Acoustic Source Separation with Microphone Arrays

Lucas C. Parra
Biomedical Engineering Department
City College of New York

CCNY

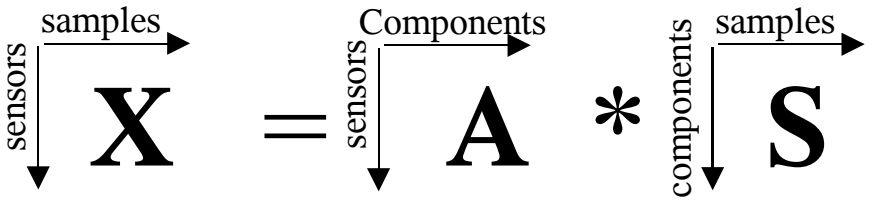
Craig Fancourt
Clay Spence
Chris Alvino

SARNOFF Corporation

Montreal Workshop, Nov 6, 2004

Blind Source Separation and ICA

The goal of Independent Component Analysis (ICA) is to factorize an observation matrix **X** into a basis vectors **S** and coefficients **A** such that the columns of **S** are statistically independent:



The notion is that the observations $\mathbf{x}(t)$ have been generated by linear mixing **A** of independent sources $\mathbf{s}(t)$, and the recovered component are the originating sources.

Separation Based on Independence

Statistical independence implies for all $i \neq j,t,l,n,m$:

$$\mathrm{E}[s_i^n(t)\, s_j^m(t+l)] = \mathrm{E}[s_i^n(t)]\mathrm{E}[s_j^m(t+l)]$$

For M sources and N sensors each tuple $\{t,l,n,m\}$ this equation gives $M(M\text{-}1)/2$ conditions for the NM unknowns in $\boldsymbol{A}$.

Sufficient conditions if we use multiple:

use	sources assumed*	condition	statistic	algorithm
t	non-stationary	$\mathbf{R_x}(t) = \mathbf{A}\,\mathbf{R_s}(t)\,\mathbf{A}^{\mathrm{T}}$	covariance	decorrelation
l	non-white	$\mathbf{R_x}(l) = \mathbf{A}\,\mathbf{R_s}(l)\,\mathbf{A}^{\mathrm{T}}$	cross-correlation	SOBI
n, m	non-Gaussian	$\mathbf{C_x}(i,j) = \mathbf{A}\mathbf{C_s}(i,j)\mathbf{A}^{\mathrm{T}}$	4th cumulants	JADE (ICA)

* zero mean

'Quickie BSS' – GEV of two Cross-Statistics

The independence assumption establishes that there are some cross-statistics of the observations say $\mathbf{R}_1$, $\mathbf{Q}_2$ that are diagonalized by $\mathbf{W}=\mathbf{A}^{-1}$:

$$\mathbf{R_S} = \mathbf{W}\,\mathbf{R_X}\,\mathbf{W}^T = diag$$
$$\mathbf{C_S} = \mathbf{W}\,\mathbf{C_X}\,\mathbf{W}^T = diag$$

This can be combined to a standard generalized Eigenvalue equation

$$\mathbf{R_X}^{-1}\mathbf{C_X}\,\mathbf{W} = \mathbf{W}\,\mathbf{D}$$

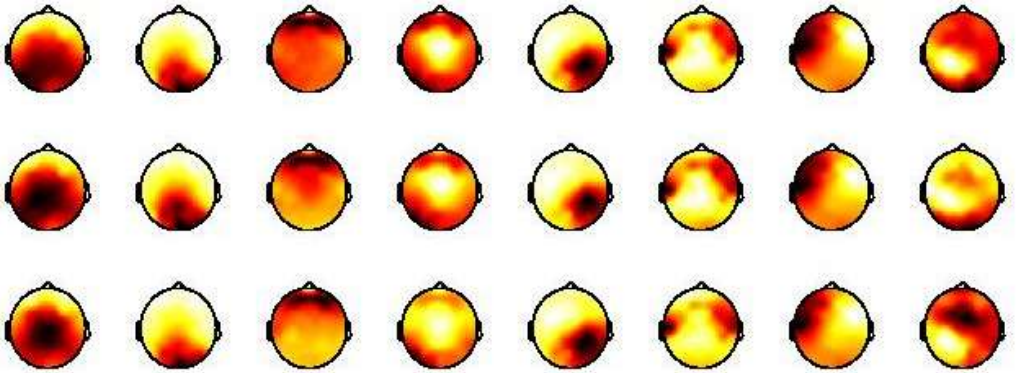
```
>> [W,D] = eig(X*X',C);
>> S = W' * X;
```

