Blind Speech Separation based on Independent Component Analysis and Sparse Component Analysis

Shoji Makino
University of Tsukuba, Japan

Hiroshi Sawada
NTT CS Labs, Japan

Outline

- Part I

  Basic concepts of BSS and ICA

- Part II

  Detailed procedures for

  FD approach

Outline

- Part I
  - Basic concepts of BSS and ICA
  - Convolutive BSS
  - Frequency-domain approach
  - BSS and adaptive beamformer

April 1-4, 2003 Nara, Japan
General Chair: Shun-ichi Amari(RIKEN)
Organizing Chair: Shoji Makino(NTT)

Participants: 247
Oversea: 119
21 Countries

ICA2003

(**Fourth** International Symposium on **I**ndependent **C**omponent **A**nalysis and **B**lind **S**ignal **S**eparation)

April 1-4, 2003 Nara

General Chair: Shun-ichi Amari(RIKEN)
Organizing Chair: Shoji Makino(NTT)
Program Chair: Andrzej Cichocki(RIKEN)
Noboru Murata(Waseda)

May 14-19, 2006 Toulouse, France

Special Session on

Audio Source Separation
with CASA and ICA

Organizers: Shoji Makino
Hiroshi Sawada

April 15-20, 2007 Hawaii, USA

Tutorial on

Audio Source Separation based on Independent Component Analysis

Organizers: Shoji Makino

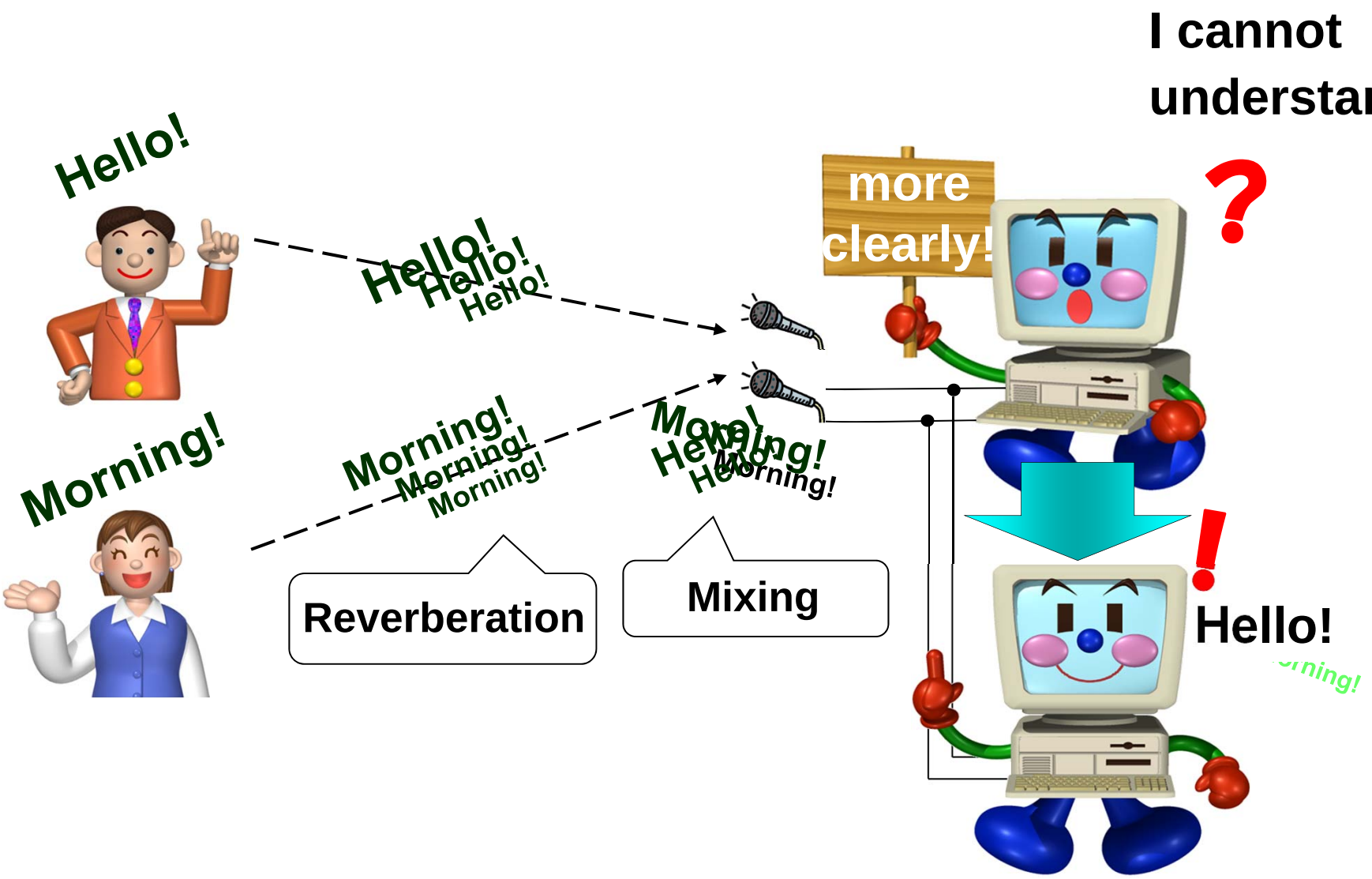
Hiroshi Sawada

Shou-Toku-Taishi

Could separate ten speeches.

At a Cocktail Party

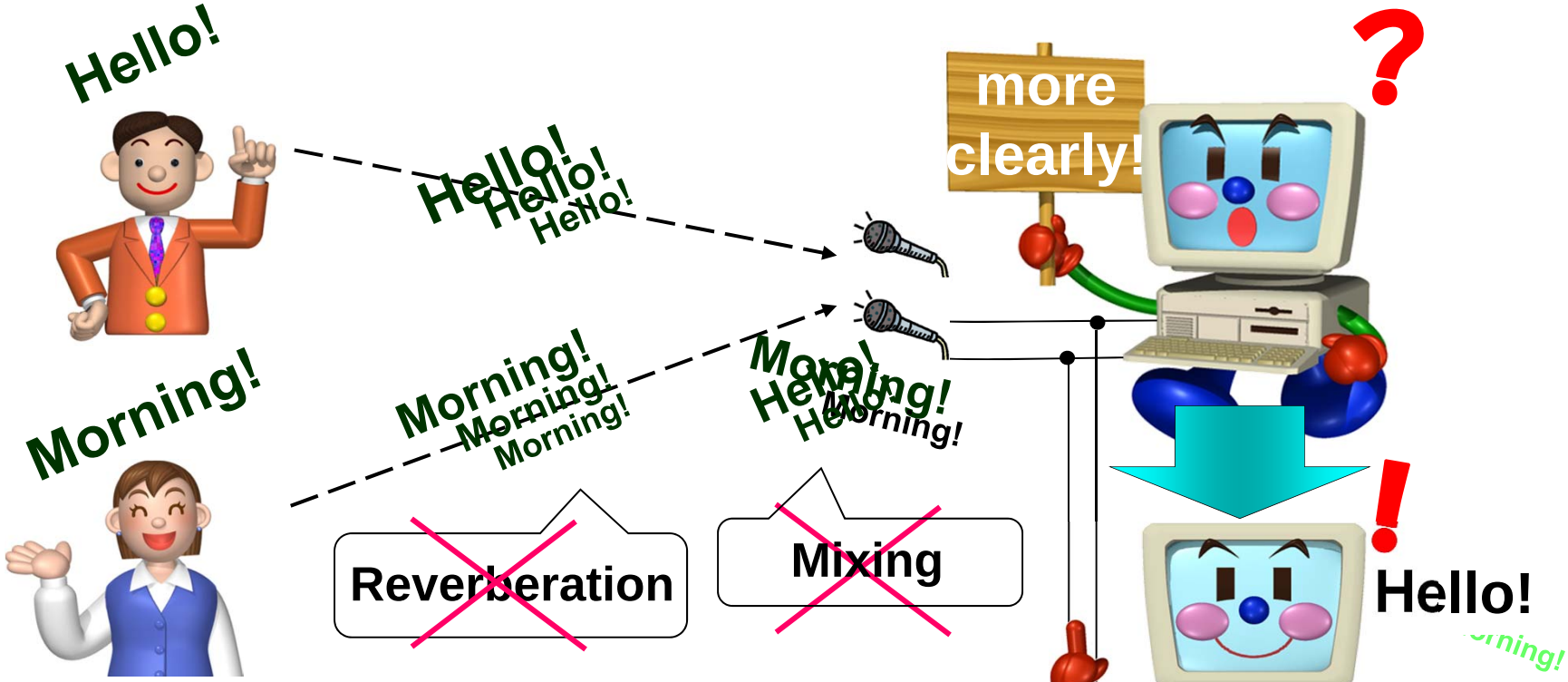
When two people talk to a computer



Blind Source Separation in a Real Environment

VIDEO

Task of Blind Source Separation

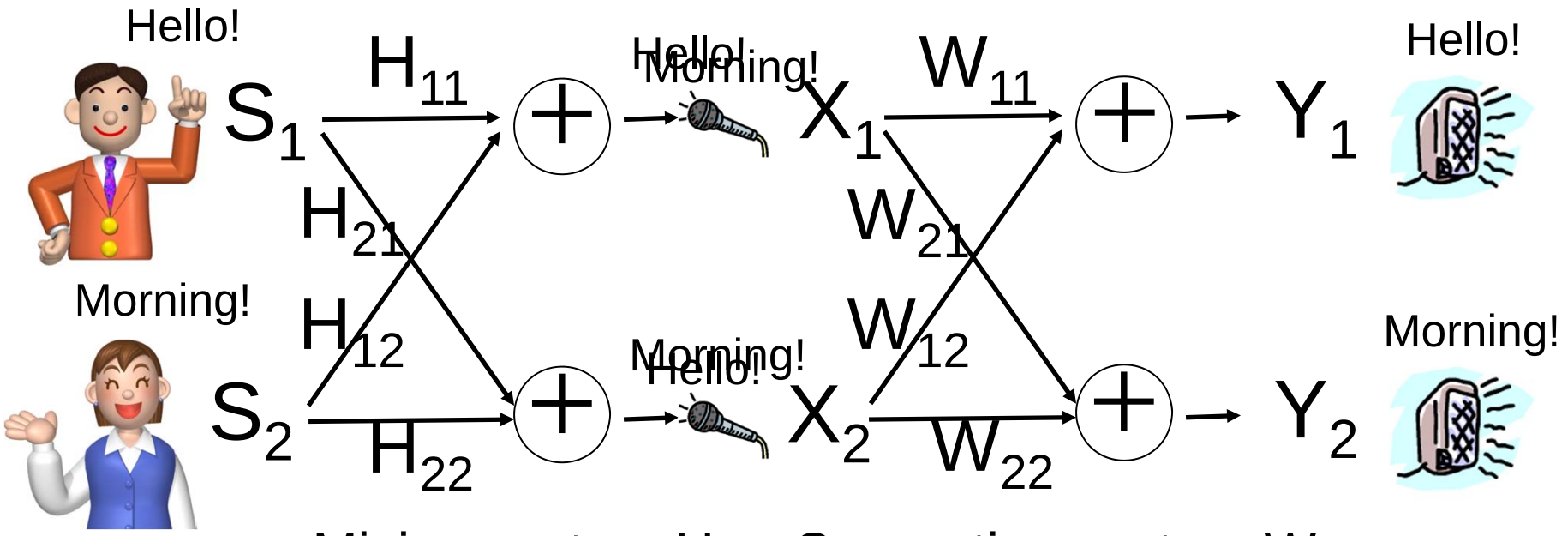


Speech recognition（Robots, Car navi. …）

People（TV conferences, Hearing aids, …）

Model of Blind Source Separation

Source signal S | **Observed signal X** | **Separated signal Y**



Source S

Independent

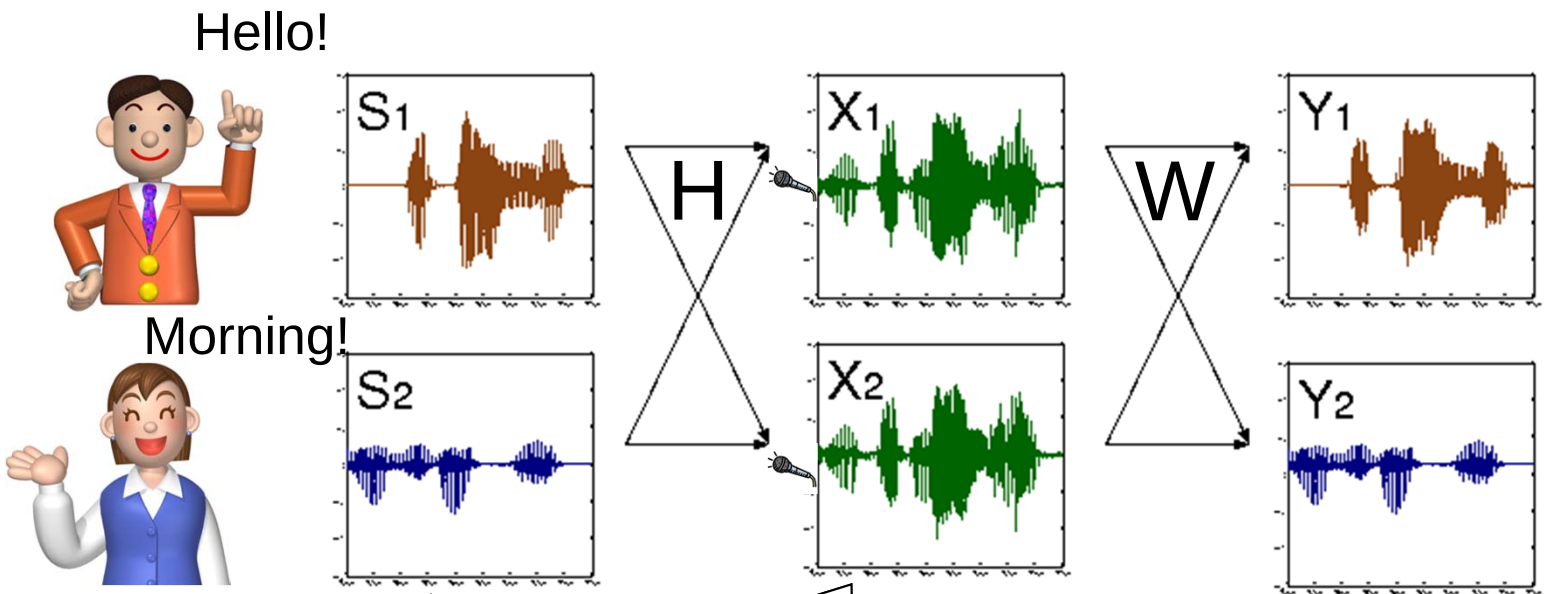
Mixing system H

Delay
Attenuation
Reverberation

Separation system W

Estimate S_1 and S_2 using only X_1 and X_2

Blind Source Separation using Independent Component Analysis



Source signals S_1 and S_2 are statistically independent

Observed signals X_1 and X_2 are correlated

Extract Y_1 and Y_2 so that they are mutually independent

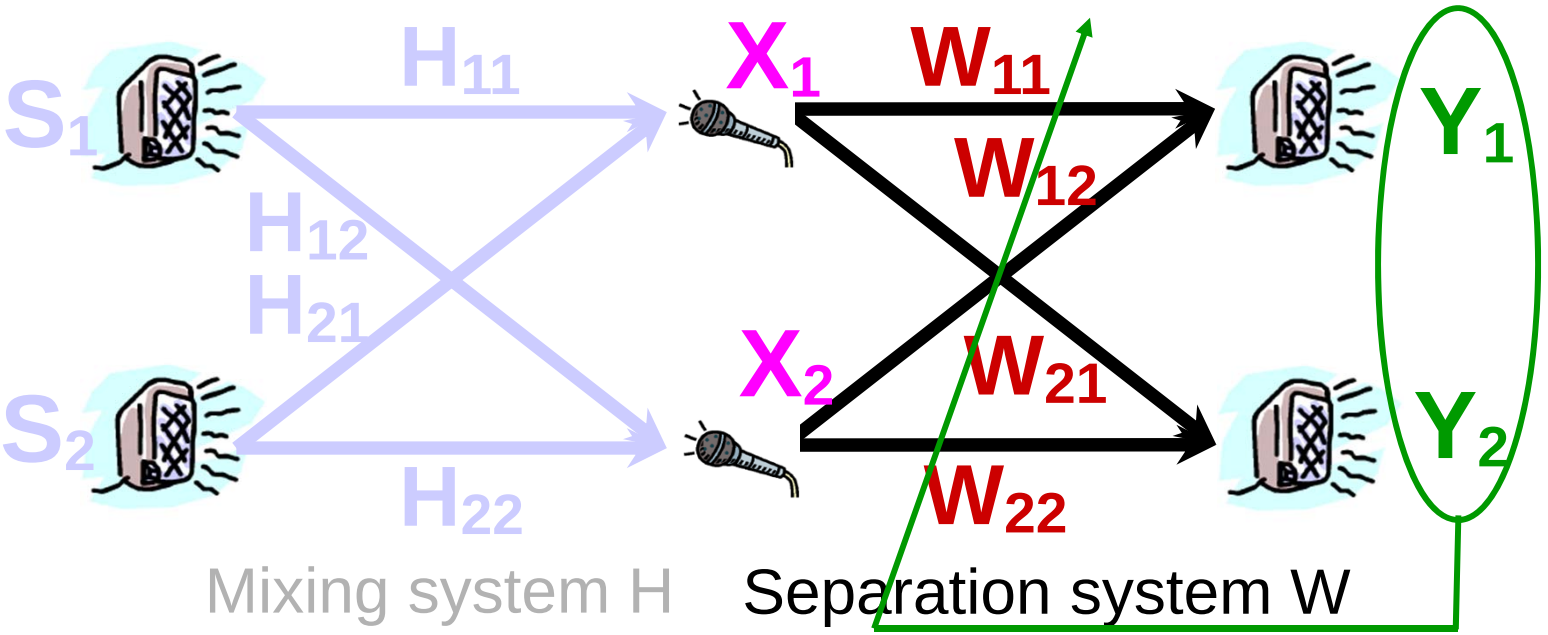
No need for information on the source signals or mixing system (location or room acoustics) ⇒ **Blind Source Separation**

Unsupervised Learning by ICA

Source signal S

Observed signal X

Separated signal Y



Source signals S_1 and S_2 are statistically independent

Estimate W so that Y_1 and Y_2 become mutually independent

What's ICA?

