and mic m. $\mathbf{S}(\omega) = \left[S^1(\omega) \cdots S^P(\omega)\right]^T$ is the source spectrum matrix and the noise spectrum $\eta(\omega)$ is assumed zero mean complex white Gaussian. The maximum likelihood estimation of the source locations and source signals results from the solution of $\min_{x_p, y_p,\ p\,=\,1,\cdots,P} = \|\mathbf{X}(\omega) - \mathbf{A}(\omega)\,\mathbf{S}(\omega)\|$ After solving for $\mathbf{S}(\omega)$ and substituting, the optimization criterion for a single source p becomes [2]

$$\max_{x_p, y_p} I_{ML} = \sum_{\omega_{lb}}^{\omega_{ub}} \|\bar{\mathbf{a}}^{\mathbf{p}}(\omega, x_p, y_p)^H\, \mathbf{X}(\omega)\| \qquad (6)$$

with $\bar{a}^p = \frac{a^p}{\sqrt{\sum_{m=1}^{M} g_m^p}}$. Since we consider near-field sources, the signal strength at each sensor can be different due to nonuniform spatial loss in the near-field geometry. It is assumed here that $g_m^p = 1$ i.e sensor gains are uniform and spatial loss negligible.

By selectively backprojecting each single separated IVA output Y_j and substituting $\mathbf{X}$ with $\hat{\mathbf{X}}(Y_j)$, the corresponding source p source can be localized using this single source case optimization objective. It therefore avoids having to solve a multiple source ML objective with the recorded $\mathbf{X}$ involving intense computations and local minima in the case of highly correlated sources [2]. As discussed in sections 3.1 and 3.2, lower and upper bounds for ω apply.

4. EXPERIMENTS

Experiments were carried out in an office environment (3m × 5m × 3m, T60 = 340 ms) with 6 omni directional microphones to separate 3 speaker signals each playing back continuous prerecorded sentences(Figure 2). The mic locations were [-0.095,0.095] (#1), [-0.12 -0.02] (#2), [-0.035 -0.12] (#3), [0.08 -0.125] (#4), [0.12 0.01] (#5) and [0.035 0.11] (#6). Source separation using the Infomax algorithm with activation function (2) was obtained using 20 sweeps on 15 second 8 kHz mixture recordings using different filter initial conditions to assure complete convergence. The filter length was 256, the FFT length 512 with an overlap of 350 samples. After computing the frequency domain filter taps in $\mathbf{W}(\omega)$, equivalent time domain filters were computed and the time domain mixture signals filtered to obtain the separated source signals. The separation is quantified by the SIR values (defined as as the ratio of the signal power of the target signal to the signal power from the interfering signals) given in Table 1. Since the mixing scenario is undercomplete (6 mics, 3 speakers), the outputs that contained the separated sources were determined in a supervised manner. Using the pseudo-inverse of the unmixing matrix $\mathbf{W}(\omega)$, the individual source signals were then selectively backprojected.

SIR (dB)	Source 1	Source 2	Source 3
	Ex 1 / Ex 2	Ex 1 / Ex 2	Ex 1 / Ex 2
Recording	-5.03 / -4.12	-3.51 / -9.64	-7.34 / -3.44
Separation	20.02 / 19.22	18.55 / 11.37	17.25 / 12.96

Table 1. Source SIR in recorded and separated data (Example 1/2)

As shown in Figure 3, RD estimates obtained from eq. 3 to infer source localization from the pseudo inverse $\mathbf{W}^{-1}$ are only reliable in a limited frequency range $[\omega_{lb}, \omega_{ub}]$. In addition to spatial aliasing shown in Figure 3 defining the upper bound, the RD values below 0.5 kHz show large variations in both examples, indicating a lower bound as discussed in section 3.1. A final average RD in band $[0.5 - 0.9]$ kHz for mic pair (2,4) and band $[0.5 - 1.3]$ kHz for mic pair (3,4) is estimated using a Kalman filter with the estimates from eq. 3 as the input sequence to be smoothed.