Example: ICA in Electro-Encephalography

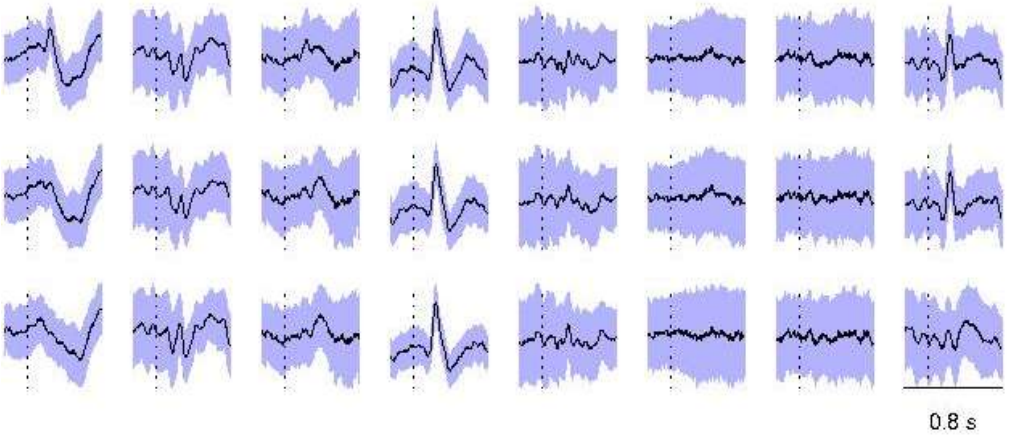
Interestingly, diagonalization algorithms based on these three different properties of the signals give often similar results in EEG:

- non-white
- Non-Gaussian
- non-stationary

EEG sensor projections $\mathbf{A}=\mathbf{W}^{-1}$



Trial averaged components $\mathbf{s}(t) = \mathbf{W}\,\mathbf{x}(t)$



This supports the interpretation of EEG as linear superposition of independent signals.

Acoustic Mixing and Separation

Acoustic environments are characterized by multi-path propagation which can be represented as a convolutive mixture:

$$\boldsymbol{x}(t) = \sum_{l} \boldsymbol{A}(l)\,\boldsymbol{s}(t-l)$$

Source $s_i(t)$ is coupled to microphone $x_j(t)$ by the room response $A_{ji}(l)$.

The goal of acoustic source separation is to find filters $W_{ij}(l)$ that generate model sources $y_i(t)$, corresponding to the sources $s_i(t)$

$$\boldsymbol{y}(t) = \sum_{l} \boldsymbol{W}(l)\,\boldsymbol{x}(t-l)$$

Frequency Domain Separation

For efficiency reasons this problem is often transformed into the frequency domain:

$$\boldsymbol{x}(t)=\boldsymbol{A}(t)*\boldsymbol{s}(t) \quad \Leftrightarrow \quad \boldsymbol{x}(\omega)=\boldsymbol{A}(\omega)\boldsymbol{s}(\omega)$$

Source separation aims to find **filters** $\boldsymbol{W}(\omega)$ such that model source $\boldsymbol{y}(\omega)$ correspond to the original sources $\boldsymbol{s}(\omega)$.

$$\boldsymbol{y}(\omega)=\boldsymbol{W}(\omega)\boldsymbol{A}(\omega)\boldsymbol{s}(\omega)$$