ICA: Independent Component Analysis

- Statistical method
- Neural Network, Communication

BSS: Blind Source Separation

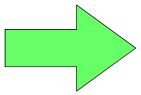
- Sounds → Speech Recognition
- Images → People
- CDMA wireless communication signals
- fMRI and EEG signals

Background Theory

- Minimization of Mutual Information
  (Minimization of Kullback-Leibler Divergence)

- Maximization of Non-Gaussianity

- Maximization of Likelihood



All solutions are identical

Background Theory

Minimization of Mutual Information

Maximization of Non-Gaussianity

Maximization of Likelihood

$$I(Y_1, Y_2) = \sum_{i=1}^{2} H(Y_i) - H(Y_1, Y_2)$$

Mutual Information

Marginal Entropy

Joint Entropy

All solutions are identical

H(·) : Entropy

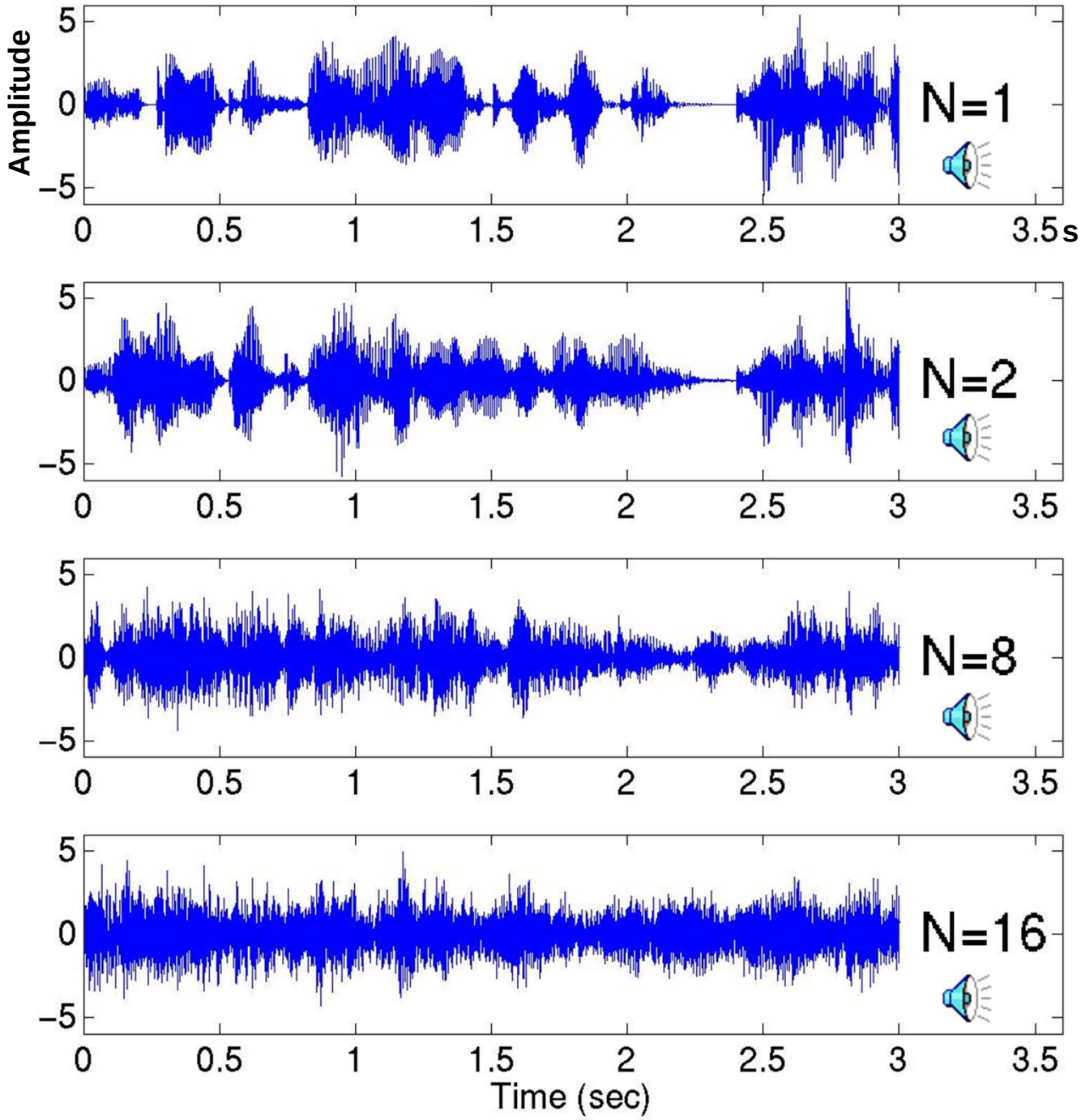
Background Theory

- Maximization of Non-Gaussianity

  - Make the output pdf away from Gaussian

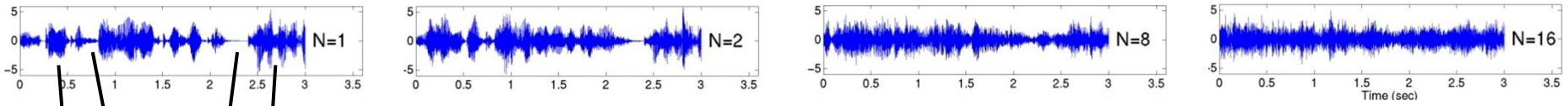
Central Limit Theorem

Mix independent components $\Rightarrow$ Gaussian

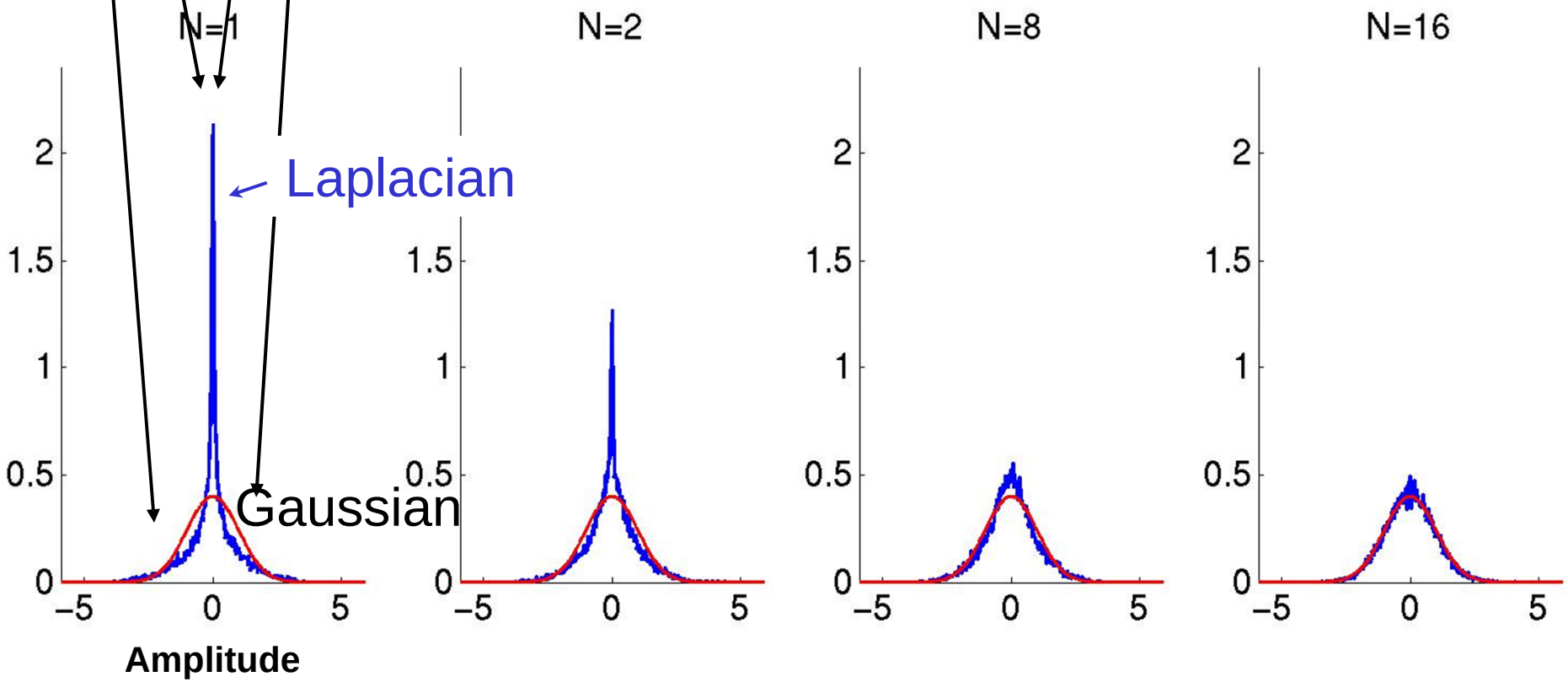


Wave forms

Wave forms



Histograms



Central Limit Theorem

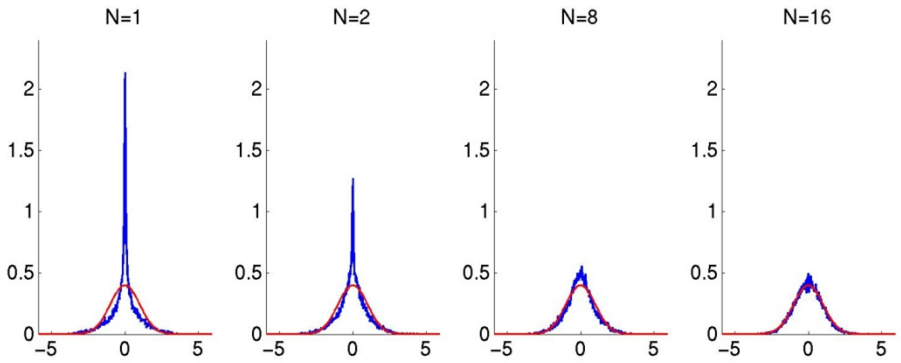
Mix independent components ⇒ Gaussian

Find independent component ⇒ Non-Gaussian

Non-Gaussianity measures

- Negentropy
- Kurtosis

Maximization of Negentropies

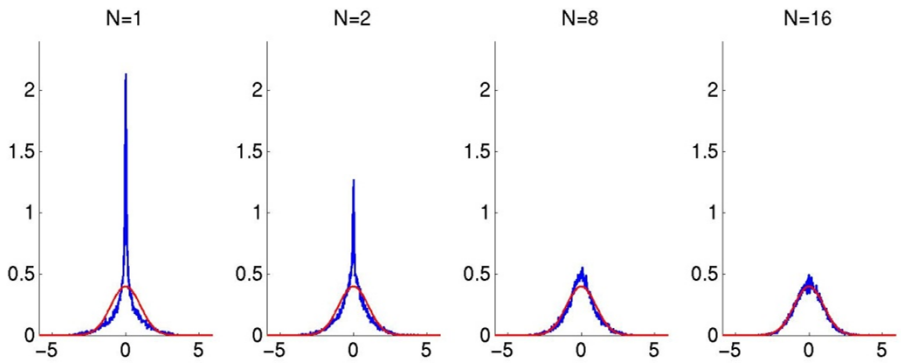


# sources N	1	2	8	16		Gaussian
Entropy *H*	1.19	1.33	1.39	1.40		1.41
Negentropy *N*	0.225	0.087	0.025	0.012		0

$$H(y) = \sum_{i=1}^{n} p_i \log \frac{1}{p_i}$$

$$N(y) = H(x_{\text{gauss}}) - H(y)$$

Maximization of Kurtosis



N	1	2	8	16		Gaussian
Kurtosis	2.1	1.8	0.70	0.39		0

$$kurt(y) = E\{|y|^4\} - 3(E\{|y|^2\})^2$$

FastICA

Newton's method

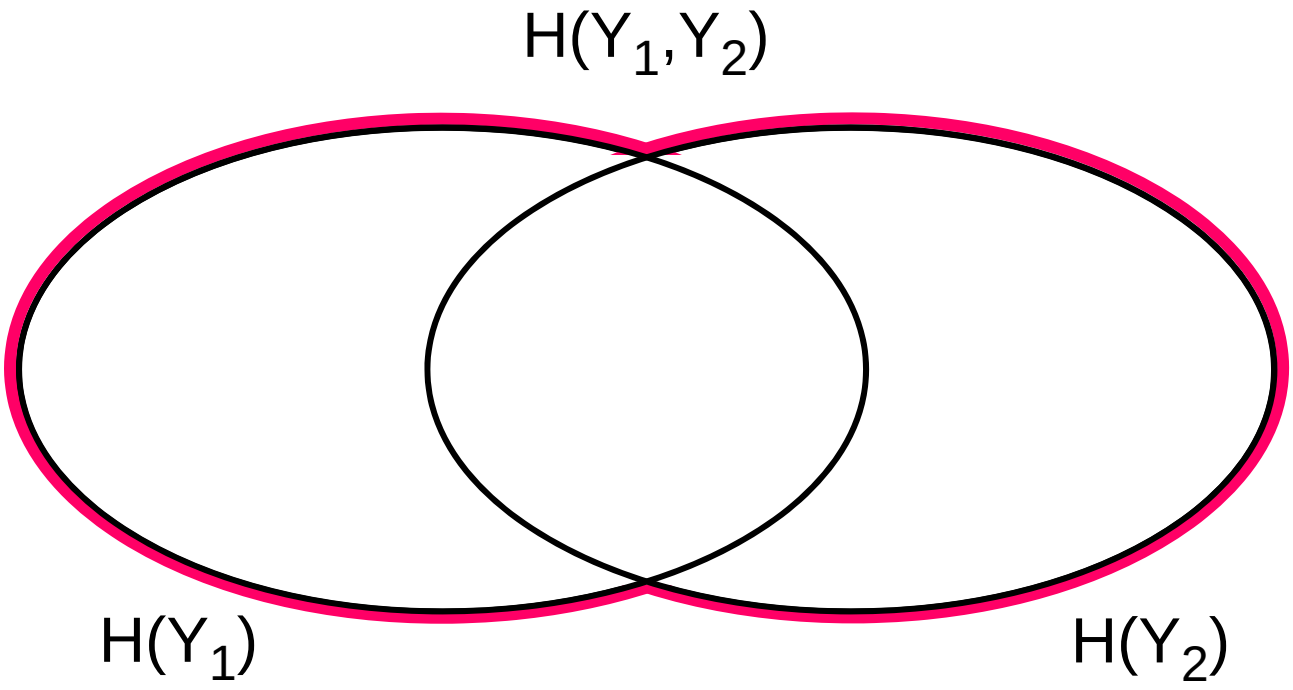
Background Theory

- Minimization of Mutual Information

  (Minimization of Kullback-Leibler Divergence)

  - Make the output “decorrelated”

Mutual Information



$$I(Y_1, Y_2) = \sum_{i=1}^{2} H(Y_i) - H(Y_1, Y_2)$$

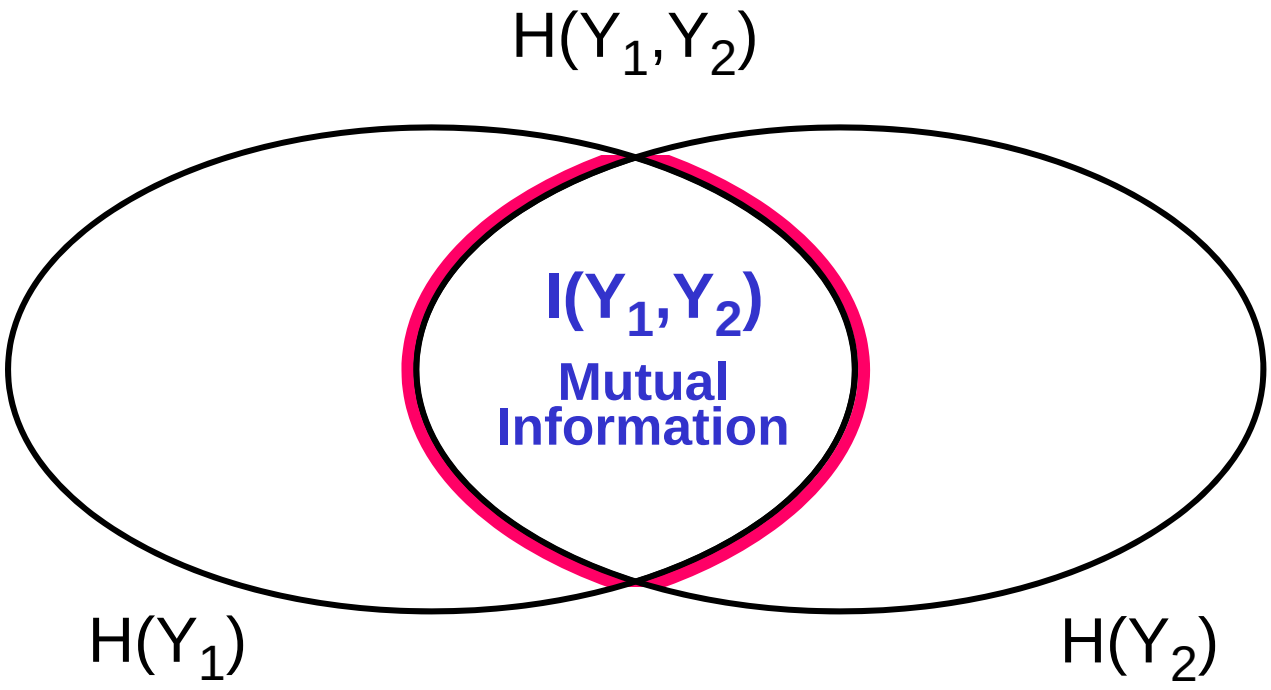
Mutual Information

Marginal Entropy

Joint Entropy

H(▪) : Entropy

Mutual Information



$$I(Y_1, Y_2) = \sum_{i=1}^{2} H(Y_i) - H(Y_1, Y_2)$$

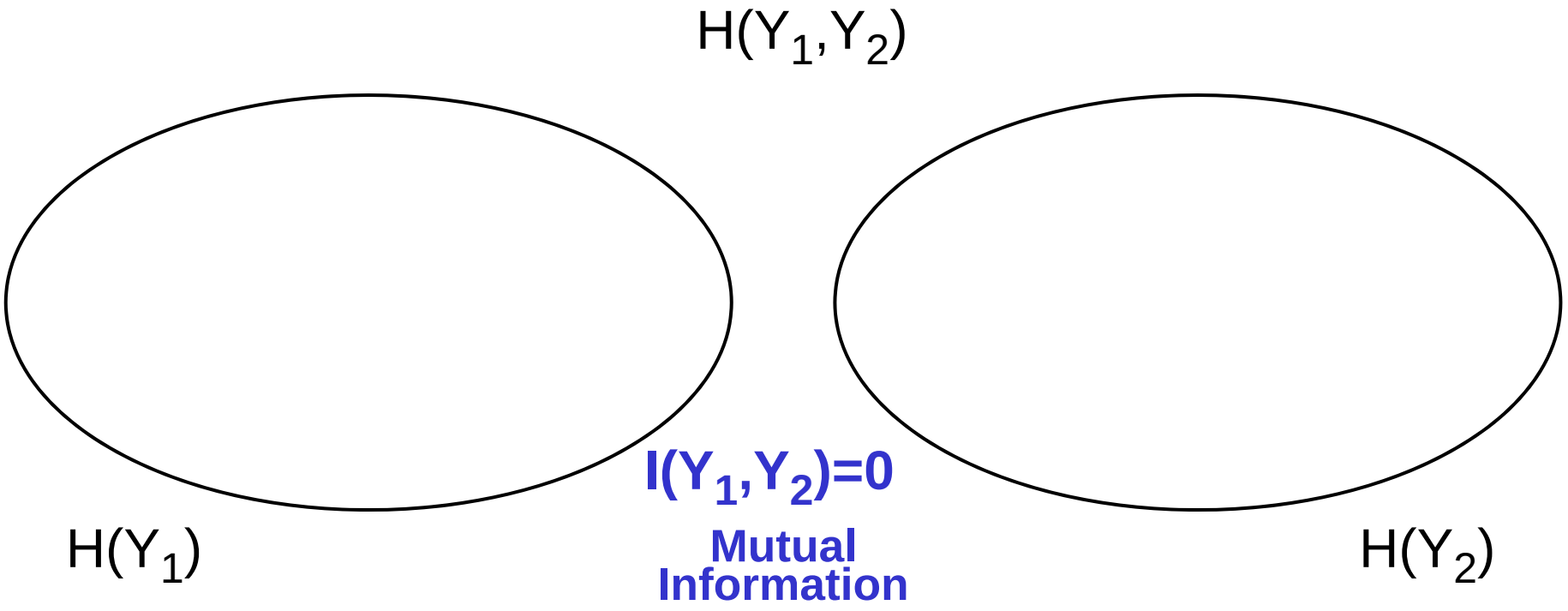
Mutual Information

Marginal Entropy

Joint Entropy

H(·) : Entropy

Minimization of Mutual Information



$$I(Y_1,Y_2)=\sum_{i=1}^{2}H(Y_i)-H(Y_1,Y_2)$$

Mutual Information

Marginal Entropy

Joint Entropy

$H(\cdot)$: Entropy

Minimization of Mutual Information

Mutual Information

Marginal Entropy

Joint Entropy

$$I(Y_1, Y_2) = \sum_{i=1}^{2} H(Y_i) - H(Y_1, Y_2)$$

$$= \int p(Y_1) \log \frac{1}{p(Y_1)} dY_1 + \int p(Y_2) \log \frac{1}{p(Y_2)} dY_2$$

$$- \int p(Y_1, Y_2) \log \frac{1}{p(Y_1, Y_2)} dY_1 dY_2$$

$$= \int p(Y_1, Y_2) \log \frac{p(Y_1, Y_2)}{p(Y_1)\, p(Y_2)} dY_1 dY_2$$

Kullbuck-Leibler Divergence

Minimization of Mutual Information

Mutual Information

Kullbuck-Leibler Divergence

$$I(Y_1, Y_2) = \int p(Y_1, Y_2) \log \frac{p(Y_1, Y_2)}{p(Y_1)\,p(Y_2)}\, dY_1 dY_2$$

- Search for ***W*** which minimizes Mutual Information ***I***
- Gradient of ***W*** can be derived by differenciating ***I*** with ***W***, using $y = Wx$;

Natural Gradient

$$\Delta \mathbf{W} \propto -\frac{\partial I(Y_1, Y_2)}{\partial \mathbf{W}} \mathbf{W}^{-\mathrm{T}}\mathbf{W} = \left(\mathbf{I} - \mathrm{E}[\boldsymbol{\phi}(\mathbf{Y})\mathbf{Y}^{\mathrm{T}}]\right)\mathbf{W}$$

InfoMax

Gradient method

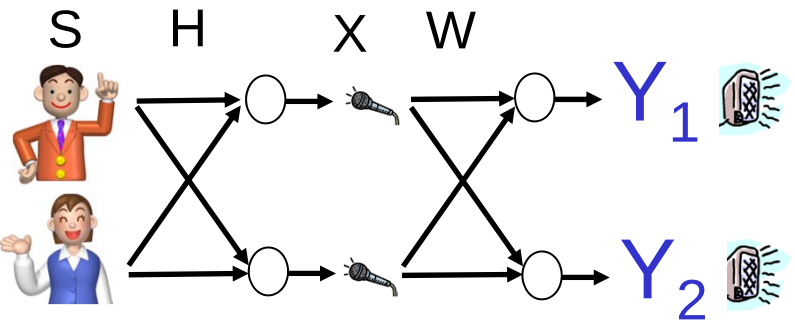
Laplacian

where $\boldsymbol{\phi}(\mathbf{Y}) = -\frac{\mathrm{d}\log p(\mathbf{Y})}{\mathrm{d}\mathbf{Y}}$

How can we separate speech?