Table 2 summarizes the results obtained with the different source localizations techniques discussed. The optima for objectives 4, 5 and 6 were found using an implementation of the simplex function optimization algorithm [2]. Maximum likelihood estimation on the original multi-source recorded data using single source search grids [2] failed to resolve speakers 1 and 2 and is therefore inappropriate for addressing closely spaced sources (see Figure 4). MUSIC applied to the recorded data of example 1 yielded a signal subspace of 3 principal components which were used to determine the spatial location of all sources with good accuracy. However, in example 2, only 2 signal subspace components were obtained resulting in the detection of only one source located in between source 1 and source 2 (see Figure 6). On the other hand, maximum likelihood and MUSIC applied to the individually backprojected IVA separated outputs yielded good source localization accuracy. As shown in Figure 5 & 7, the IVA-maximum likelihood method correctly resolved sources 1 and 2 although located spatially close together. Morover it can be seen that the least squares methods based on RD estimates tended to over estimate the range of the separated sources due to the flat optima (see Figure 4-7). Also, the least squares solutions obtained through the use of $\mathbf{W}^{-1}$ are more accurate than using RD estimates obtained with GCC-PHAT on the backprojected data. The corresponding DOA errors (Δ DOA = $\|$ atan($\frac{y}{x}$)-atan($\frac{y_{true}}{x_{true}}$)$\|$) are however quite small (less than 6 degrees on average) so the RD estimated through

Example 1	ML $\mathbf{X}$	MC $\mathbf{X}$	ML IVA	MC IVA	LS IVA	GCC IVA
Source1 rms(m)	N/A	0.11	0.03	0.03	0.17	0.11
Source2 rms(m)	N/A	0.04	0.04	0.11	0.05	0.14
Source3 rms(m)	0.18	0.04	0.04	0.11	0.13	0.18
Mean rms(m)	N/A	0.06	0.04	0.08	0.12	0.14
Mean DOA error	N/A	4.4^o	2.3^o	2.7^o	4.4^o	4.1^o
Example 2						
Source1 rms(m)	N/A	N/A	0.03	0.11	0.03	0.05
Source2 rms(m)	N/A	N/A	0.03	0.08	0.29	0.25
Source3 rms(m)	0.12	0.04	0.10	0.07	0.12	0.22
Mean rms (m)	N/A	N/A	0.05	0.09	0.14	0.17
Mean DOA error	N/A	N/A	0.7^o	0.9^o	2.0^o	6.0^o

Table 2. Range(rms)/DOA estimation errors for Example 1 & 2: ML-$\mathbf{X}$ = Maximum likelihood estimation applied to original recorded data $\mathbf{X}$; MC-$\mathbf{X}$ = MUSIC applied to $\mathbf{X}$; ML-IVA = ML applied to backprojected data $\hat{\mathbf{X}}(Y_j)$; MC-IVA = MUSIC applied to backprojected data $\hat{\mathbf{X}}(Y_j)$; LS-IVA = Least squares interpolation (eq.4) using RDs from $\mathbf{W}^{-1}$; GCC-IVA = Least squares (eq.4) using RDs from GCC-PHAT on backprojected data $\hat{\mathbf{X}}(Y_j)$;

$\mathbf{W}^{-1}$ and GCC-IVA are useful. Hence the limited range precision of the least squares techniques with this particular microphone array can be improved in better conditioned source-microphone arrangements.

5. CONCLUSIONS

It was shown that using a blind source separation algorithm as a front end to traditional multi-source localization techniques can significantly enhance spatial resolution of such techniques by transforming a multi-source localization problem into single source scenarios. In particular a source localization technique was presented where range difference information is retrieved directly from the blind source separation unmixing matrix for subsequent use in least squares spherical interpolation.

6. REFERENCES

[1] Kim, T., Eltoft, T., Lee, T.-W., Independent vector analysis (IVA): an extension of ICA to multivariate components, Proc. ICA and BSS 2006, pp. 165-172, 2006

[2] Chen, J.C., Hudson, R.E., Yao, K., Maximum likelihood source localization and unknown sensor location estimation for wideband signals in the near-field, IEEE Trans. Sig. Proc., Vol 50, No 8, 2002

[3] Schmidt, R.O., Multiple emitter location and signal parameter estimation, Trans. Ant. Prop, vol AP34, pp.276-280, 1986

[4] Knapp, C.H., Carter, G.C., The generalized correlation method for estimation of time delay, IEEE Trans. ASSP, vol ASSP-24, pp. 320-327, 1976

[5] Tung, T.L., Yao, K., Chen, D., Hudson, R.E., Reed, C.W., Source localization and spatial filtering using wideband music and maximum power beamforming for multimedia applications, in Proc IEEE SiPS, Oct. 1999, pp. 625-634