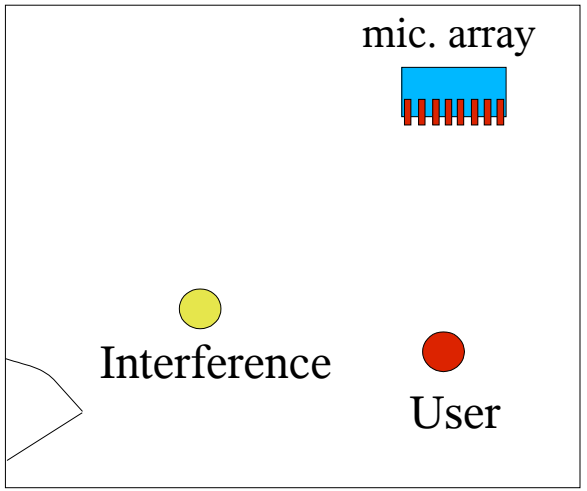
Separation based on Non-Stationarity

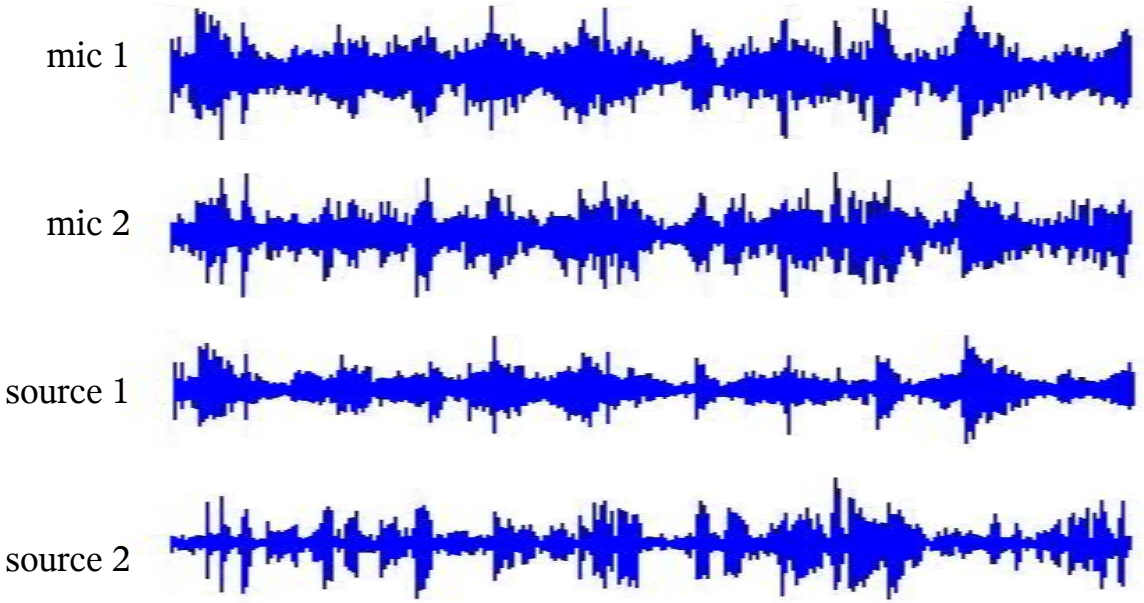
For non-stationary signals this can be achieved by minimizing cross-powers $\boldsymbol{R}_{yy}(t,\omega)=\mathrm{E}[\boldsymbol{y}(\omega)\boldsymbol{y}^{\mathrm{H}}(\omega) \mid \mathrm{t}\ ]$ estimated over multiple estimation periods:

$$\underset{\boldsymbol{W}(\omega)}{argmin} \sum_t \left\| offdiag\left(\boldsymbol{R}_{yy}(t,\omega)\right)\right\|^2$$

$$\boldsymbol{R}_{yy}(t,\omega)=\boldsymbol{W}(\omega)\boldsymbol{R}_{xx}(t,\omega)\boldsymbol{W}^H(\omega)$$

Separation for Speech Controlled TV





Ambiguities of the Separation Criteria

Independence criteria specifies result up to permutation $\boldsymbol{P}(\omega)$ and scaling $\boldsymbol{D}(\omega)$

$$\boldsymbol{W}(\omega)\boldsymbol{A}(\omega)=\boldsymbol{P}(\omega)\boldsymbol{D}(\omega)$$

- Convolution: Scaling $\boldsymbol{D}(\omega)$ corresponds to convolution of each source.
- Permutation: $\boldsymbol{P}(\omega)$ may be different for every frequency bin!
- Subspace: $N \leq M$ dimensional space of equivalent solutions $\boldsymbol{W}(\omega)$.

Convolutive BSS Summary

State-of-the-art **(?)**: **Frequency domain SOS** achieving **15dB-20dB** interference reduction for mixtures of up to **3 sources**.

Second order statistics (SOS) sufficient for non-stationary signals such as speech.

SOS algorithms easier to implement.

Frequency domain formulations give fast algorithms.

However SOS in the frequency domain leads to frequency permutation ambiguity.

Criteria that have been proposed to fix this problem are:

- Limit support in time domain (smooth spectrum)
- Exploit frequency co-modulation
- Impose geometric constraints

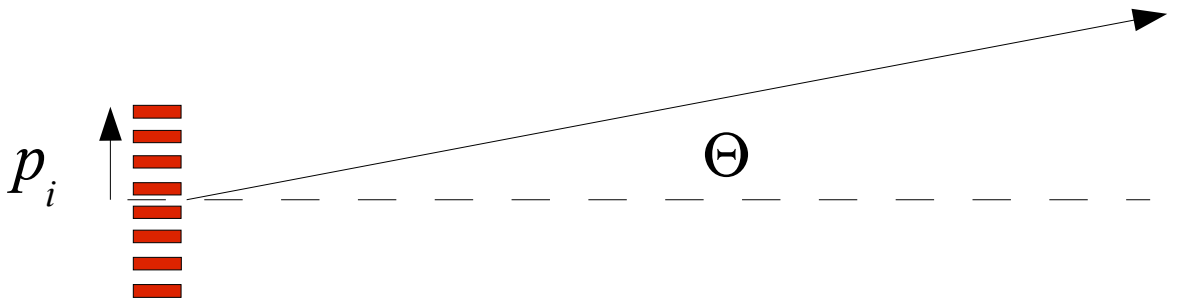
Farfield Beam Response

In order to interpret the resulting filter geometrically consider the farfield array response $\boldsymbol{r}(\Theta,\omega)$ for filters $\boldsymbol{W}(\omega)$:

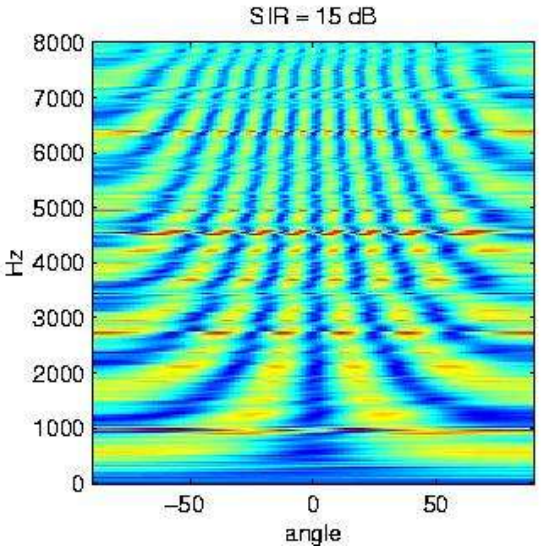
$$\boldsymbol{r}(\Theta,\omega)=\boldsymbol{W}(\omega)\boldsymbol{d}(\Theta,\omega)$$