Diagonalize R_Y

$$R_Y = \begin{bmatrix} \langle \phi(Y_1) Y_1 \rangle & \langle \phi(Y_1) Y_2 \rangle \\ \langle \phi(Y_2) Y_1 \rangle & \langle \phi(Y_2) Y_2 \rangle \end{bmatrix}$$



$\phi(\cdot)$ activation function

$\langle \cdot \rangle$ averaging operator

At the Convergence Point

Laplacian

$$\phi(Y_1) = -\frac{\mathrm{d}\log p(Y_1)}{\mathrm{d}Y_1}$$

Mutual independence

$$\begin{bmatrix} * & 0 \\ 0 & * \end{bmatrix} \qquad \langle \phi(Y_1) Y_2 \rangle = 0 \qquad \langle \phi(Y_2) Y_1 \rangle = 0$$

Average amplitude of Y

$$\begin{bmatrix} c_1 & * \\ * & c_2 \end{bmatrix} \qquad \langle \phi(Y_1) Y_1 \rangle = c_1 \qquad \langle \phi(Y_2) Y_2 \rangle = c_2$$

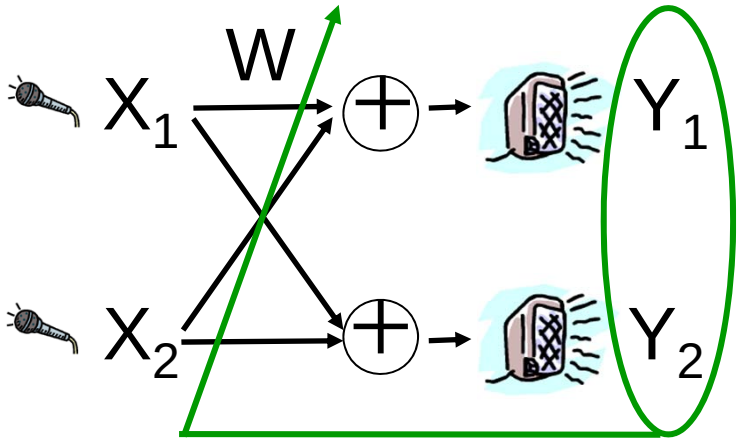
4 equations for 4 unknowns W_{ij}

ICA Learning Rule: InfoMax

$$W_{i+1} = W_i + \Delta W_i$$

$$R_Y = \begin{bmatrix} \langle \phi(Y_1) Y_1 \rangle & \langle \phi(Y_1) Y_2 \rangle \\ \langle \phi(Y_2) Y_1 \rangle & \langle \phi(Y_2) Y_2 \rangle \end{bmatrix}$$

$$\Delta W_i = \mu \begin{bmatrix} c_1 - \langle \phi(Y_1) Y_1 \rangle & \langle \phi(Y_1) Y_2 \rangle \\ \langle \phi(Y_2) Y_1 \rangle & c_2 - \langle \phi(Y_2) Y_2 \rangle \end{bmatrix} W_i \Rightarrow 0$$



Update W so that Y_1 and Y_2 become mutually independent

Second Order Statistics vs. Higher Order Statistics

Second Order Statistics (SOS)

colored sources

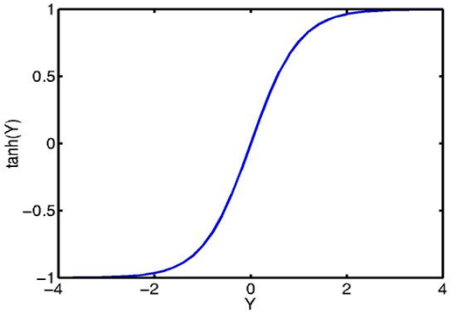
$$\phi(Y_1) = Y_1 \qquad \langle Y_1(t) Y_2(t+\tau_i) \rangle = 0$$

for multiple time delay

Higher Order Statistics (HOS)

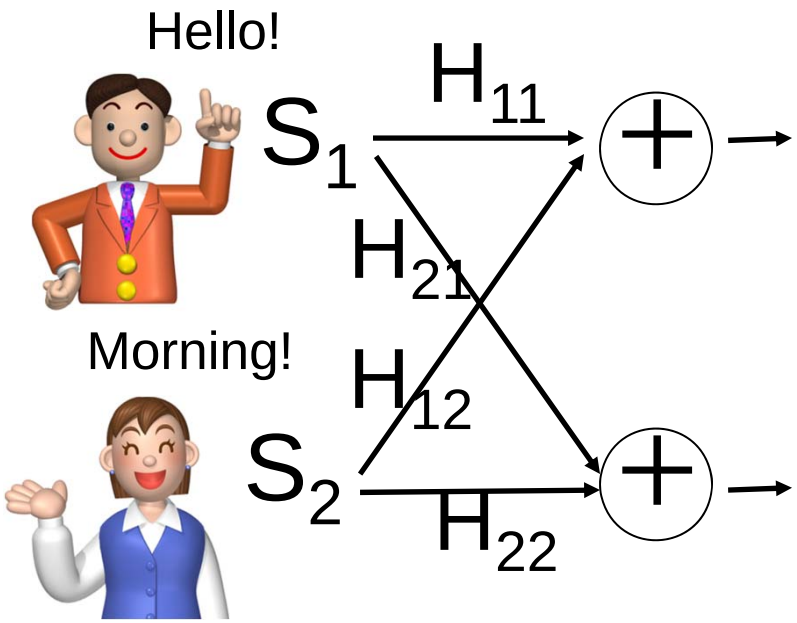
$$\phi(Y_1) = \tanh(Y_1) \qquad \langle \tanh(Y_1) Y_2 \rangle = 0$$

Tailor expansion



$$\langle Y_1 - \frac{Y_1^3}{3} + \frac{2Y_1^5}{15} \cdots) Y_2 \rangle = 0$$

Instantaneous vs. Convolutive



Instantaneous mixture

H_{ij} are **scalars**

- sounds with mixer
- images
- wireless communication signals
- fMRI and EEG signals

Well studied, good results

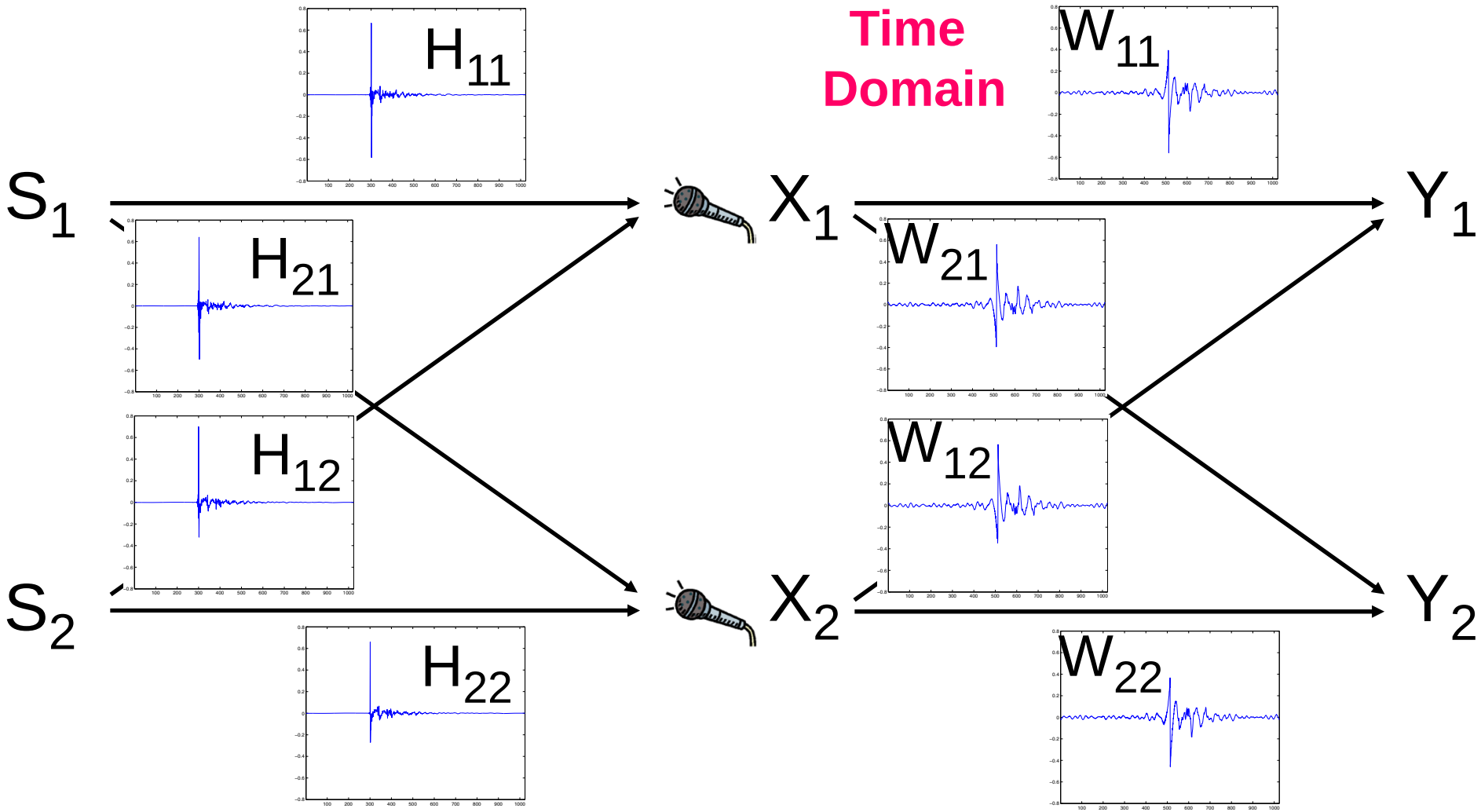
Convolutive mixture

H_{ij} are **FIR filters > 1000 taps**

- sounds in a room

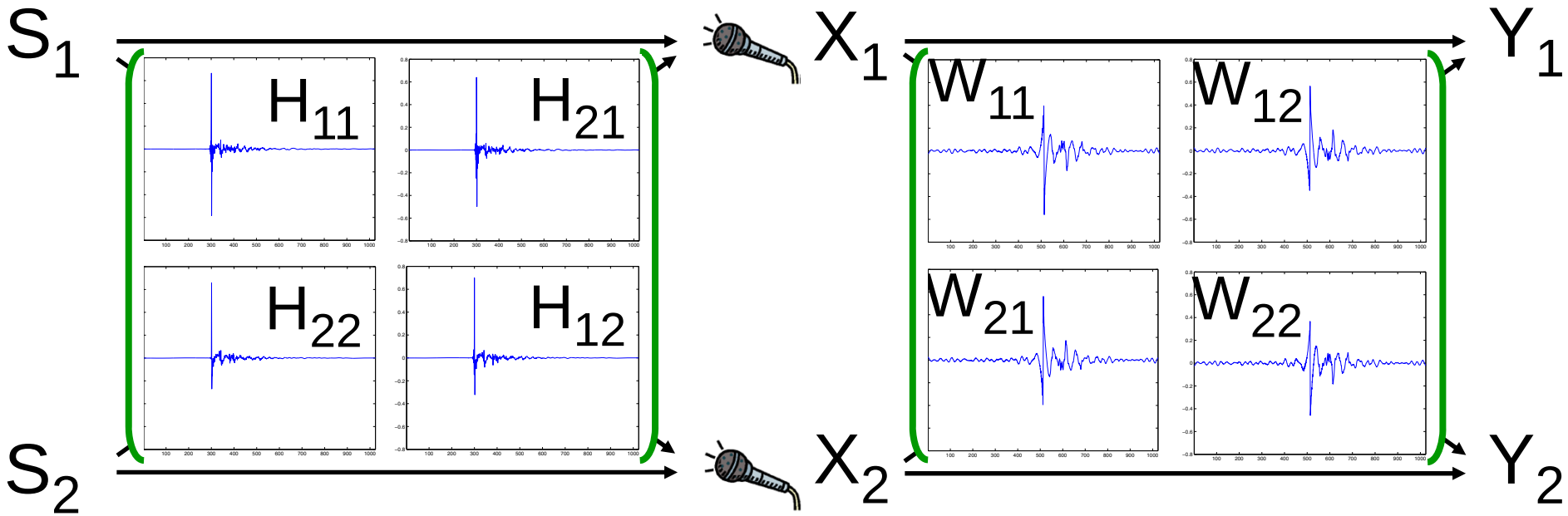
Difficult problem, relatively new

Mixing Filters and Separation Filters



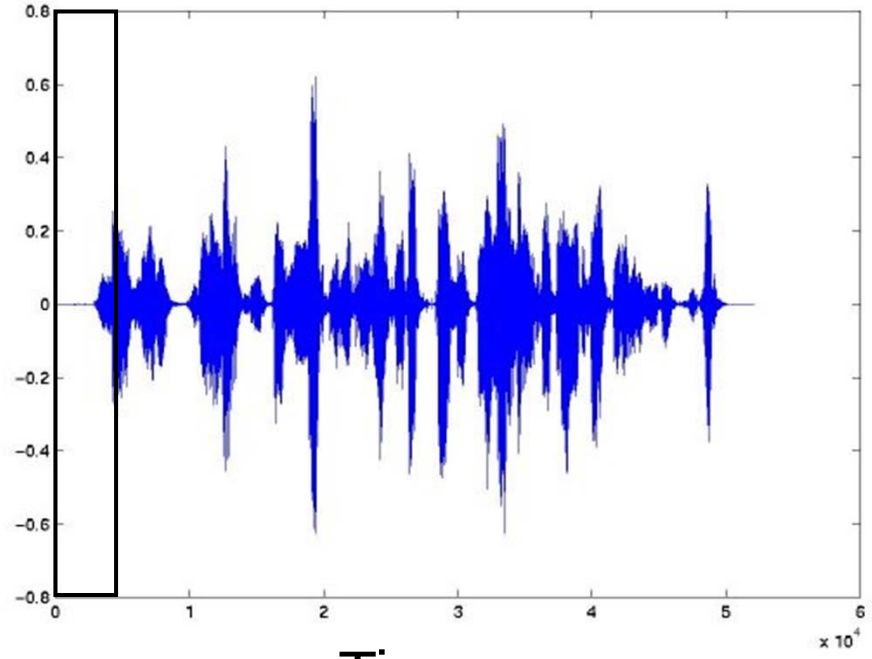
Mixing Filters and Separation Filters

Time Domain **in a Matrix**



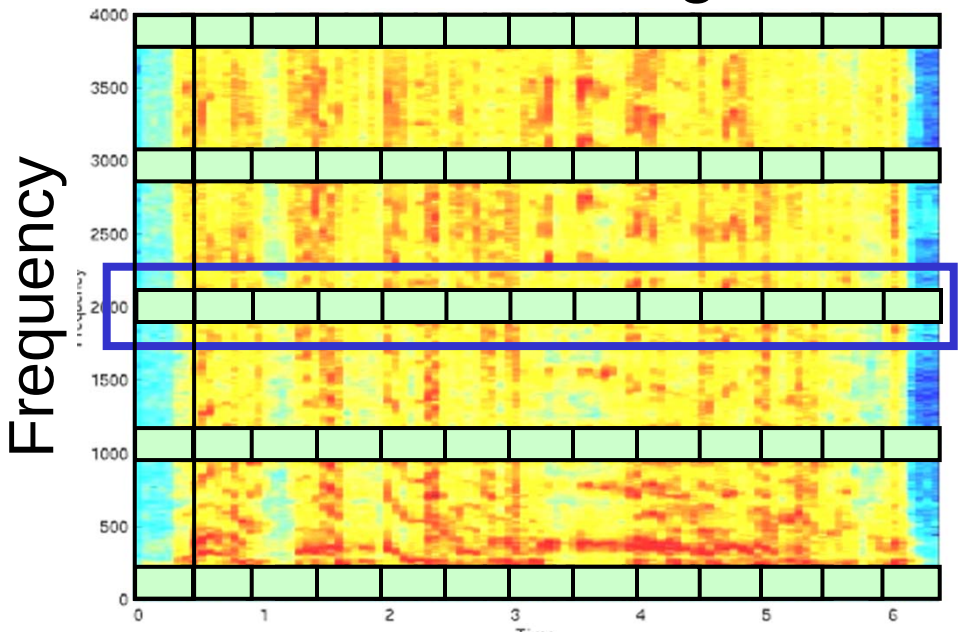
Short Time DFT (Spectrogram)