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

[7] Kennedy, R.A., Abhayapala, T.D., Ward, D.B., Broadband nearfield beamforming using a radial transformation, IEEE Trans. Sig. Proc., vol 46, no 8, 1998

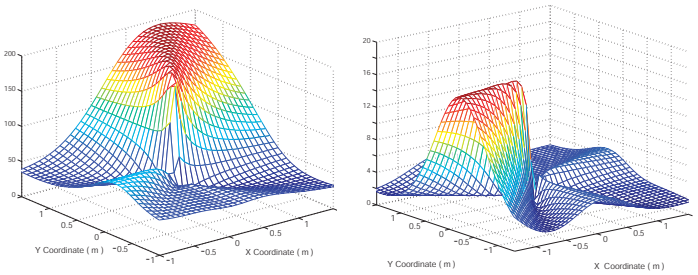

Fig. 4. Example 1, **left plot**: Maximum likelihood (ML) indicator (eq. 6) on recorded data $\mathbf{X}$: source 3 (-0.85,-0.15) is barely resolved, source 1 (0.60, 0.45) and 2 (0.26, 0.64) DOA is not clear. **right plot**: ML indicator computed on backprojected data $\hat{\mathbf{X}}(Y_j)$ corresponding to source 3 (-0.85 -0.15)

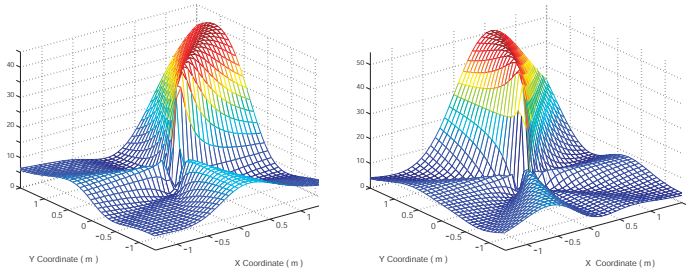

Fig. 5. Example 1: ML indicator (eq. 6) applied to backprojected data $\hat{\mathbf{X}}(Y_j)$ corresponding to source 1 (**left plot**, (0.6,0.45)) and source 2 (**right plot**, (0.26,0.64)): both sources' DOA are clearly extracted (compare to Figure 4)

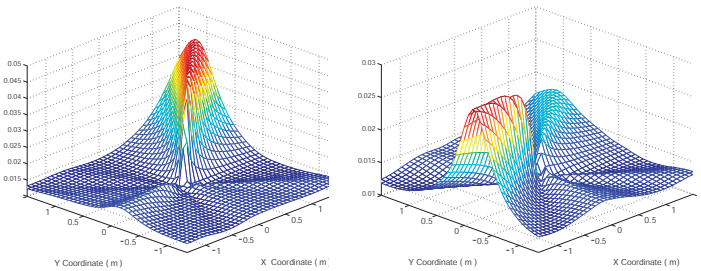

Fig. 6. Example 2: MUSIC indicator (eq. 5) computed on recorded data $\mathbf{X}$ signal subspace consisting of only 2 components: 1st component (**left plot**) yields location (0.6,0.45) between source 1 (0.65,0.39) and 2 (0.46,0.46) and does not resolve these closely spaced sources. 2nd component shows source 3 (**right plot**)

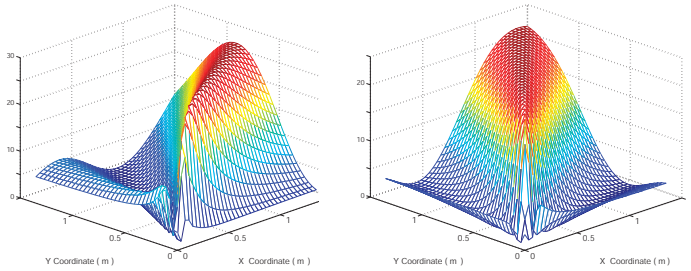

Fig. 7. Example 2: ML indicator (eq. 6) computed on IVA backprojected data $\hat{\mathbf{X}}(Y_j)$ corresponding to source 1 (**left plot**, (0.65,0.39)) and source 2 (**right plot**, (0.46,0.46)): although spatially close, a clear difference is observed in their respective DOAs.