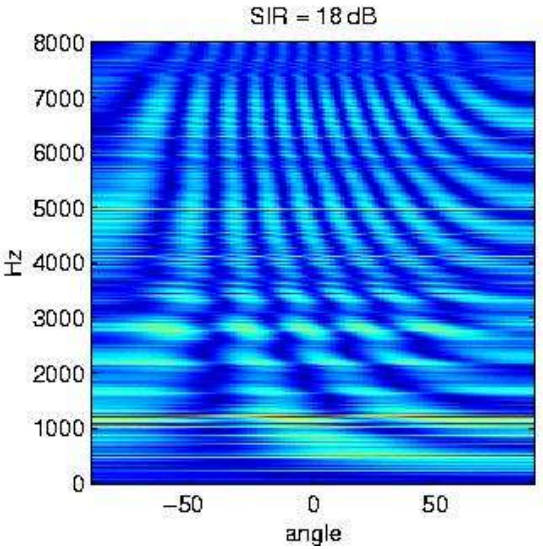
Array vector for microphones at positions p_i :

$$d_i(\Theta,\omega)=\exp\left(j\frac{2\pi\omega}{f}\frac{p_i}{c}\sin(\Theta)\right)$$



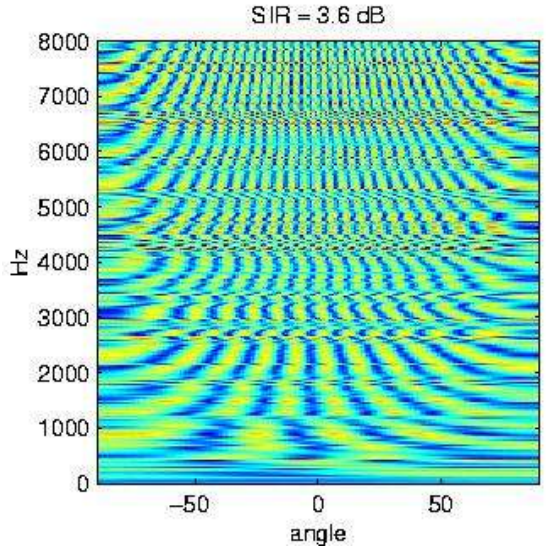
Example Beam Patterns - 2 Microphones

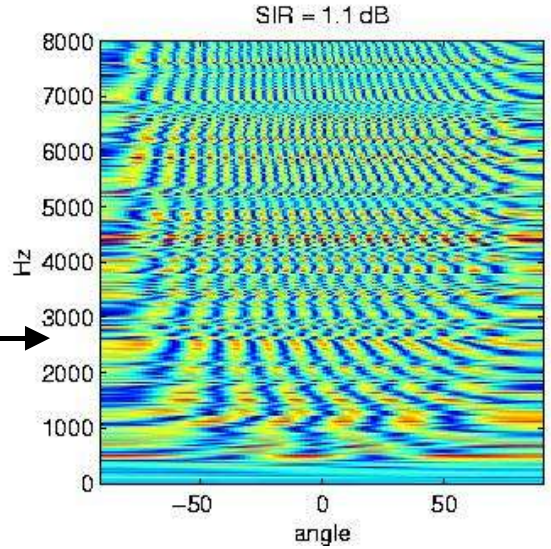




Two users at 2.5 m distance from array and 2 mics with 30 cm aperture.

Notice nulls consistent across frequency bands at user positions 0º and -45º.





Same signal and user positions with 70cm aperture.

Notice "permutation" in lower frequency band.

Numerical kurtosis measured on 15s of data is higher than above.

Beamforming: Geometry as Linear Constraints

To avoid trivial solution, $W(\omega)=0$, normalization is required, e.g.

In MSC and often in BSS we keep gain of one microphone constant:

$$W(\omega)\boldsymbol{e}_i=1$$

In linearly constraint minimum variance (LCMV) or generalized sidelobe canceling (GSC) keep one orientation constant:

$$W(\omega)\boldsymbol{d}(\omega,\Theta)=1$$

Or in some robust adaptive beam-forming the angle response is constraint to change slowly with direction:

$$W(\omega)\partial_\Theta \boldsymbol{d}(\omega,\Theta)=0$$

Power vs. Cross-Power minimization

Second order separation and adaptive beamforming have similar criteria:

Beam power and cross-power is given by

$$R_{yy}(\omega) = W(\omega) R_{xx}(\omega) W^H(\omega)$$

Adaptive beam forming minimizes power -*does not allow crosstalk*

$$\underset{W(\omega)}{argmin}\, diag\left(R_{yy}(\omega)\right)$$

Decorrelation algorithms minimizes cross-power - *allows crosstalk*

$$\underset{W(\omega)}{argmin} \left\| offdiag\left(R^t_{yy}(\omega)\right)\right\|^2$$