Time-domain signal



Time

Frequency-domain time-series signal

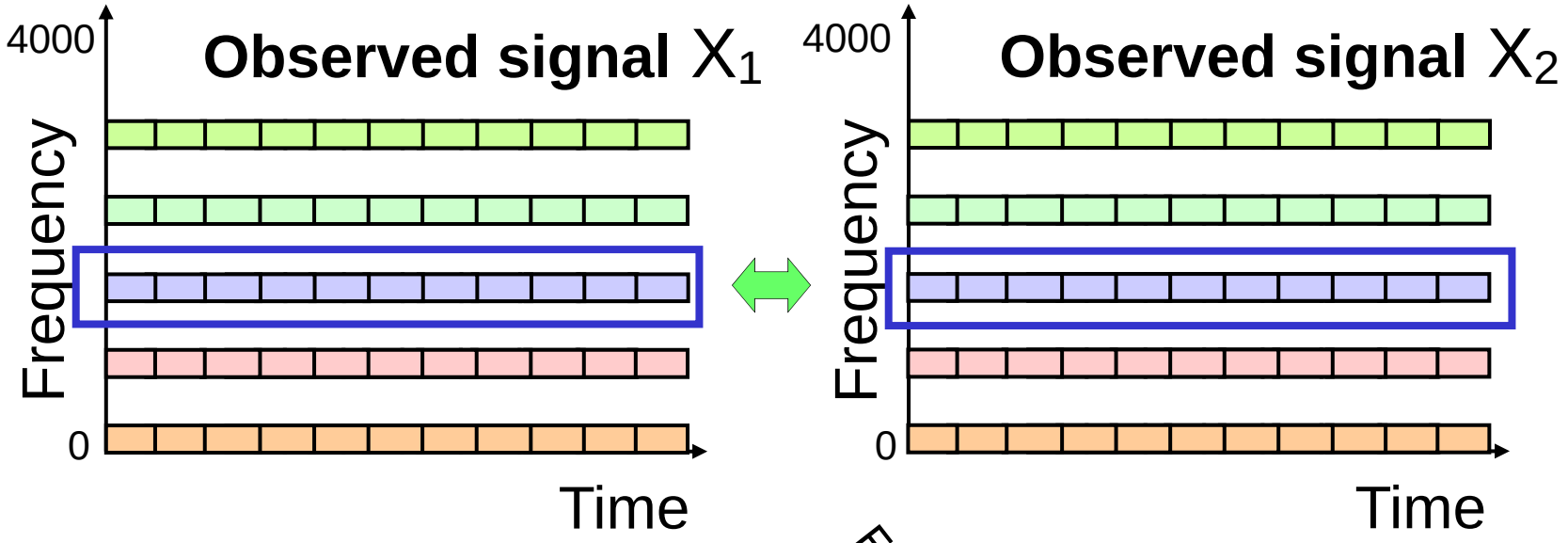


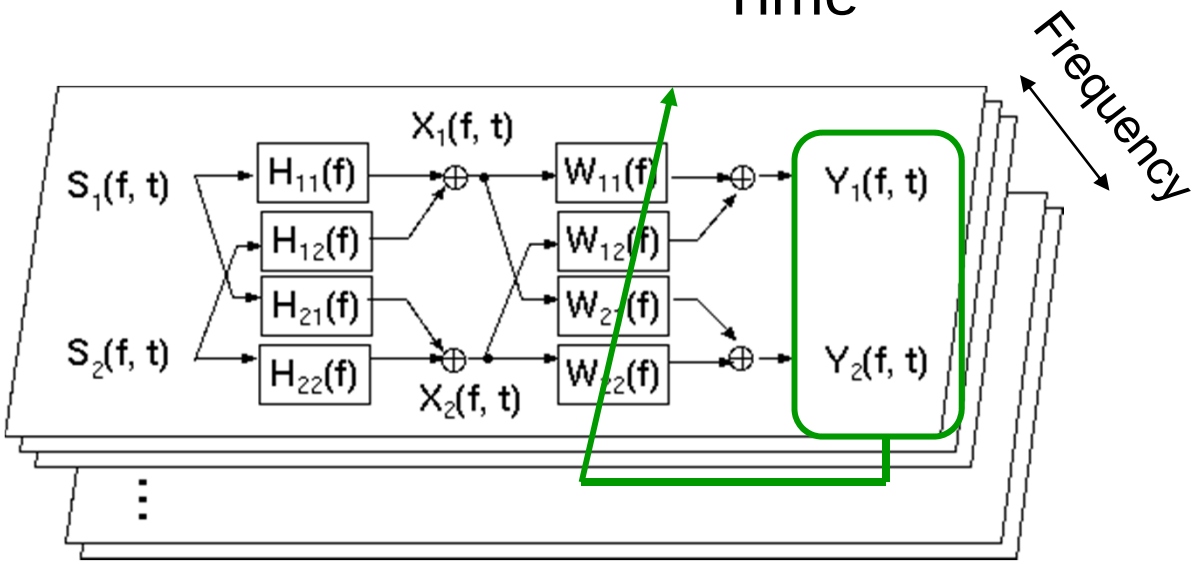
Time

Convolutive mixture in time domain

Multiple instantaneous mixtures in frequency domain

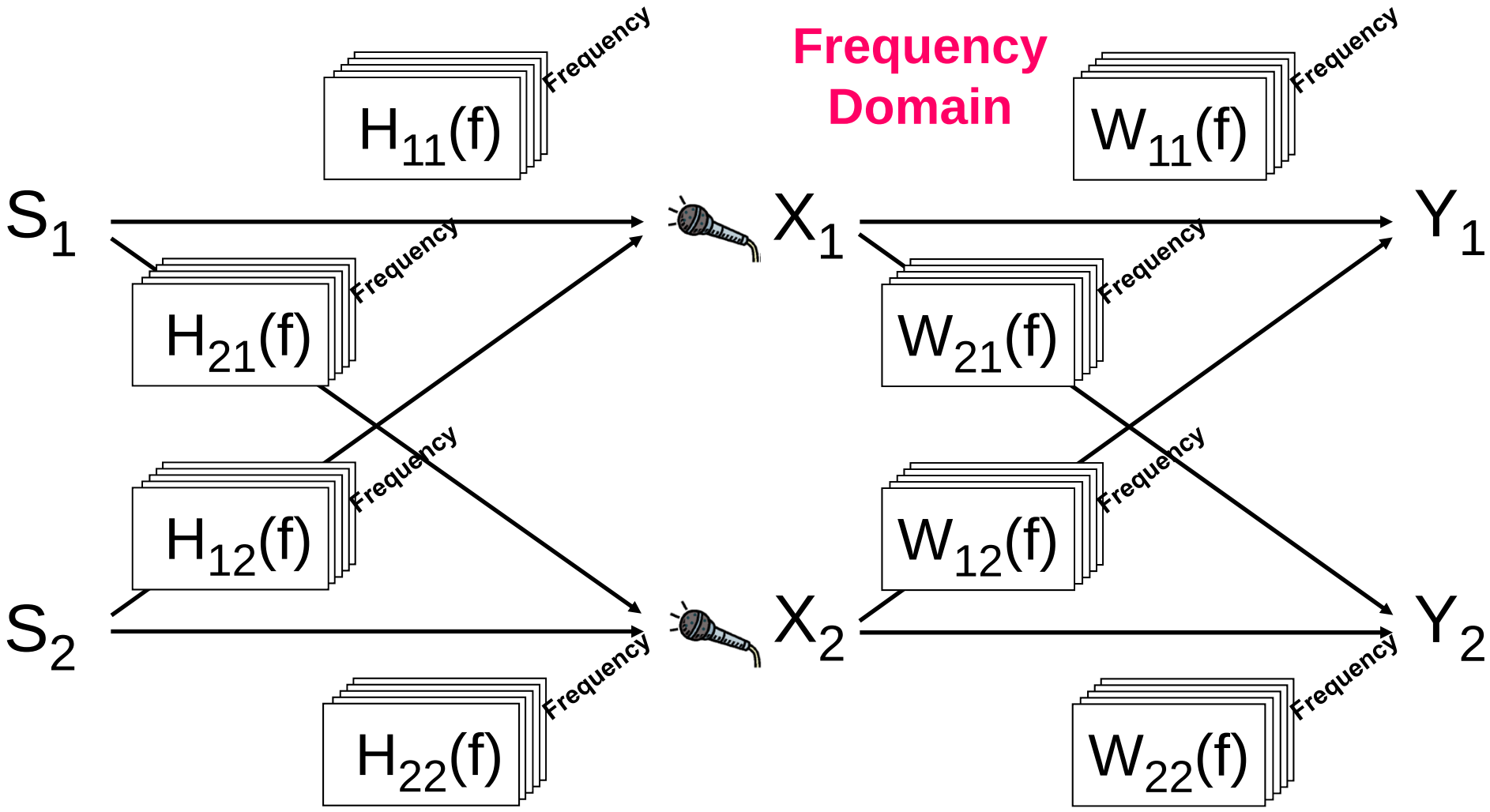
Frequency Domain BSS





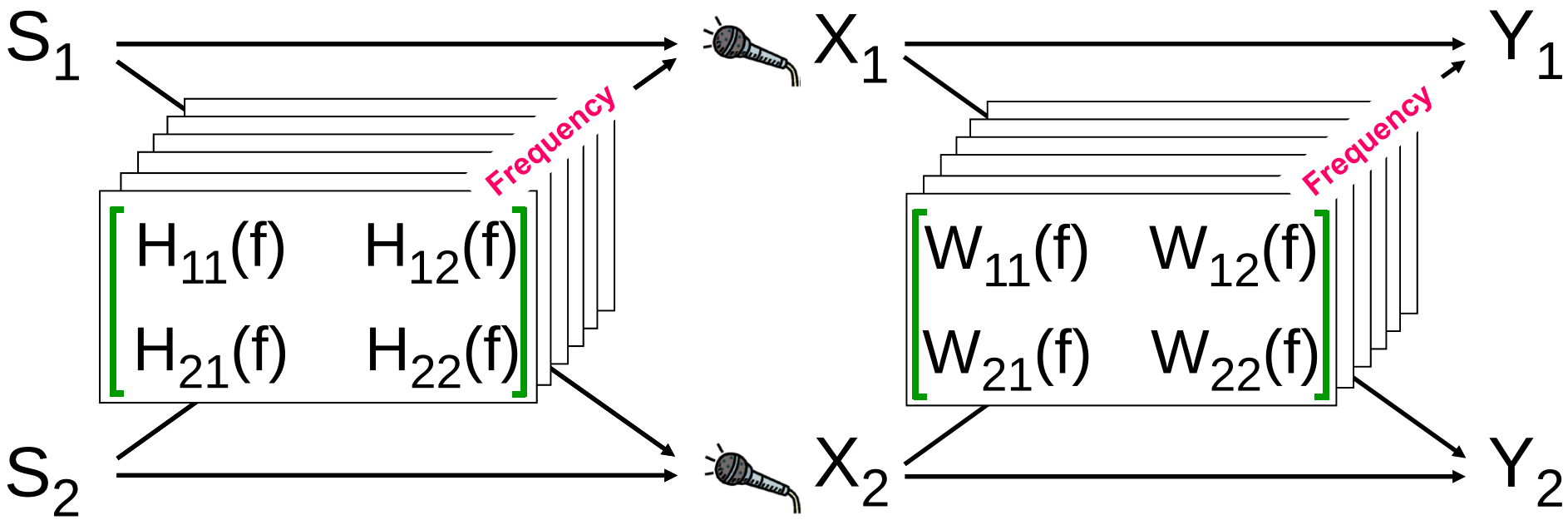
Apply instantaneous ICA approach to each frequency bin

Mixing Filters and Separation Filters

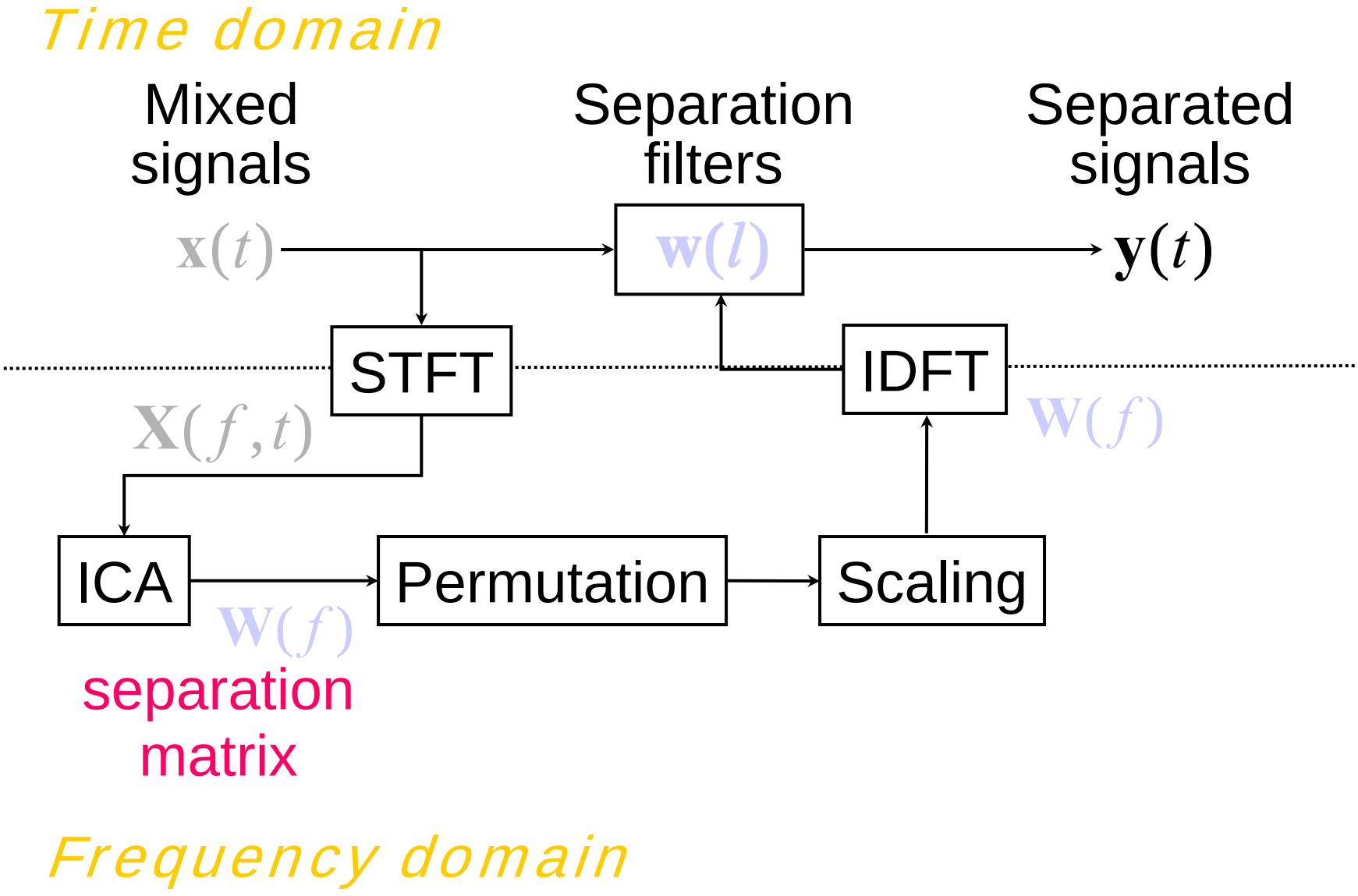


Mixing Filters and Separation Filters

Frequency Domain **in many Matrices**



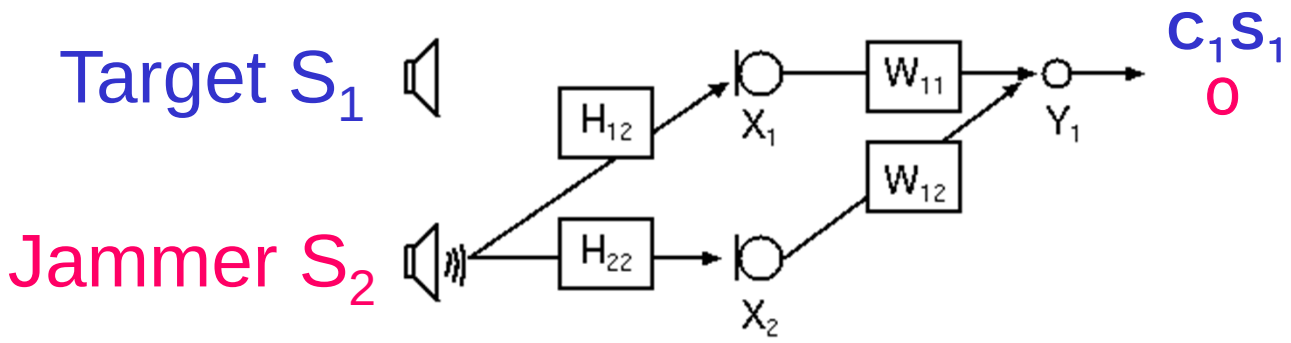
Flow of Frequency Domain BSS



Physical Interpretation of BSS

BSS = Two sets of ABF

Adaptive Beamformer (ABF)



- Assumptions

  Direction and absence period of a target is known

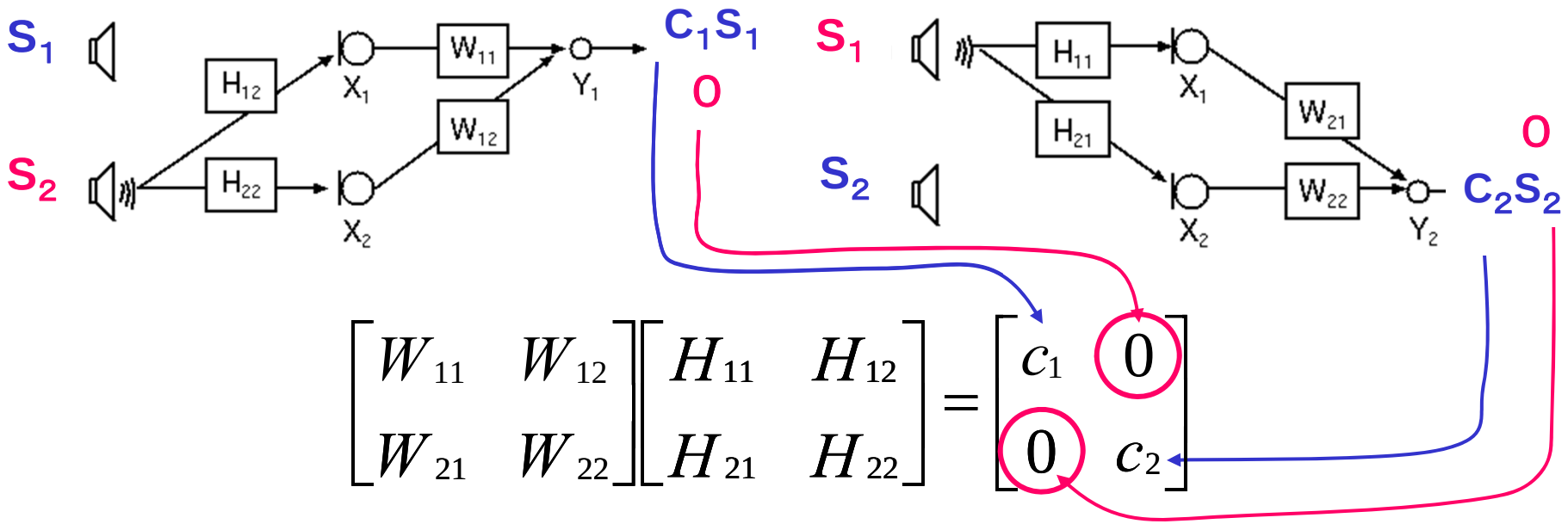
- Strategy

  Minimize the output when only a jammer is active but a target is not active

Two Sets of ABFs

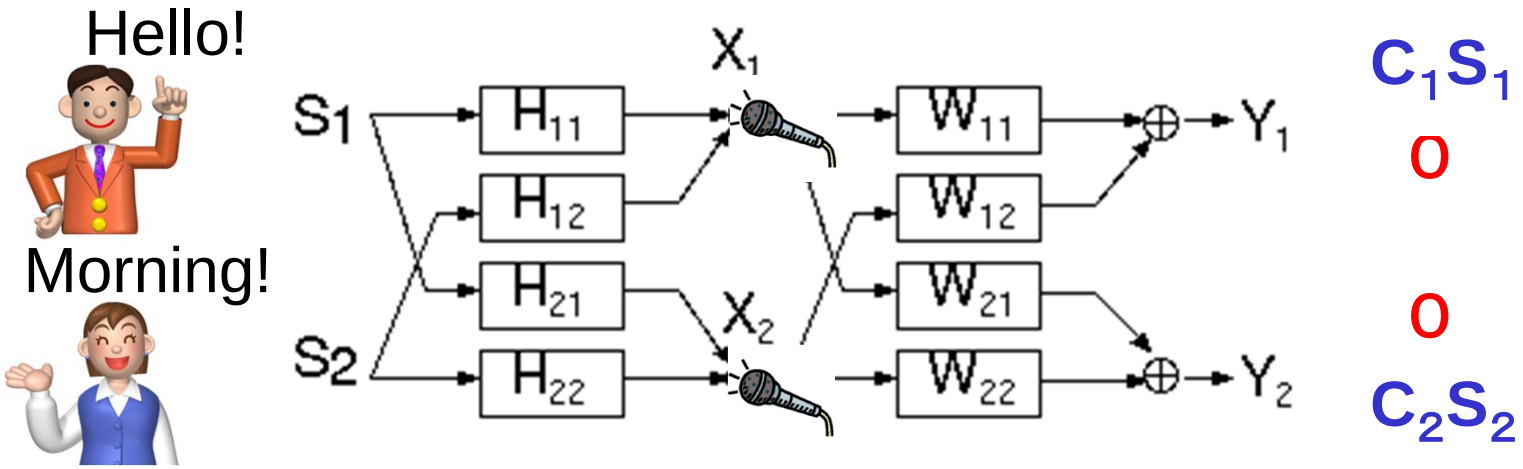
(a) ABF for target S_1 and jammer S_2

(b) ABF for target S_2 and jammer S_1



$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} = \begin{bmatrix} c_1 & 0 \\ 0 & c_2 \end{bmatrix}$$

Blind Source Separation (BSS)



- Assumptions

  Two sources are mutually independent

- Strategy

  Minimize the SOS or HOS of the outputs

Diagonalization of $\mathbf{R}_Y(f,t)$ in BSS

- The BSS strategy works to diagonalize $\mathbf{R}_Y(f,t)$

$$\mathbf{R}_{\mathbf{Y}}(f,t) = \mathbf{W}(f)\mathbf{R}_{\mathbf{X}}(f,t)\mathbf{W}^*(f)$$
$$= \mathbf{W}(f)\mathbf{H}(f)\mathbf{\Lambda}_S(f,t)\mathbf{H}^*(f)\mathbf{W}^*(f)$$
$$= E\begin{bmatrix} Y_1Y_1^* & Y_1Y_2^* \\ Y_2Y_1^* & Y_2Y_2^* \end{bmatrix}$$

- After convergence, the off-diagonal component is

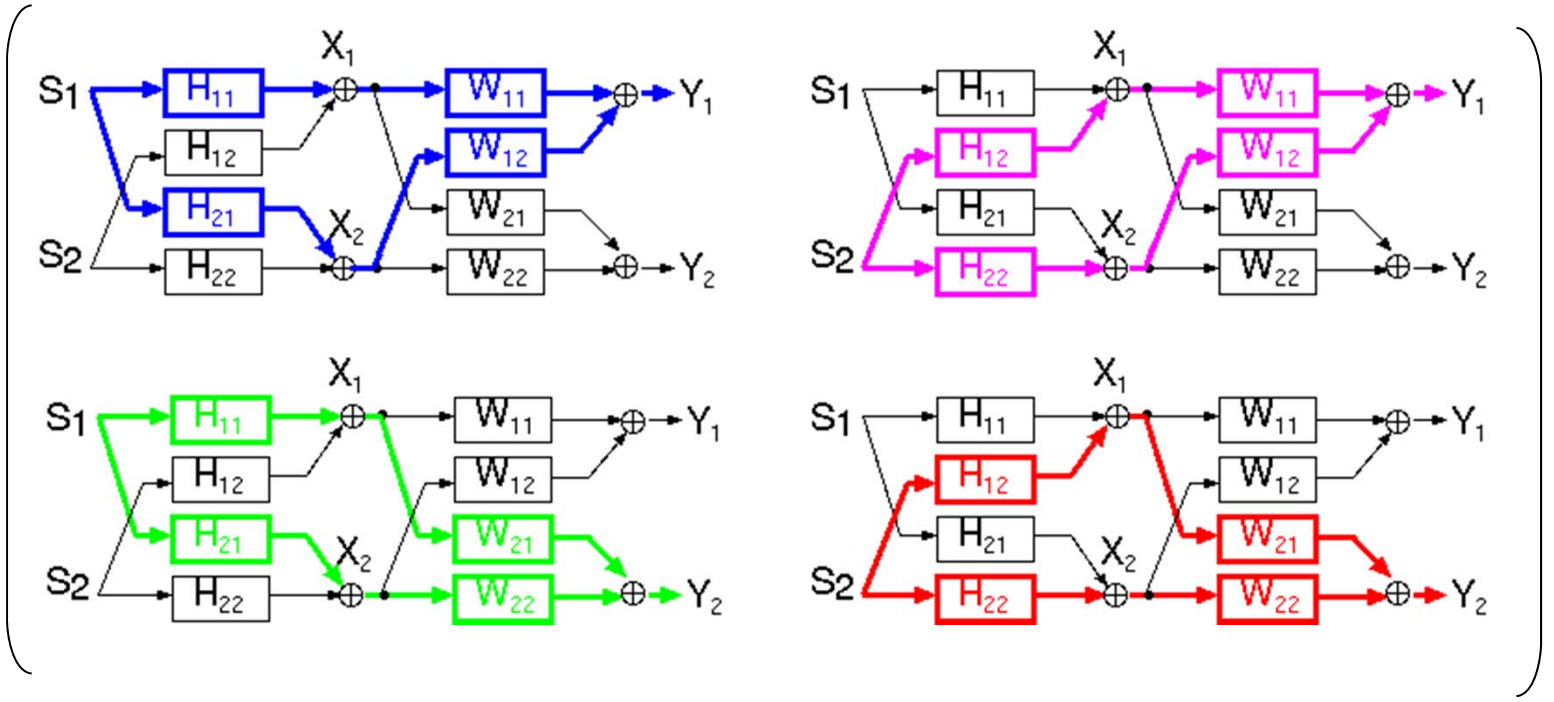
$$(E[Y_1Y_2^*])^2$$
$$= \{ad^*E[S_1S_2^*] + bc^*E[S_2S_1^*] + (ac^*E[S_1^2] + bd^*E[S_2^2])\}^2 \quad (\star)$$
$$= 0 \quad (E[S_1S_2^*] = 0,\ E[S_2S_1^*] = 0)$$

(If S_1 and S_2 are ideally independent)

where

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}$$

=



BSS Solutions

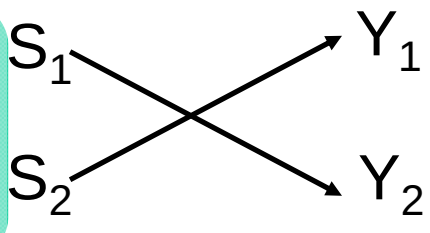
CASE 1: a=c_1, c=0, b=0, d=c_2

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}\begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}=\begin{bmatrix} c_1 & 0 \\ 0 & c_2 \end{bmatrix}$$

Same as ABF

CASE 2: a=0, c=c_1, b=c_2, d=0

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}\begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}=\begin{bmatrix} 0 & c_1 \\ c_2 & 0 \end{bmatrix}$$



(Permutation solution)

CASE 3: $\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}\begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}=\begin{bmatrix} 0 & 0 \\ c_1 & c_2 \end{bmatrix}$

CASE 4: $\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}\begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}=\begin{bmatrix} c_1 & c_2 \\ 0 & 0 \end{bmatrix}$

do not appear

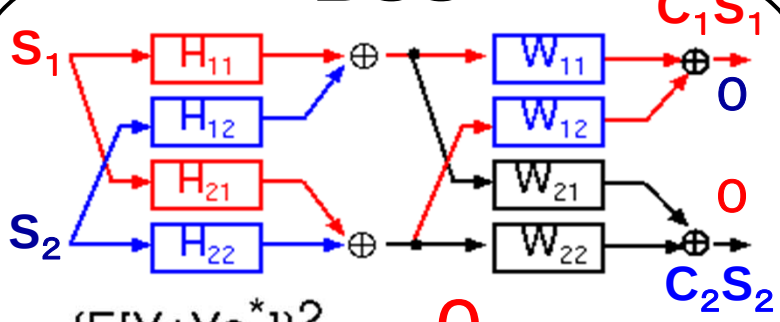
$\because$ we assume

$$|\mathbf{H}(f)|\neq 0$$

$$H_{ji}(f)\neq 0$$

Equivalence between BSS and ABF

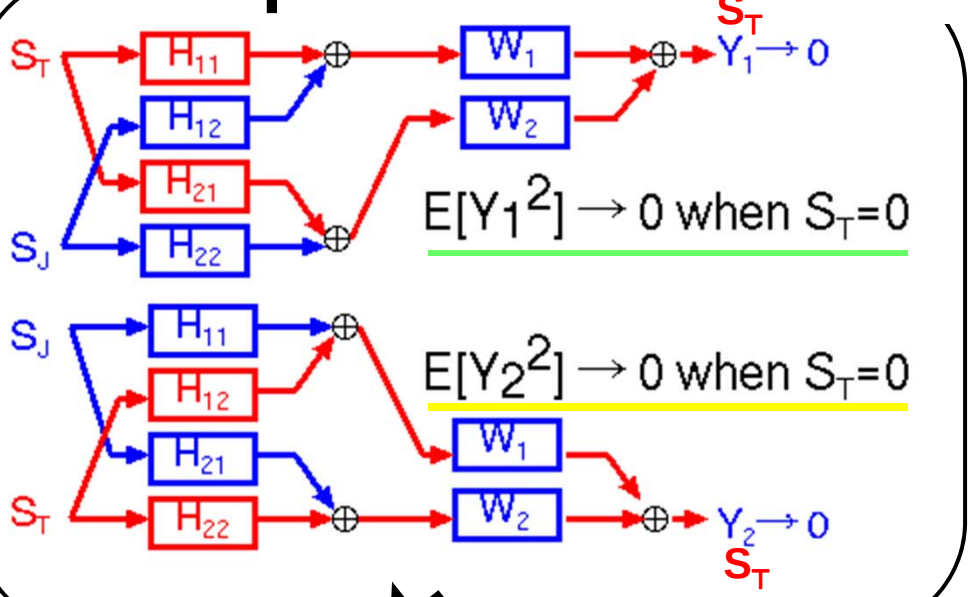
BSS



$$\{E[Y_1 Y_2^*]\}^2 = 0$$

$$= \{ad^* E[S_1 S_2^*] + bc^* E[S_2 S_1^*] + (ac^* E[S_1^2] + bd^* E[S_2^2])\}^2 \rightarrow 0$$

Adaptive Beamformers



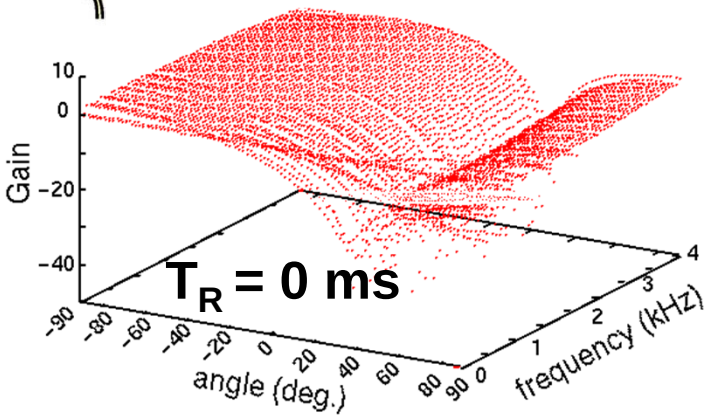
$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix} \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix} = \begin{pmatrix} C_1 & 0 \\ 0 & C_2 \end{pmatrix}$$

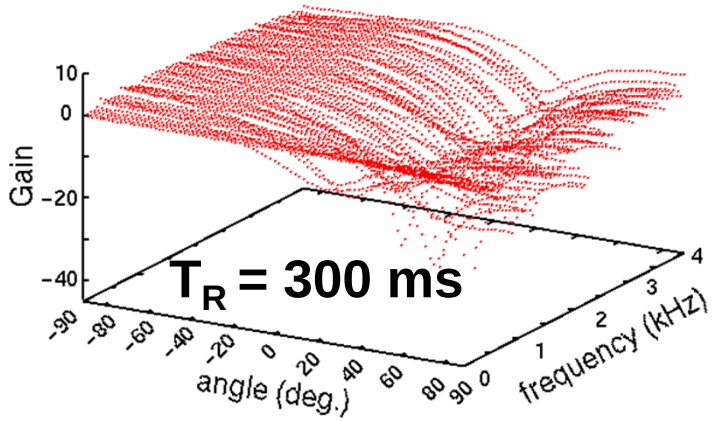
BSS and ABF are equivalent
if the independent assumption ideally holds

Physical Understanding of BSS

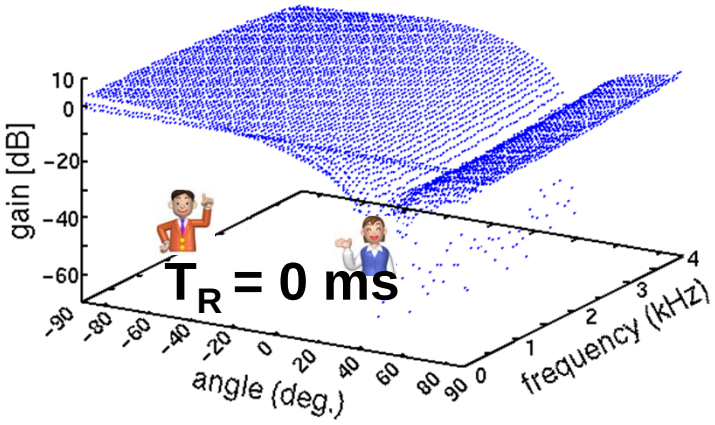
BSS, ABF ⇒ Spatial notch to jammer direction