Combination of BSS with Beamforming

Source Separation based on second order non-stationarity

Minimum Cross-Power over time: $\underset{W(\omega)}{argmin} \sum_t \left\| offdiag[R^t_{yy}(\omega)] \right\|^2$

"Geometrically *initialized* Source Separation":

Delay-sum initialization $W_{init}(\omega) = \boldsymbol{d}^{\boldsymbol{H}}(\omega, \Theta)$

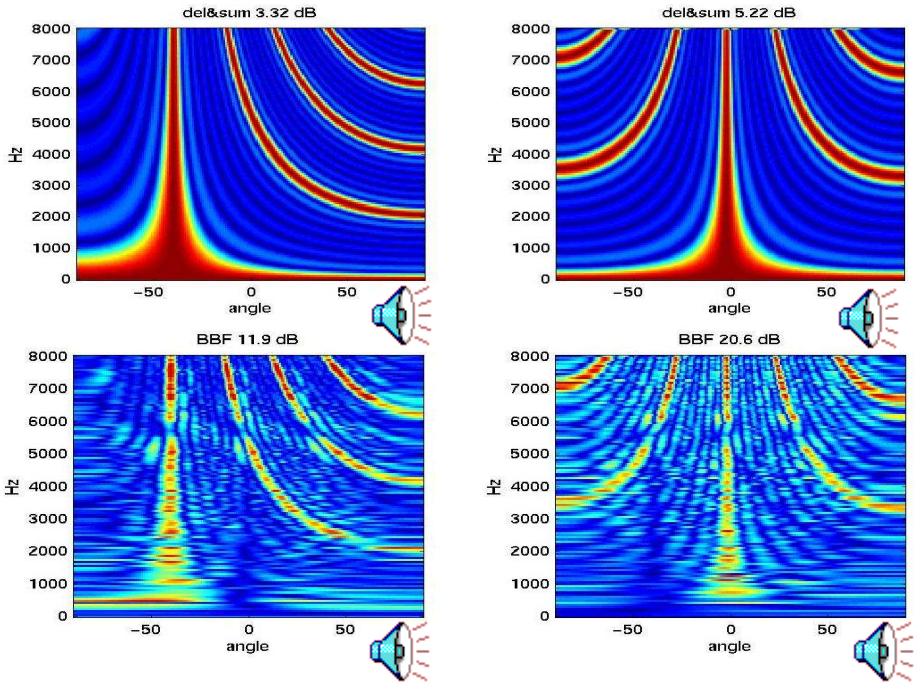
"Geometrically *constraint* Source Separation":

Unit-gain constrained in Θ $W(\omega)\boldsymbol{d}(\omega, \Theta) = 1$

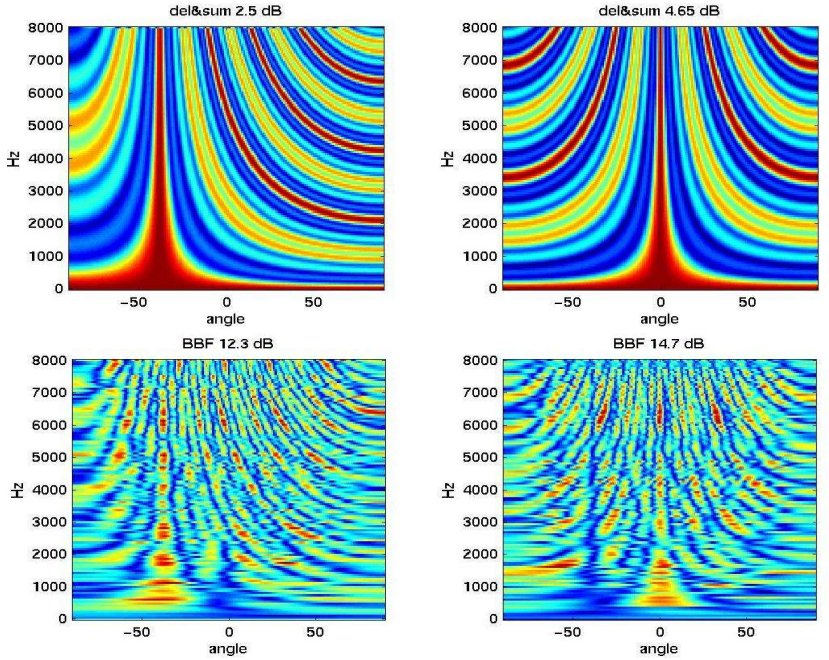
BSS with Geometric Initialization

Cross-power minimization across time can be made robust by initializing the filter structure with delay-sum beams, i.e. for each channel a maxima in the orientation of one of the sources.