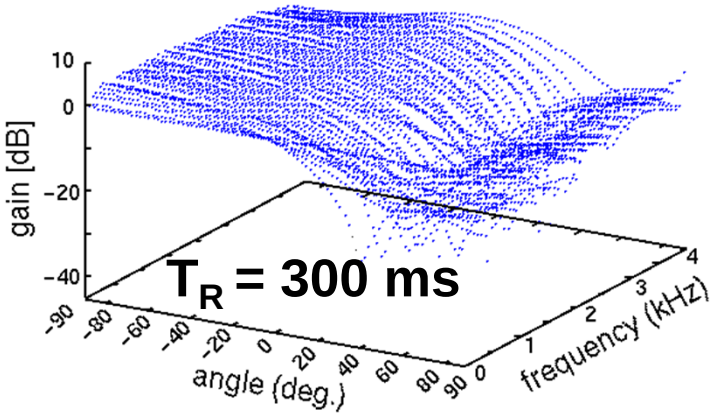
BSS





ABF

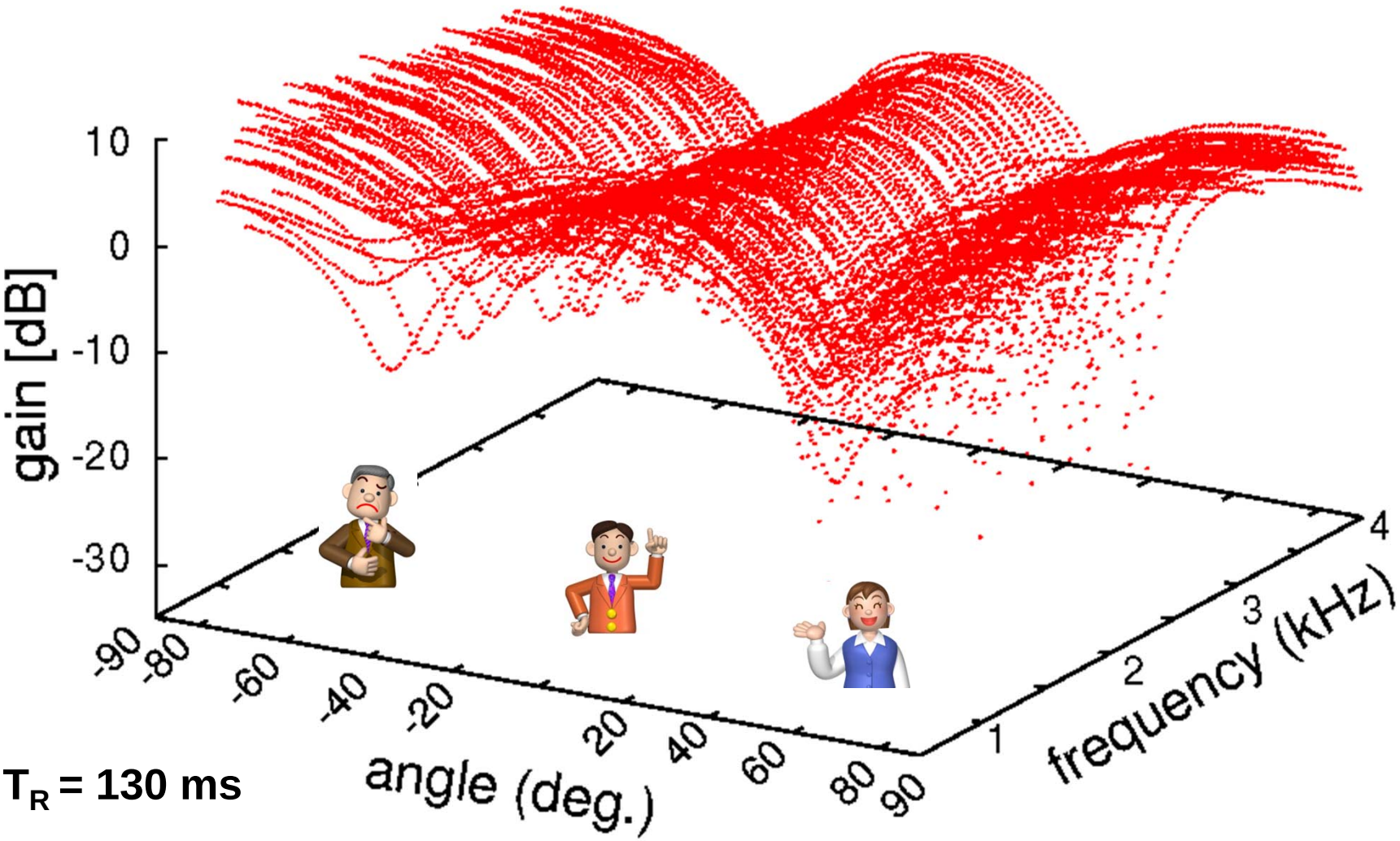




Directivity Patterns

BSS of Three Speeches

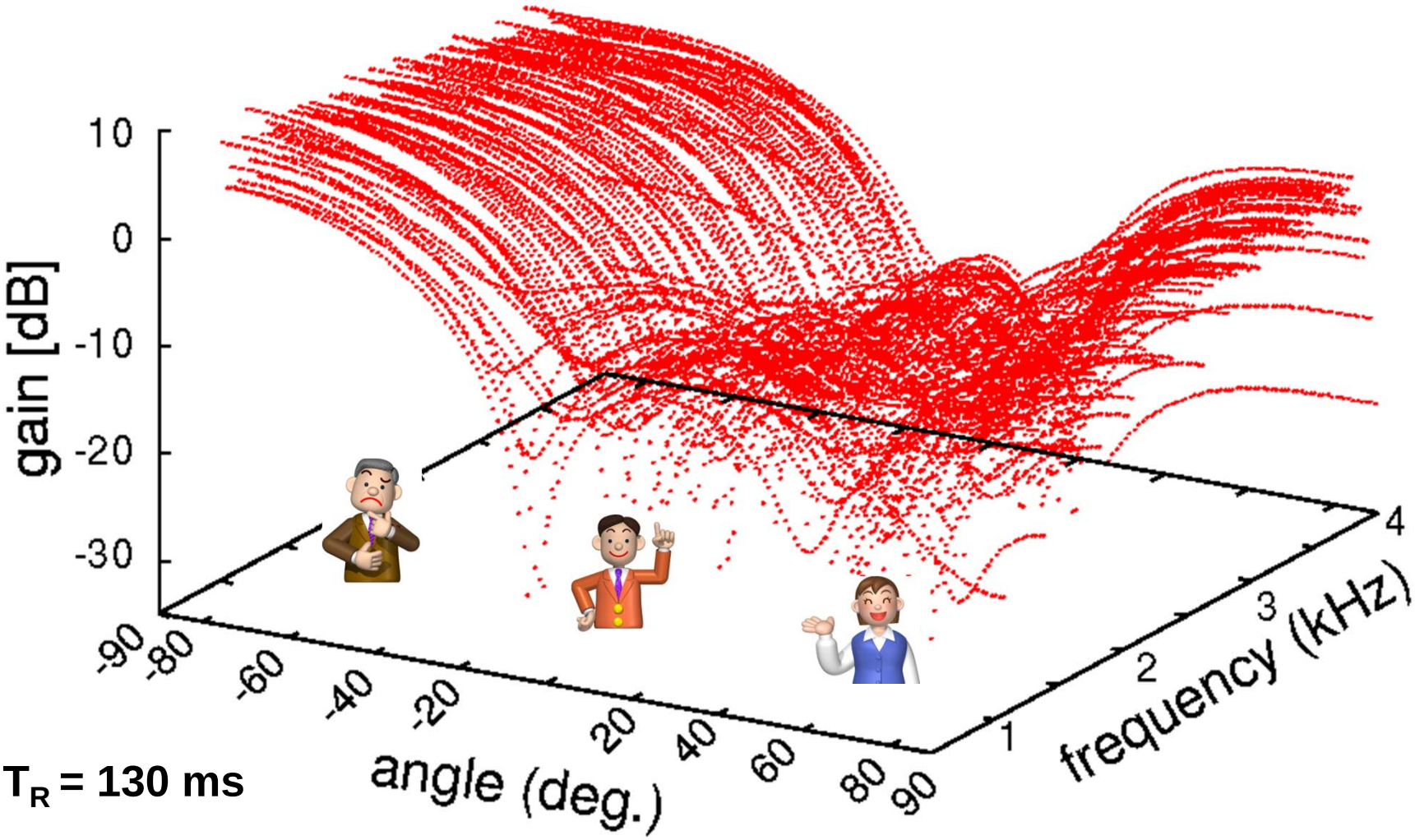
3 sources × 3 sensors BSS



T_R = 130 ms

Directivity Patterns

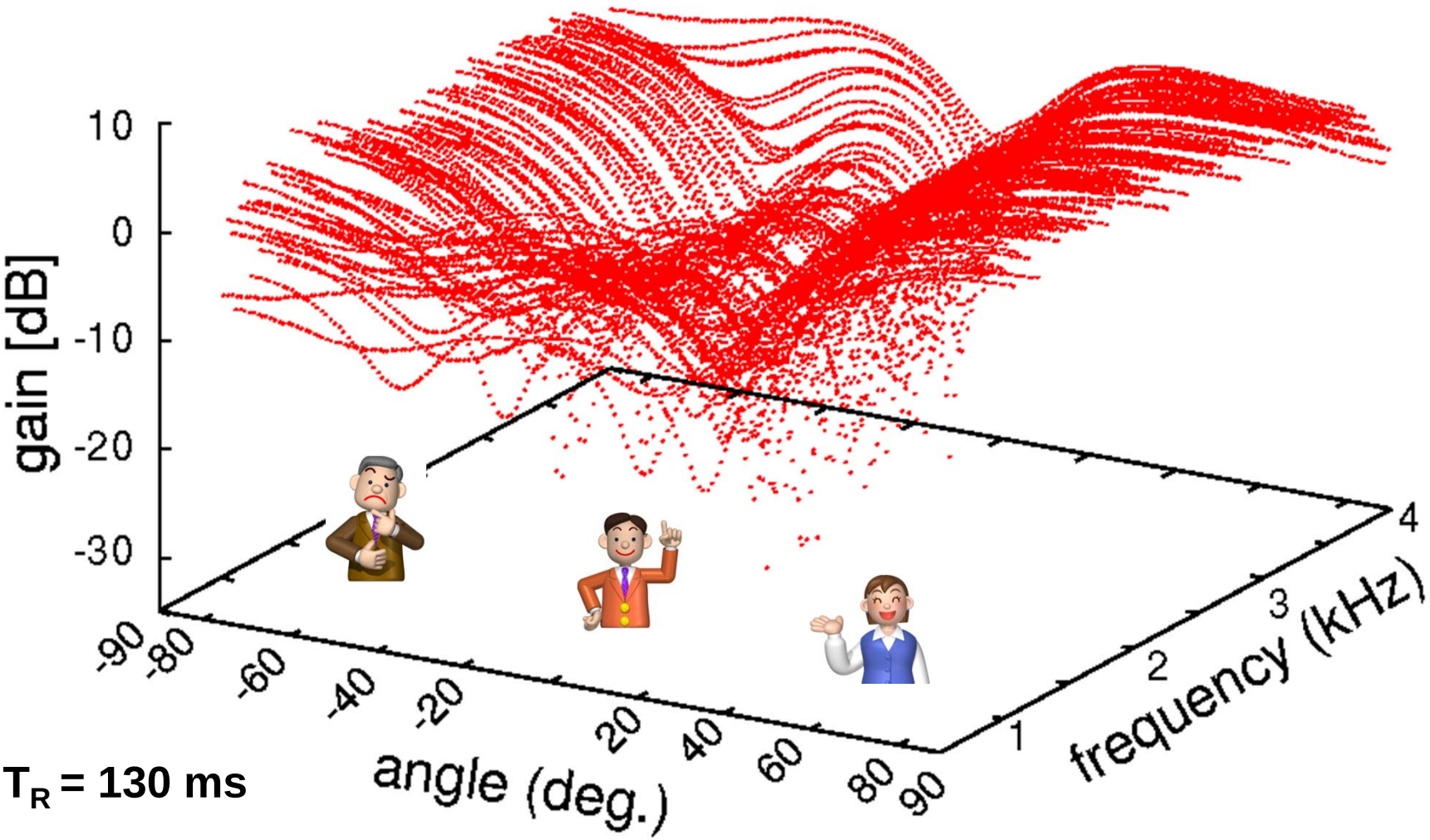
3 sources × 3 sensors BSS



T_R = 130 ms

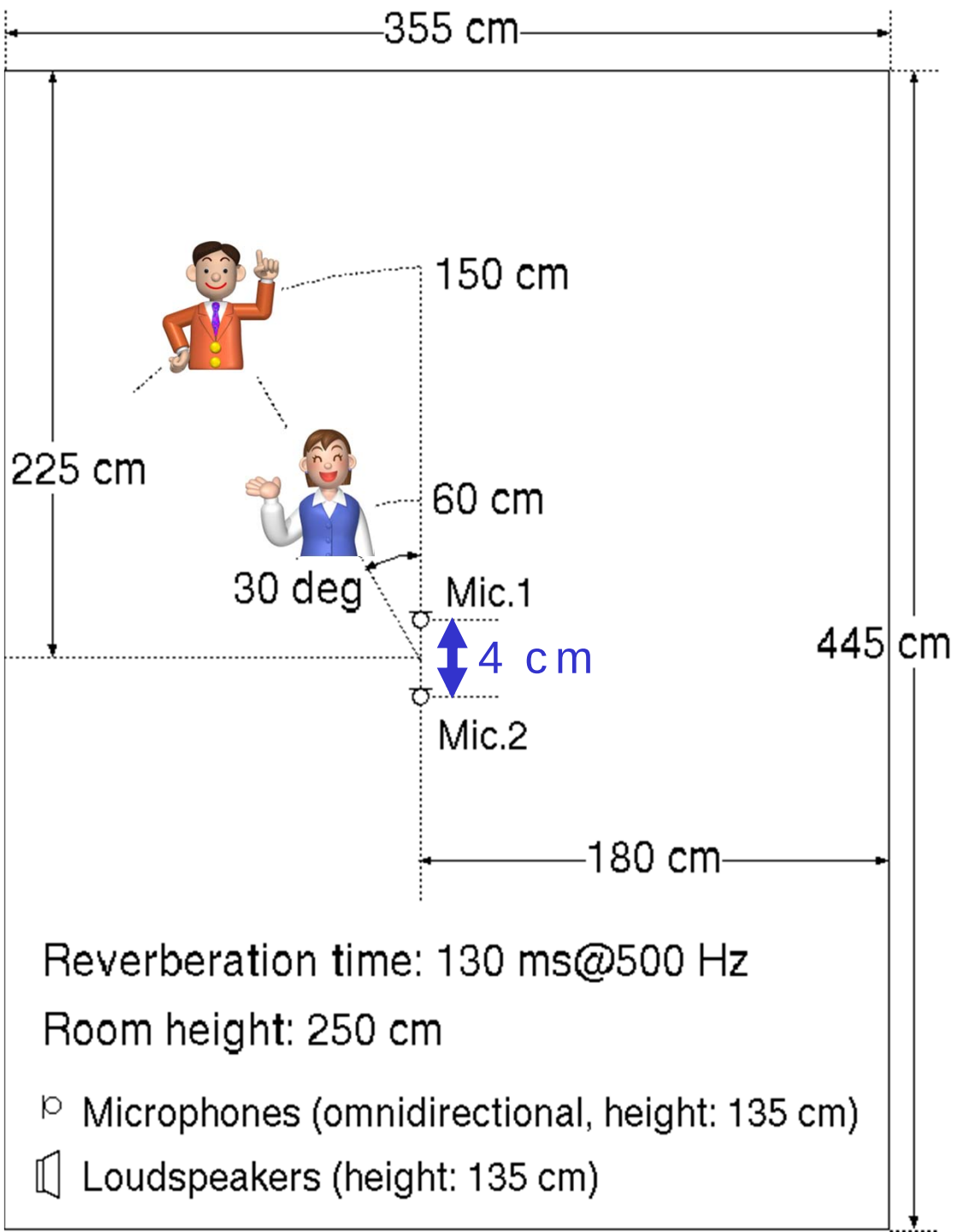
Directivity Patterns

3 sources × 3 sensors BSS



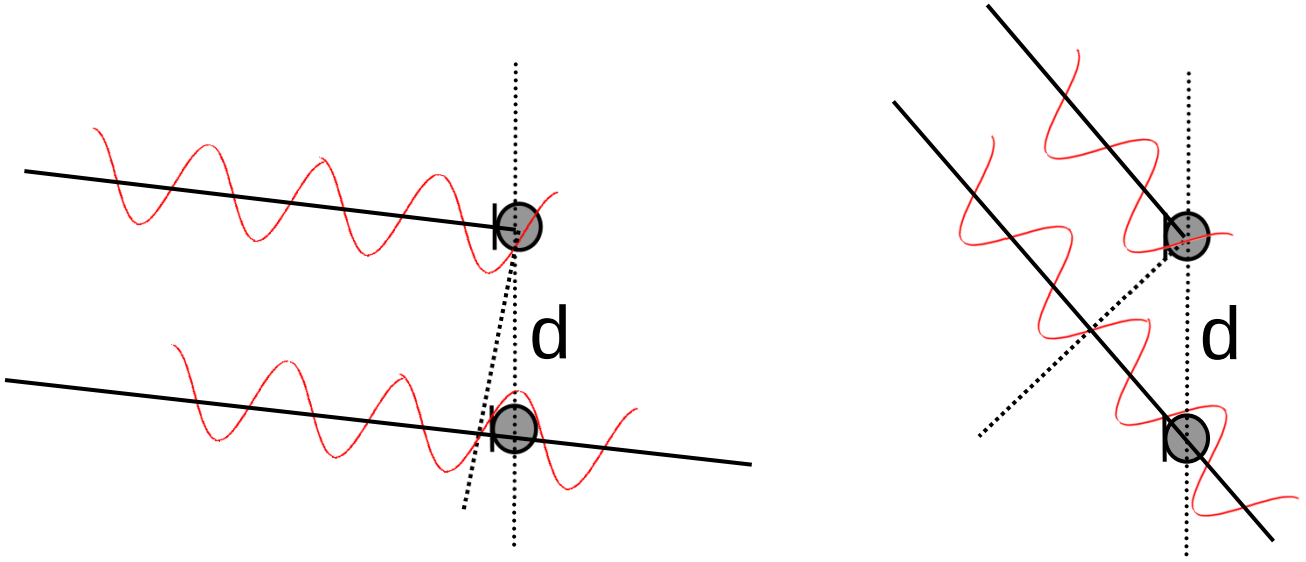
T_R = 130 ms

Directivity Patterns


355 cm
150 cm
225 cm
60 cm
30 deg
Mic.1
4 cm
445 cm
Mic.2
180 cm
Reverberation time: 130 ms@500 Hz
Room height: 250 cm
Microphones (omnidirectional, height: 135 cm)
Loudspeakers (height: 135 cm)

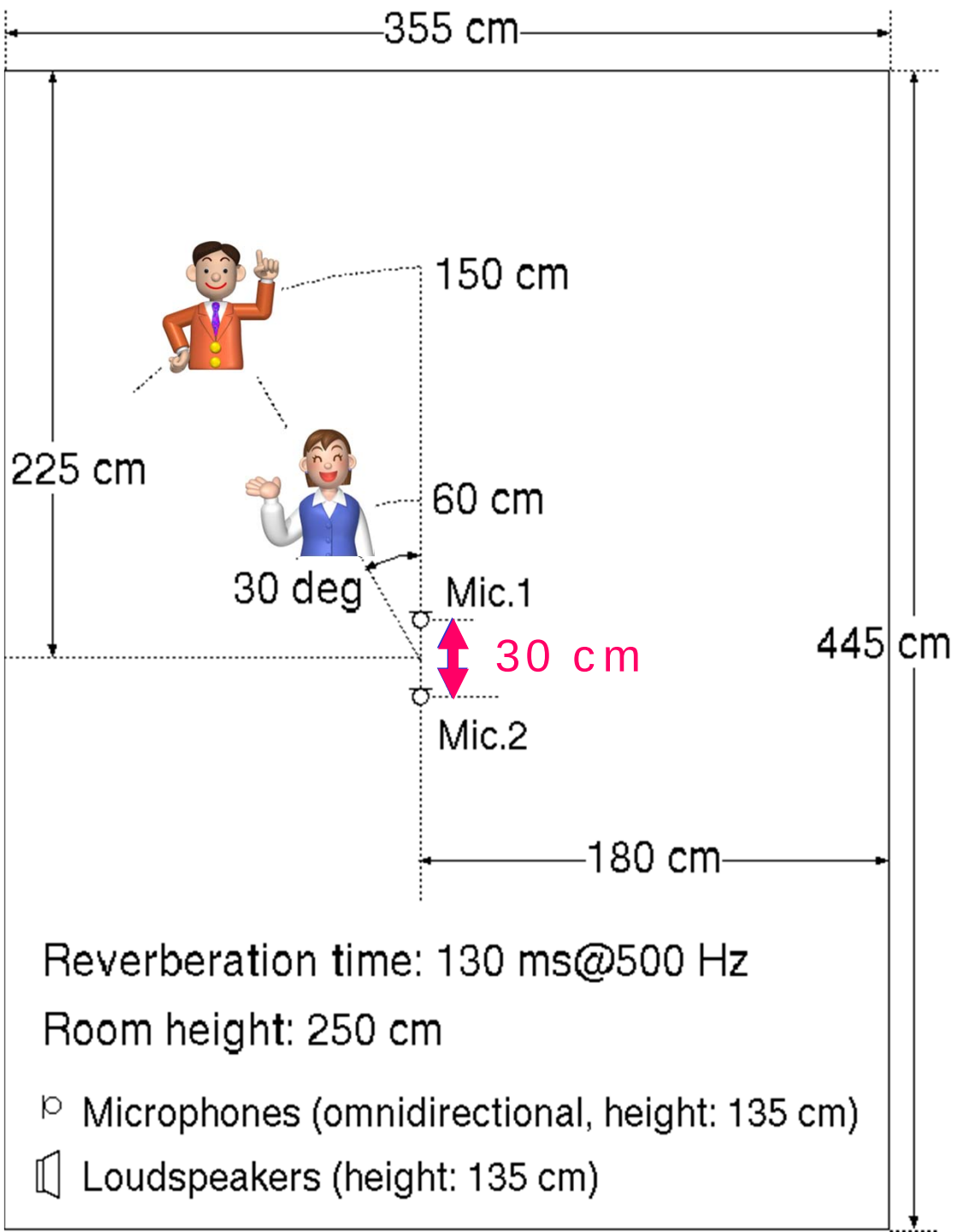
Spatial Aliasing

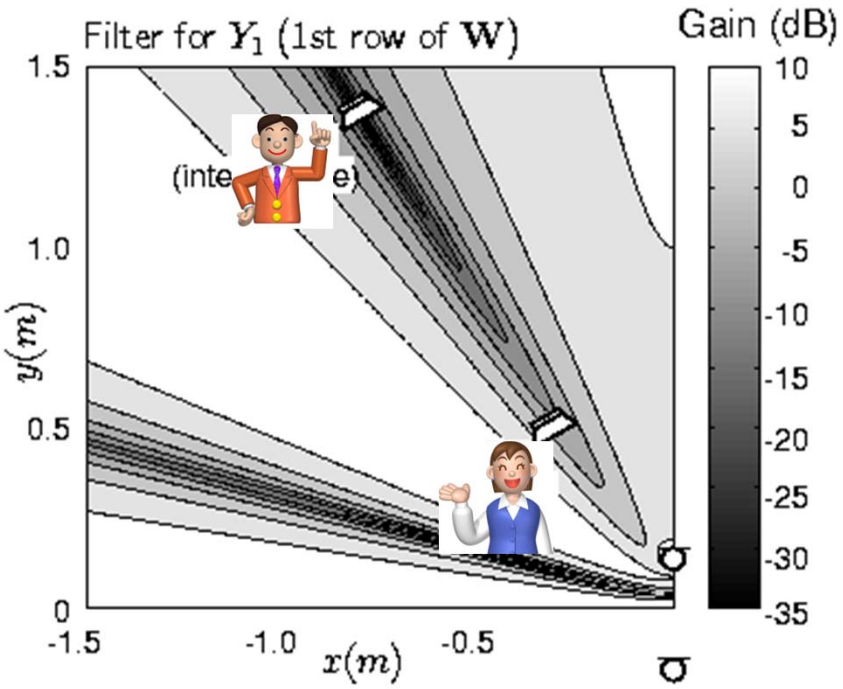
When microphone spacing *d* is too wide…

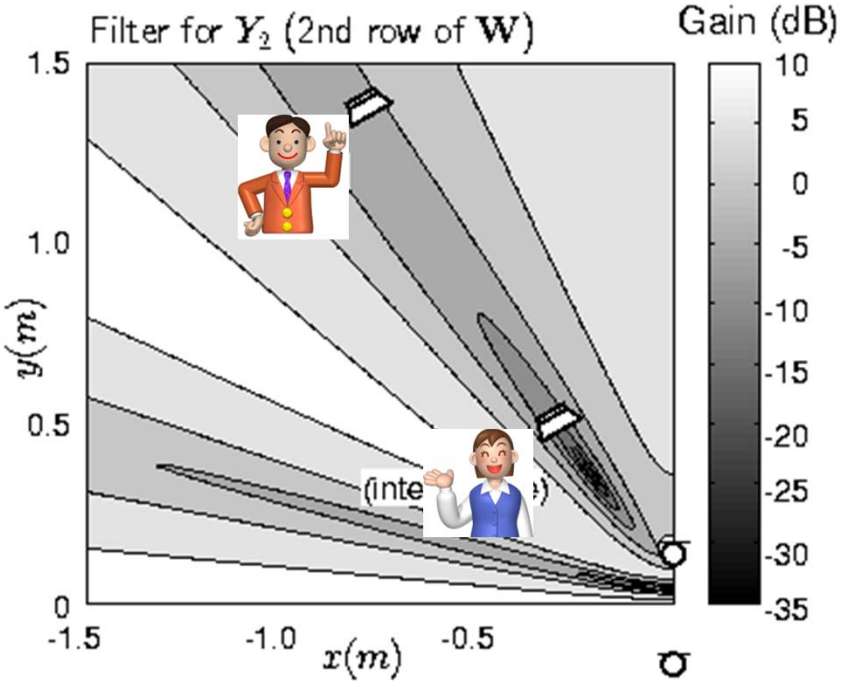


Spatial aliasing does not occer when $d < \frac{\lambda}{2}$

λ : wave length of the highest frequency

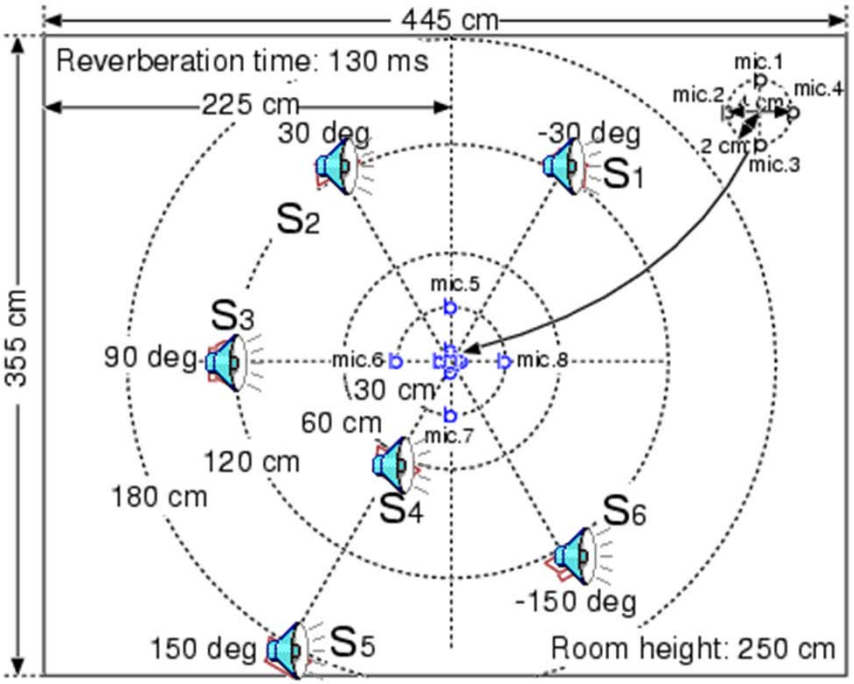

355 cm
150 cm
225 cm
60 cm
30 deg
Mic.1
30 cm
445 cm
Mic.2
180 cm
Reverberation time: 130 ms@500 Hz
Room height: 250 cm
Microphones (omnidirectional, height: 135 cm)
Loudspeakers (height: 135 cm)


Filter for Y_1 (1st row of $\mathbf{W}$)
Gain (dB)
1.5
1.0
0.5
0
-1.5
-1.0
-0.5
$y(m)$
$x(m)$
10
5
0
-5
-10
-15
-20
-25
-30
-35

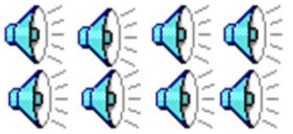

Filter for Y_2 (2nd row of $\mathbf{W}$)
Gain (dB)
1.5
1.0
0.5
0
-1.5
-1.0
-0.5
$y(m)$
$x(m)$
10
5
0
-5
-10
-15
-20
-25
-30
-35

6 sources × 8 sensors BSS

Experimental conditions



6 sources
8 microphones



Microphones (omnidirectional, height: 135 cm)

Loudspeakers (height: 135 cm)

Sampling rate	8 kHz
Data length	6 s
Frame length	2048 points (256 ms)
Frame shift	512 points (64 ms)
ICA algorithm	Infomax (complex valued)

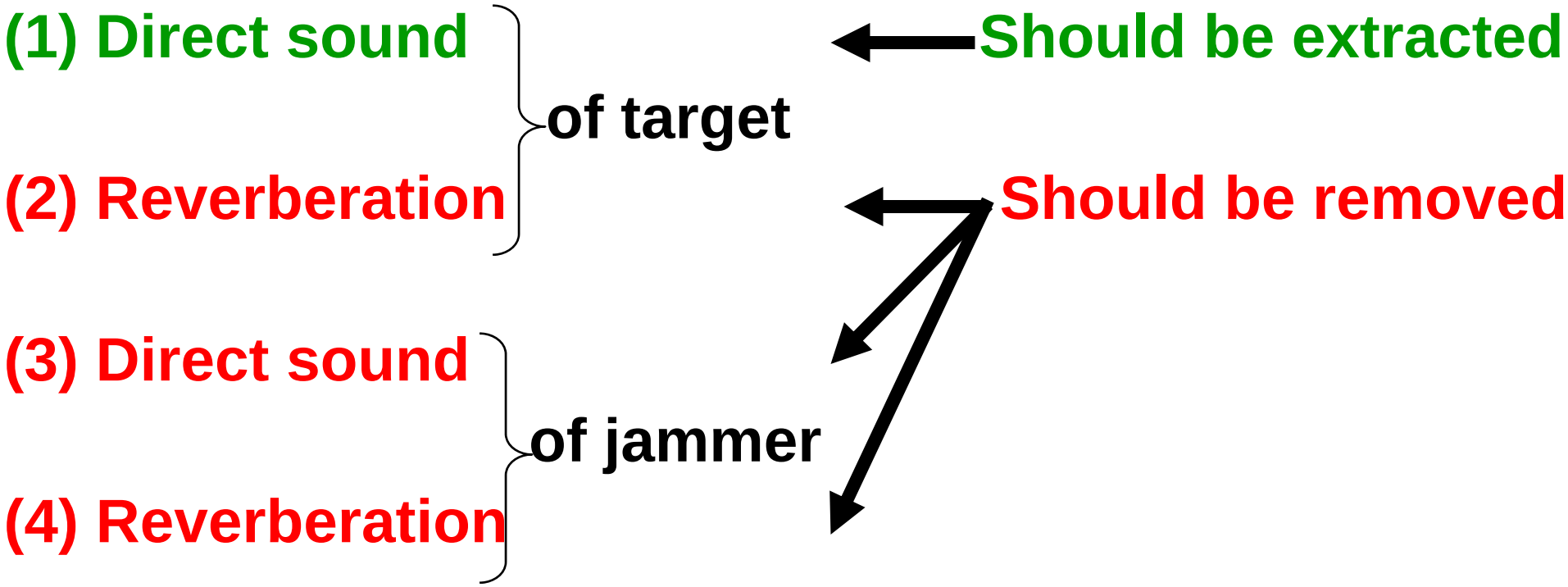
Experimental results

	S_1	S_2	S_3	S_4	S_5	S_6	ave.
Input SIR (dB)	-8.3	-6.8	-7.8	-7.7	-6.7	-5.2	-7.1
Output SIR (dB)	12.3	5.6	14.5	7.6	8.9	10.8	10.0

SIR improvement is 17.1 dB

Reverberation time: 130 ms
Computation time: about 1 min. for 6 sec. data
(Athlon 3200+, MATLAB)

What do we want?

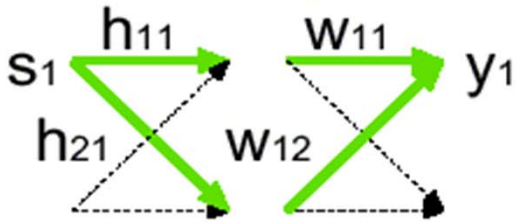


What is separated, and what remains?