8 microphones



4 microphones

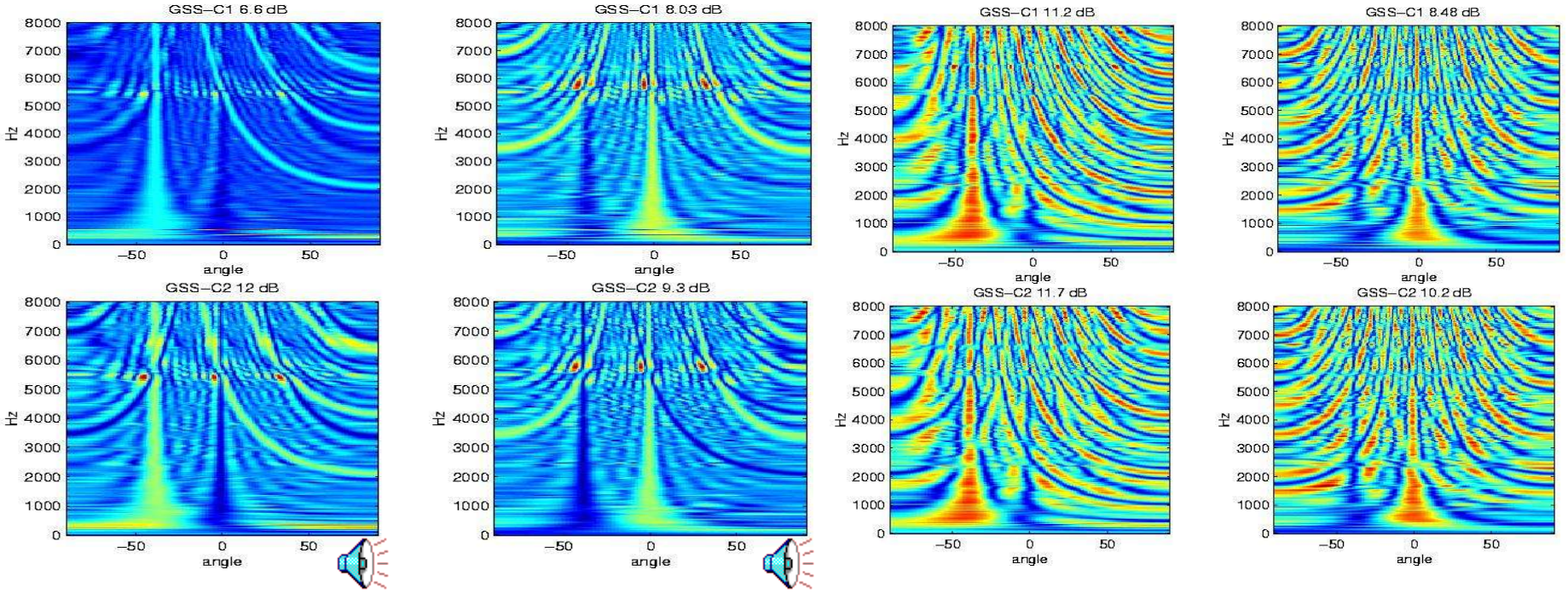


BSS with 'soft' Geometric Constraints

Cross-power minimization across time with 'soft' constraints on the response for given angles. Here we used constant unit-gain for user locations, and zero-gain for interference locations.

8 microphones

4 microphones

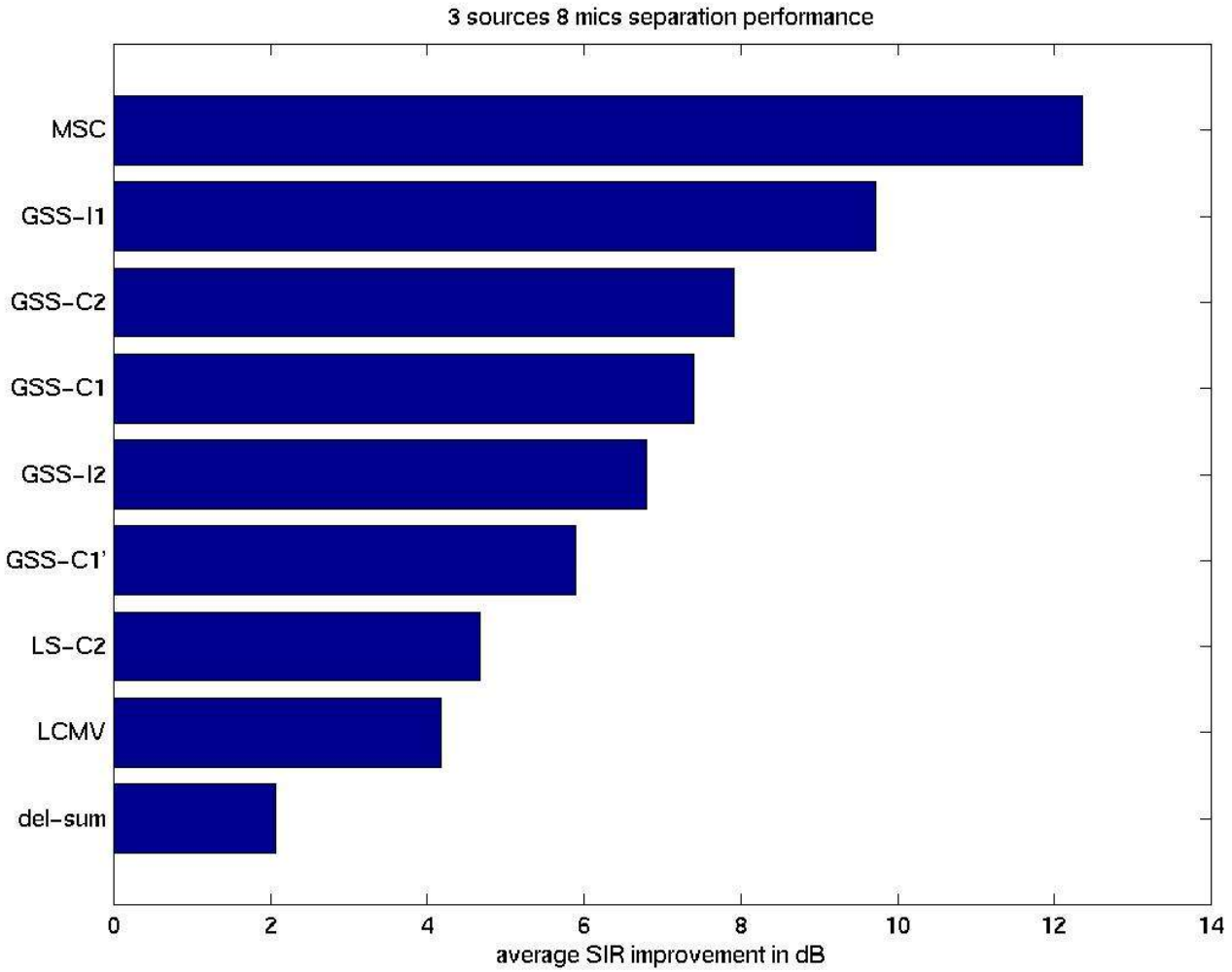


Acknowledgement

This work was performed 2000-2002 while working at

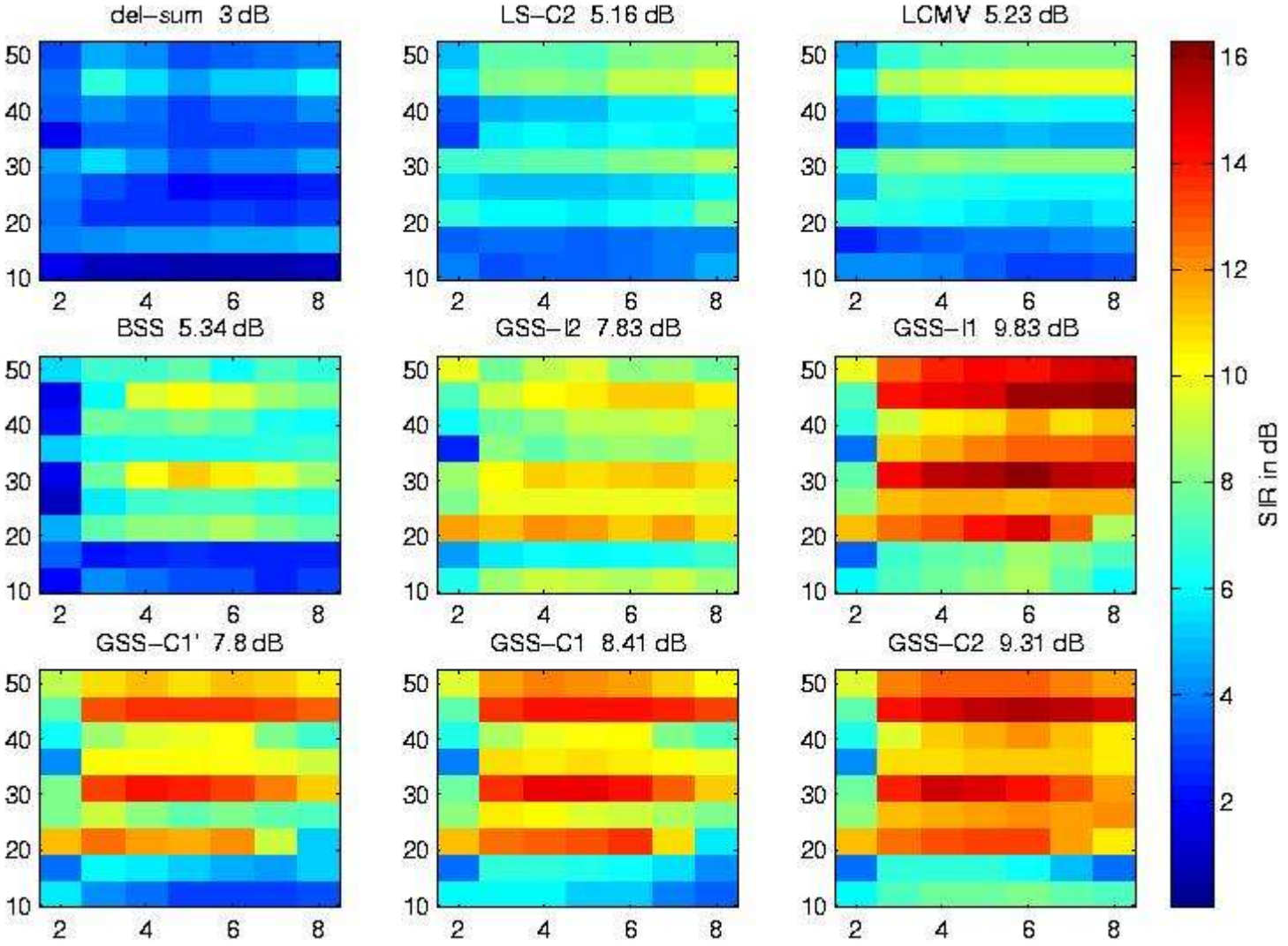
Clay Spence, Sarnoff Corporation
Chris Alvino, Rutgers
Craig Fancourt, Sarnoff Corporation