BSS = Two sets of ABF

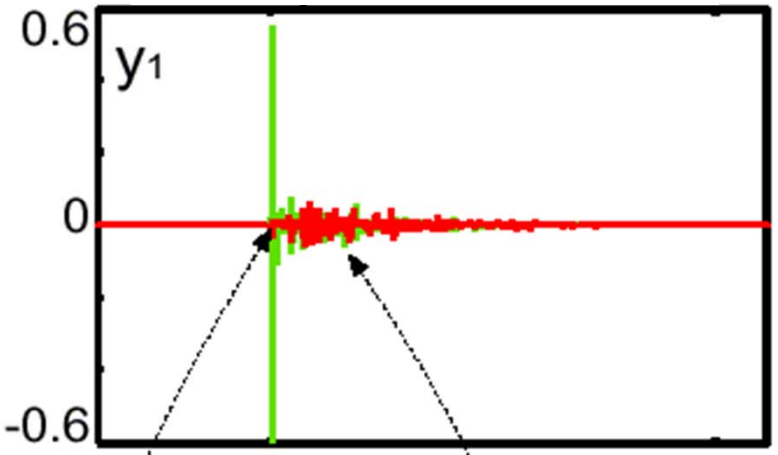
Comparison of NBF and BSS

target impulse response

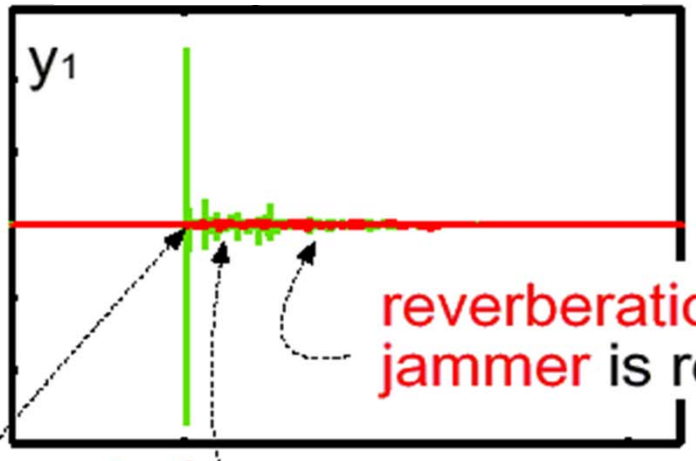


jammer impulse response

NBF output



BSS output



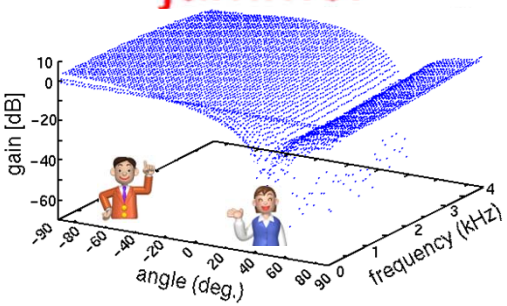
reverberation of jammer is reduce

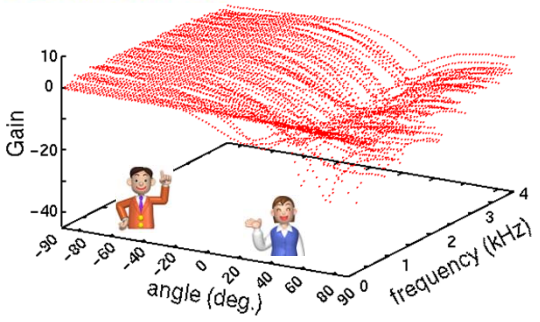
direct sound of jammer

direct sound of jammer is removed

reverberation of targe remains

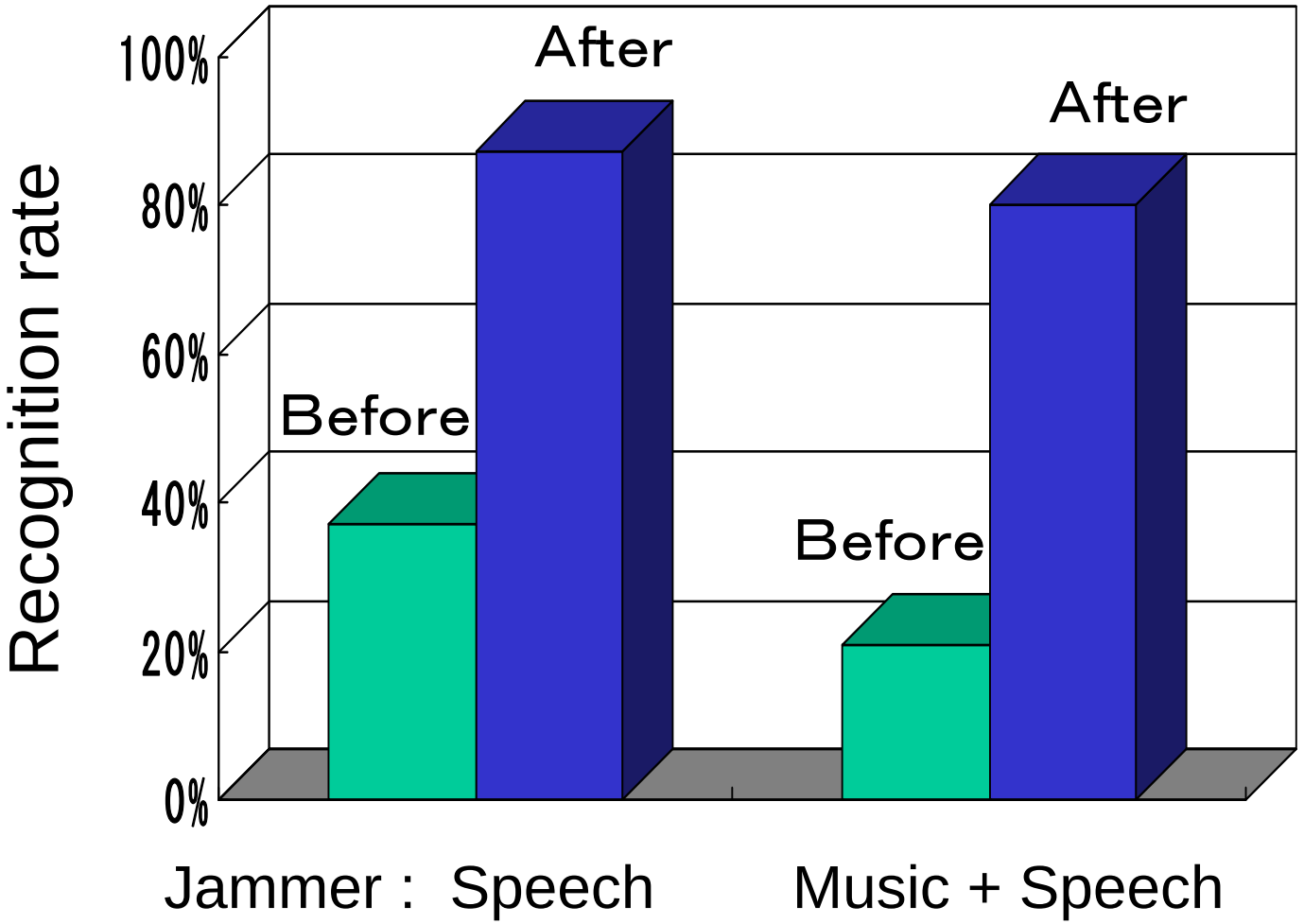
reverberation of jammer remains





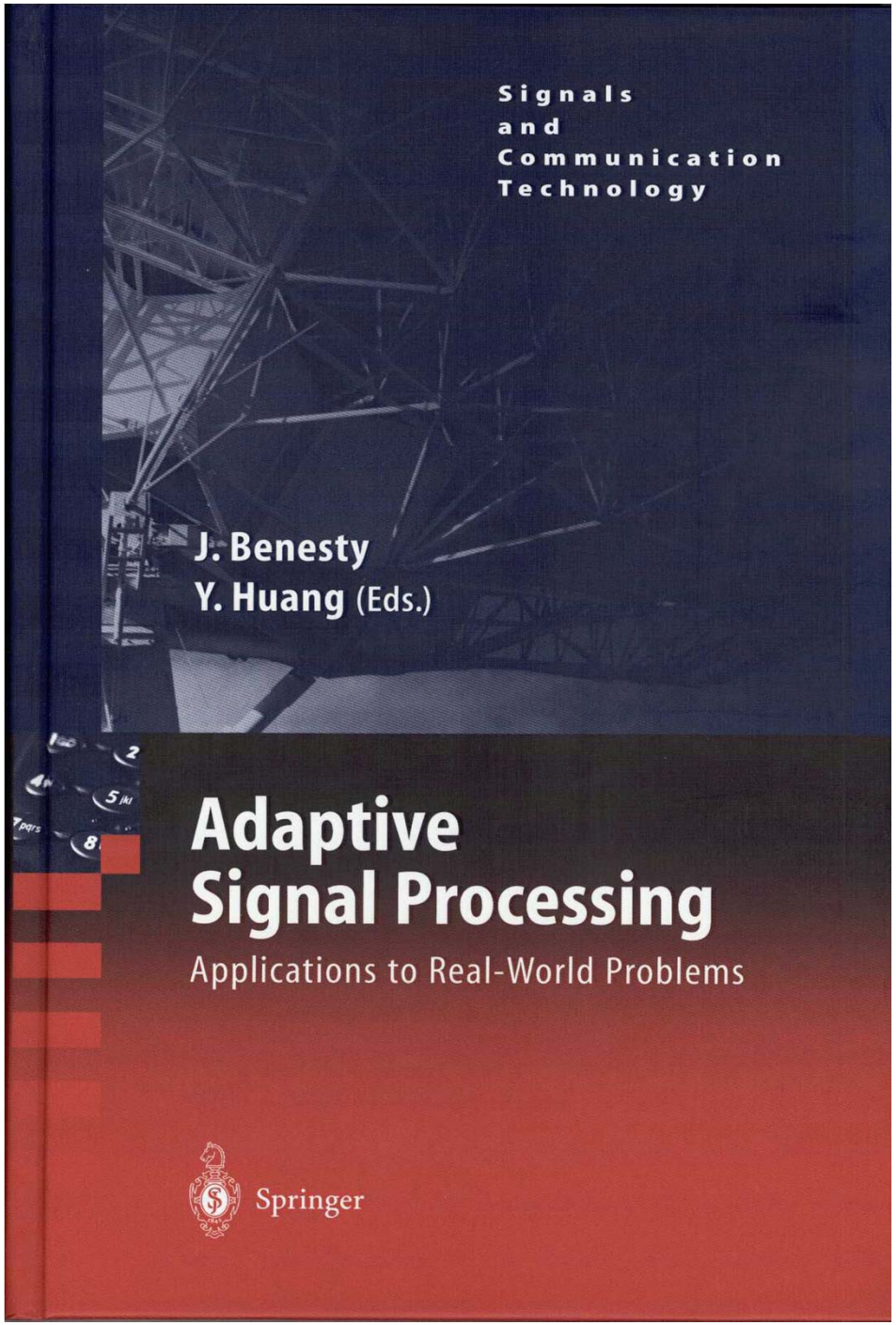
Effect on Speech Recognition

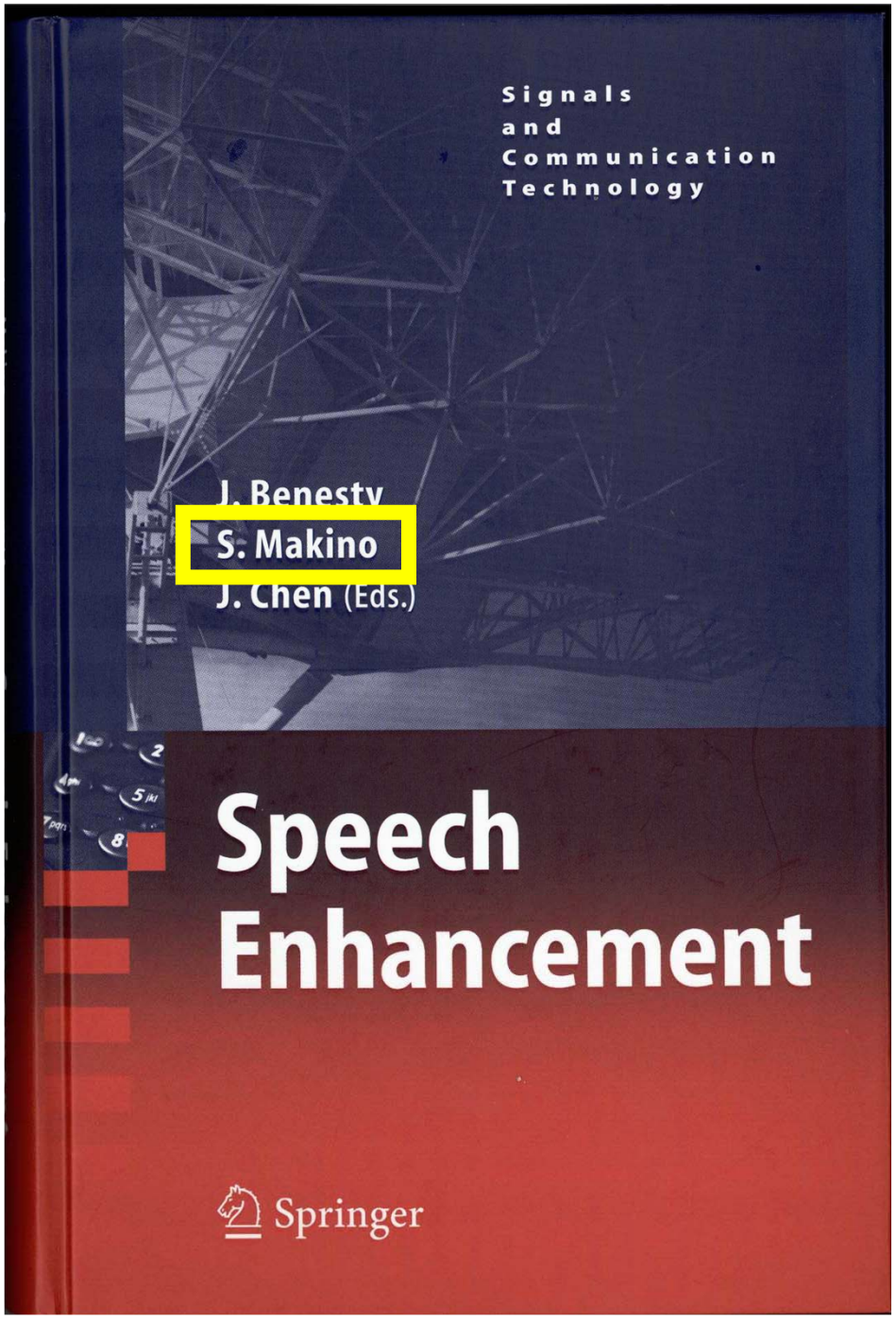
Effect on Speech Recognition

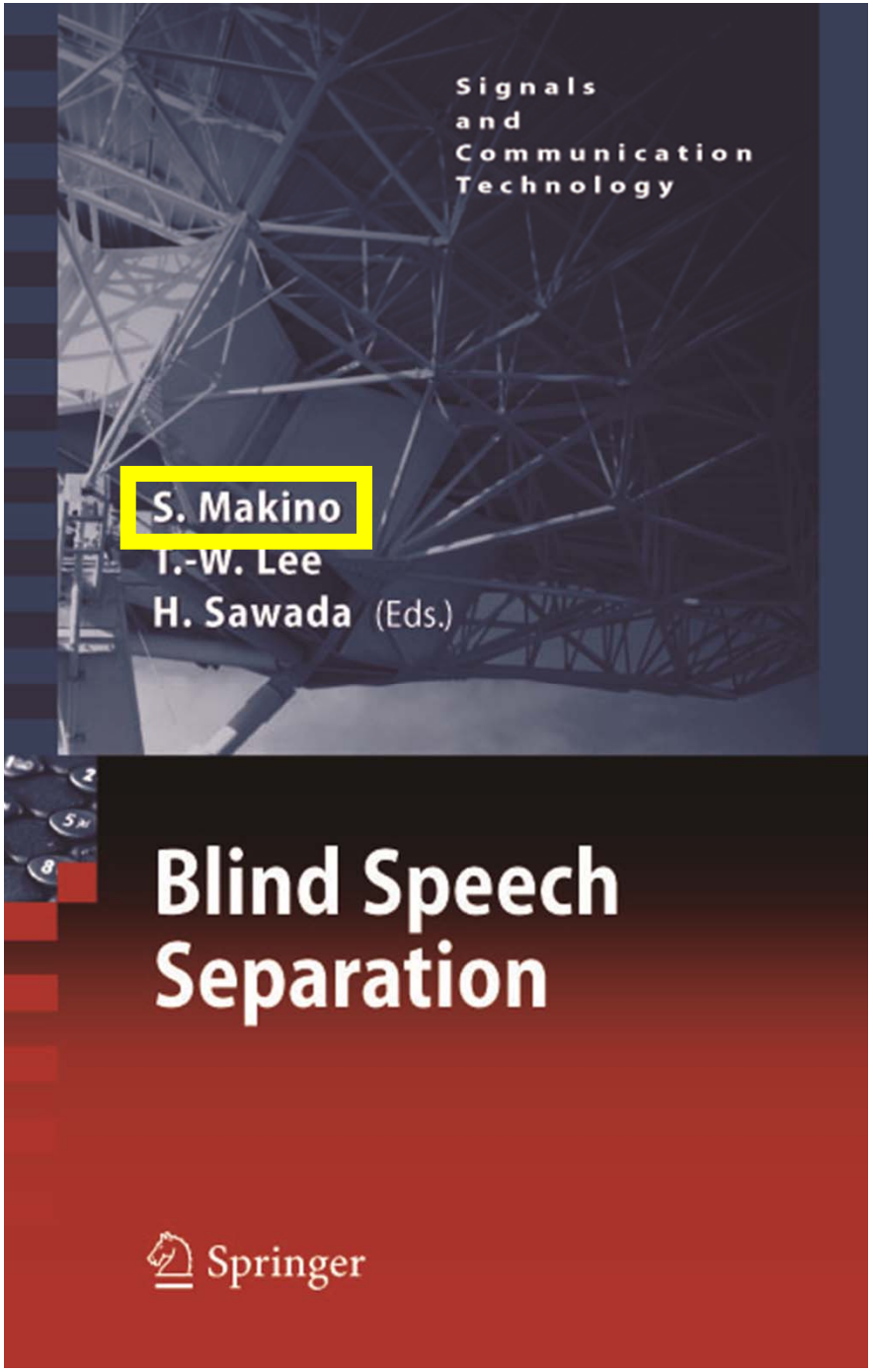


Blind Source Separation of Many Sounds

VIDEO


Signals
and
Communication
Technology
J. Benesty
Y. Huang (Eds.)
Adaptive
Signal Processing
Applications to Real-World Problems
Springer

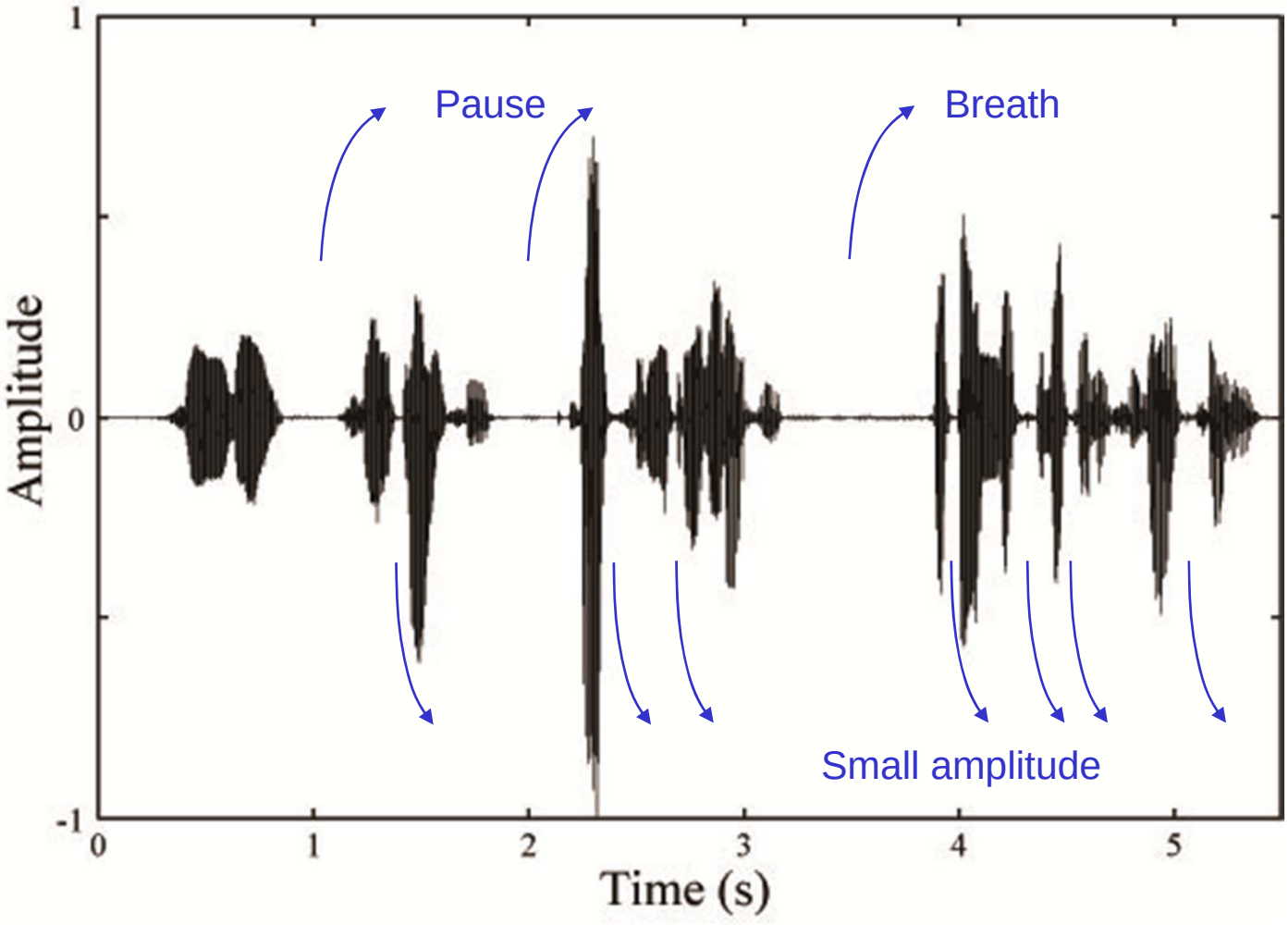

Signals
and
Communication
Technology
J. Benesty
S. Makino
J. Chen (Eds.)
Speech
Enhancement
Springer


Signals
and
Communication
Technology
S. Makino
T.-W. Lee
H. Sawada (Eds.)
Blind Speech
Separation
Springer