Performance Comparison 3 Sources



Real room performance comparison

... 2 sources for variable locations and number of microphones ...



Separation Based on Independence

Second order non-stationary

$$\boldsymbol{R}_{\boldsymbol{x}}(t)=E[\boldsymbol{x}(t)\boldsymbol{x}^{\boldsymbol{T}}(t)]=\boldsymbol{A}\,E[\boldsymbol{s}(t)\boldsymbol{s}^{\boldsymbol{T}}(t)]\,\boldsymbol{A}^{\boldsymbol{T}}=\boldsymbol{A}\,\boldsymbol{R}_{\boldsymbol{s}}(t)\,\boldsymbol{A}^{\boldsymbol{T}}$$

Second order non-white

$$\boldsymbol{R}_{\boldsymbol{x}}(l)=E[\boldsymbol{x}(t+l)\boldsymbol{x}^{\boldsymbol{T}}(t)]=\boldsymbol{A}\,E[\boldsymbol{s}(t+l)\boldsymbol{s}^{\boldsymbol{T}}(t)]\,\boldsymbol{A}^{\boldsymbol{T}}=\boldsymbol{A}\,\boldsymbol{R}_{\boldsymbol{s}}(l)\,\boldsymbol{A}^{\boldsymbol{T}}$$

4th order non-Gaussian

Assuming independence one can show that for any $k,l,\ i\neq j$

$$\begin{aligned}Cum(s_i,s_j,s_k,s_l)=&E[s_i s_j s_k s_l]-E[s_i s_j]E[s_k s_l]\\&-E[s_i s_k]E[s_j s_l]-E[s_i s_j]E[s_j s_k]=0\end{aligned}$$

Hence for any $i\neq j$ $\quad c_{i,j}(\boldsymbol{M})=\sum_{kl}Cum(s_i,s_j,s_k,s_l)m_{kl}=0$

or in matrix notation $\quad \boldsymbol{C}_{\boldsymbol{s}}(\boldsymbol{M})=diag$

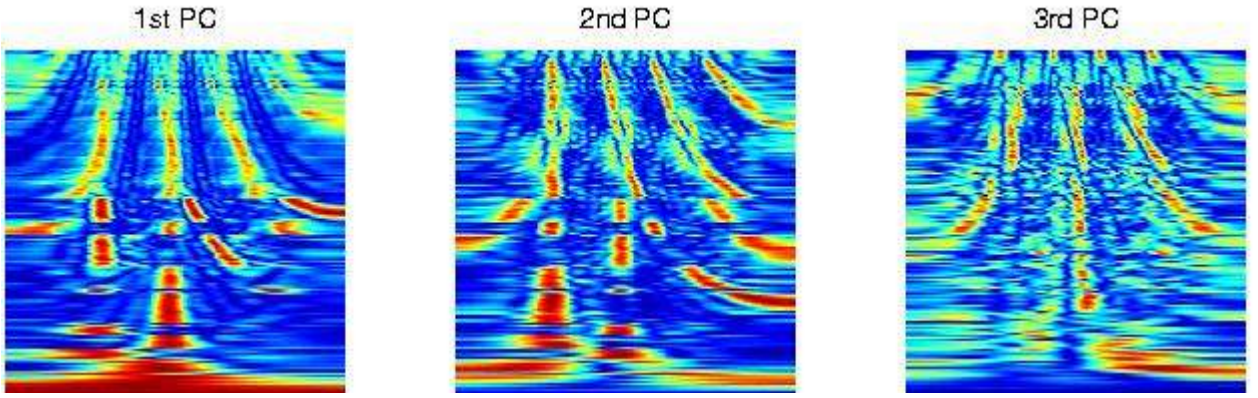
after some algebra $\quad \boldsymbol{C}_{\boldsymbol{x}}(\boldsymbol{I})=\boldsymbol{A}\,\boldsymbol{C}_{\boldsymbol{s}}(\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{A})\,\boldsymbol{A}^{\boldsymbol{T}}$

Source Subspace

... extract signal from noise from additional microphones ...

N□M dimensional space of equivalent solutions for $\boldsymbol{W}$(□) because there are more observations than signals. Signal exists in a M dimensional subspace.

Recover source subspace with PCA for each □❍



Problem:

- Orthogonality within source subspace may be meaningless in terms of beam geometry.
- Sources of interest may have small powers.