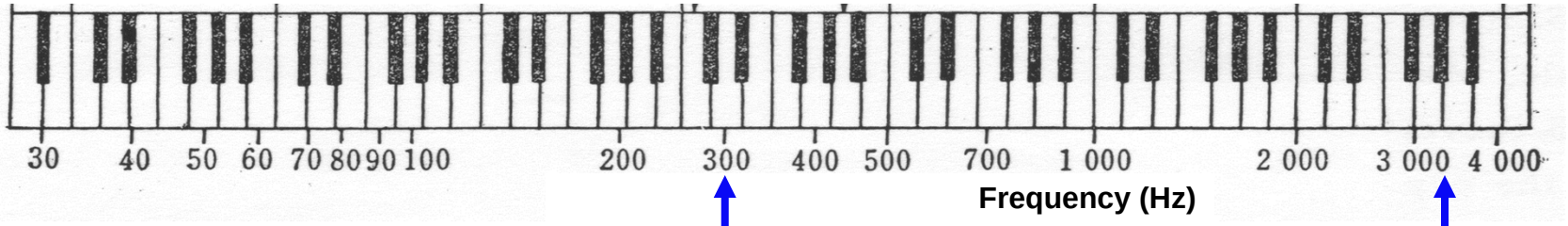
Outline

- Sparseness of speech sources

Speech in Time Domain



Speech in Frequency Domain



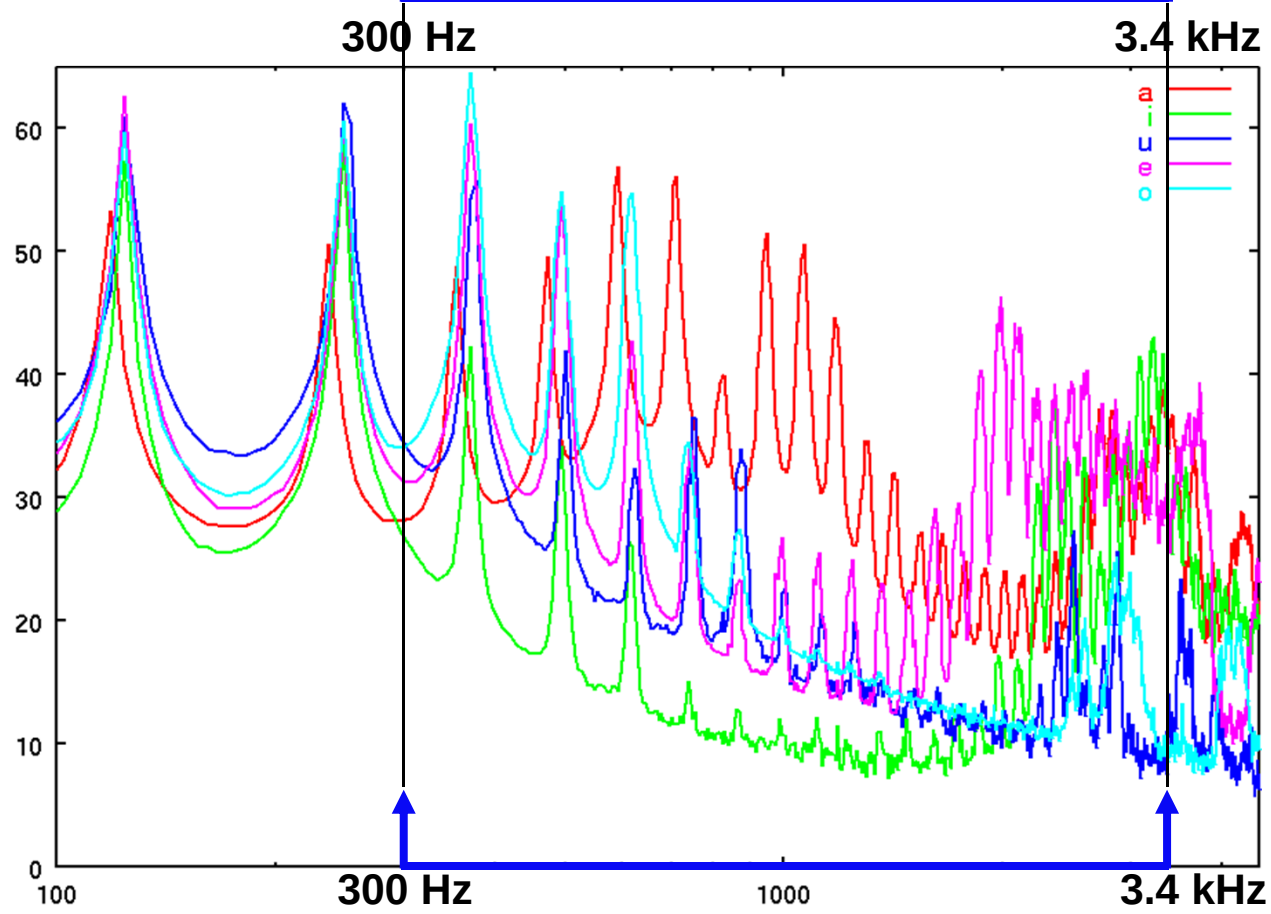
”a”

“i”

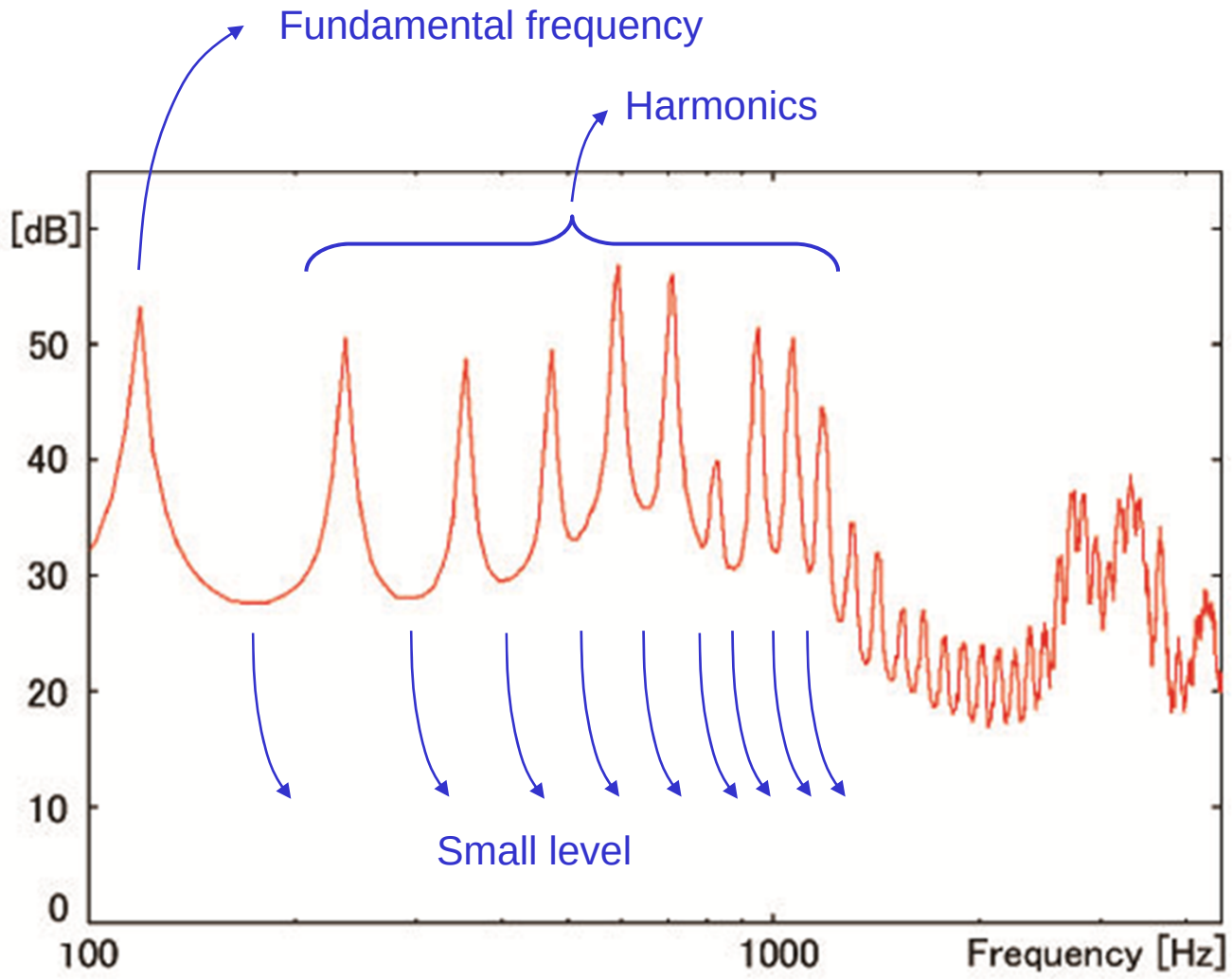
“u”

“e”

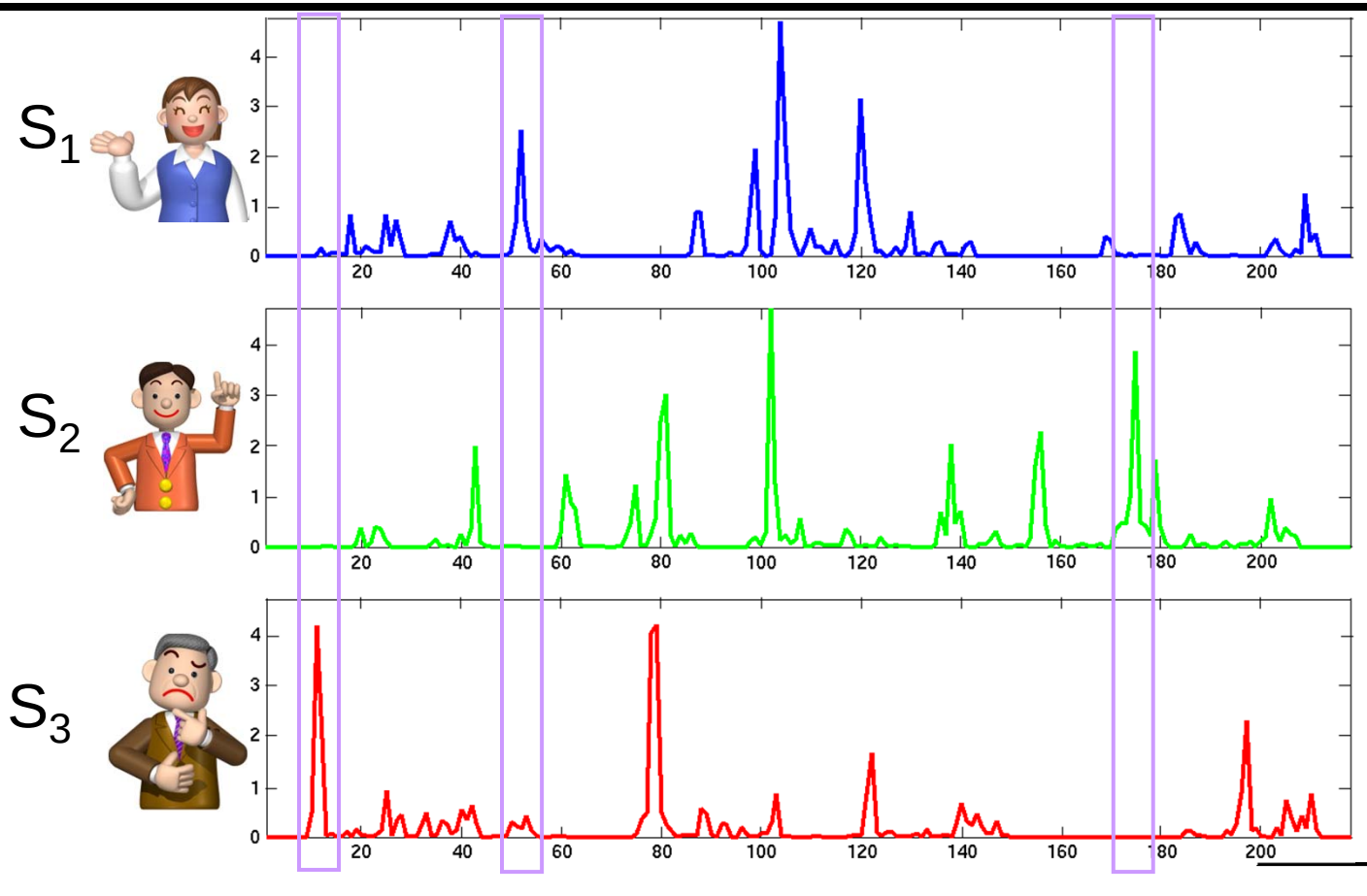
“o”



Speech in Frequency Domain



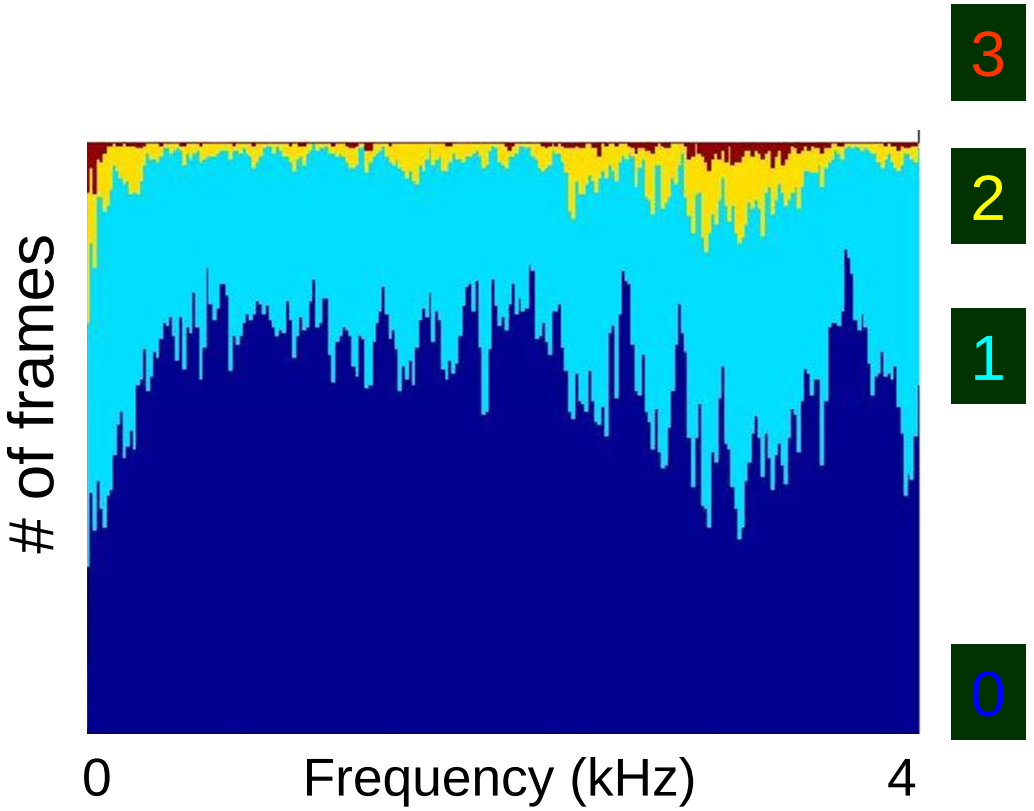
Sparseness



- Most of the samples are almost zero

⇒ ASSUMPTION Sources rarely overlap

Number of sources in each frame



More than two sources are rarely active simultaneously.

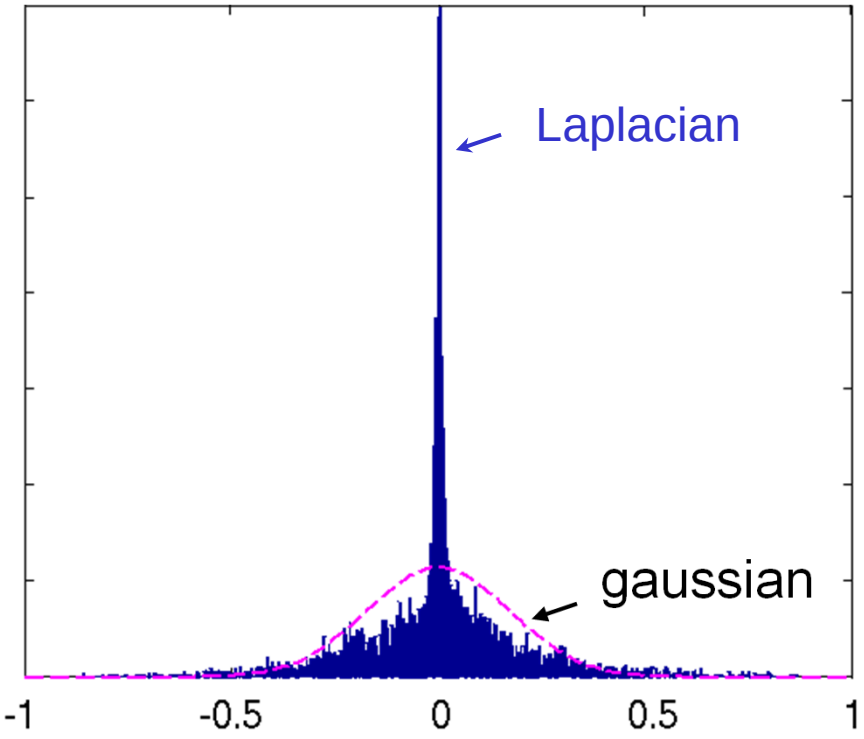
Anechoic, male-male-female, threshold = -20dB from the maximum

Sparseness

- Most of the samples of a sparse signal are almost zero.

- ASSUMPTION
  Sources overlap at rare intervals.

- Information on mixing process is not lost.

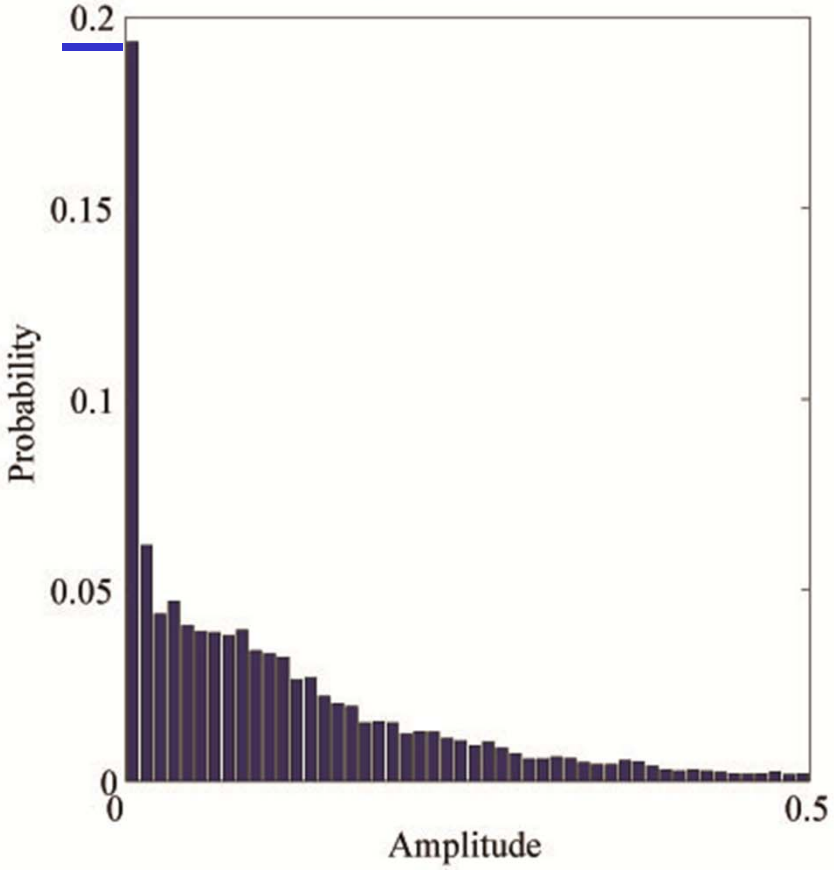


Example pdf of a sparse signal

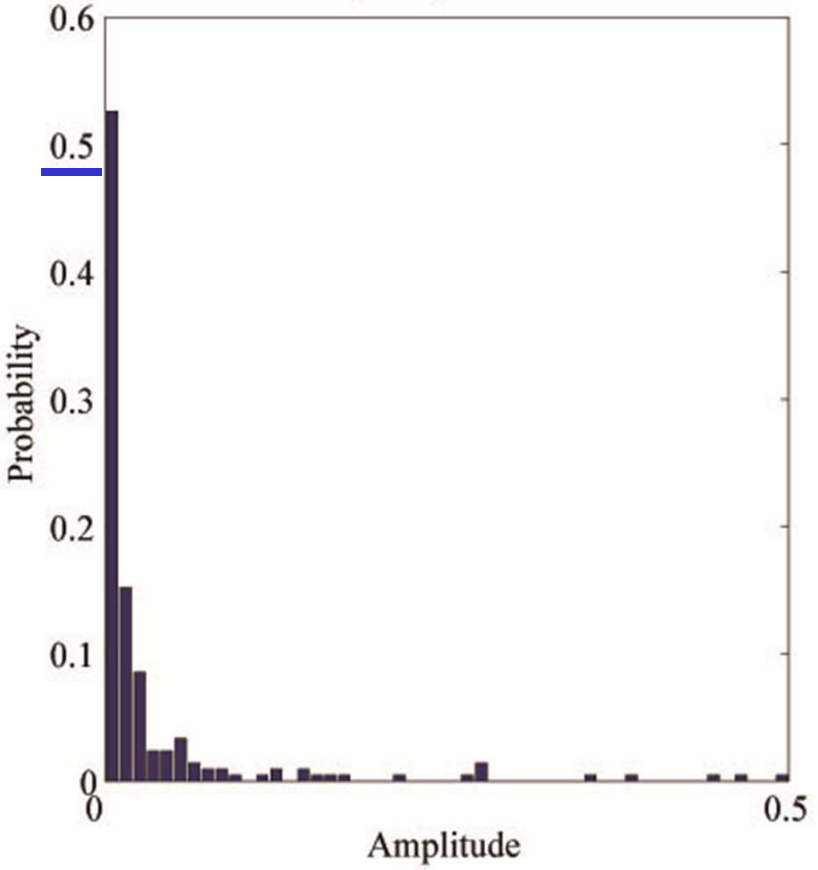
Sparseness in Time & Frequency Domains

Speech is sparser in time-frequency domain.

Time domain

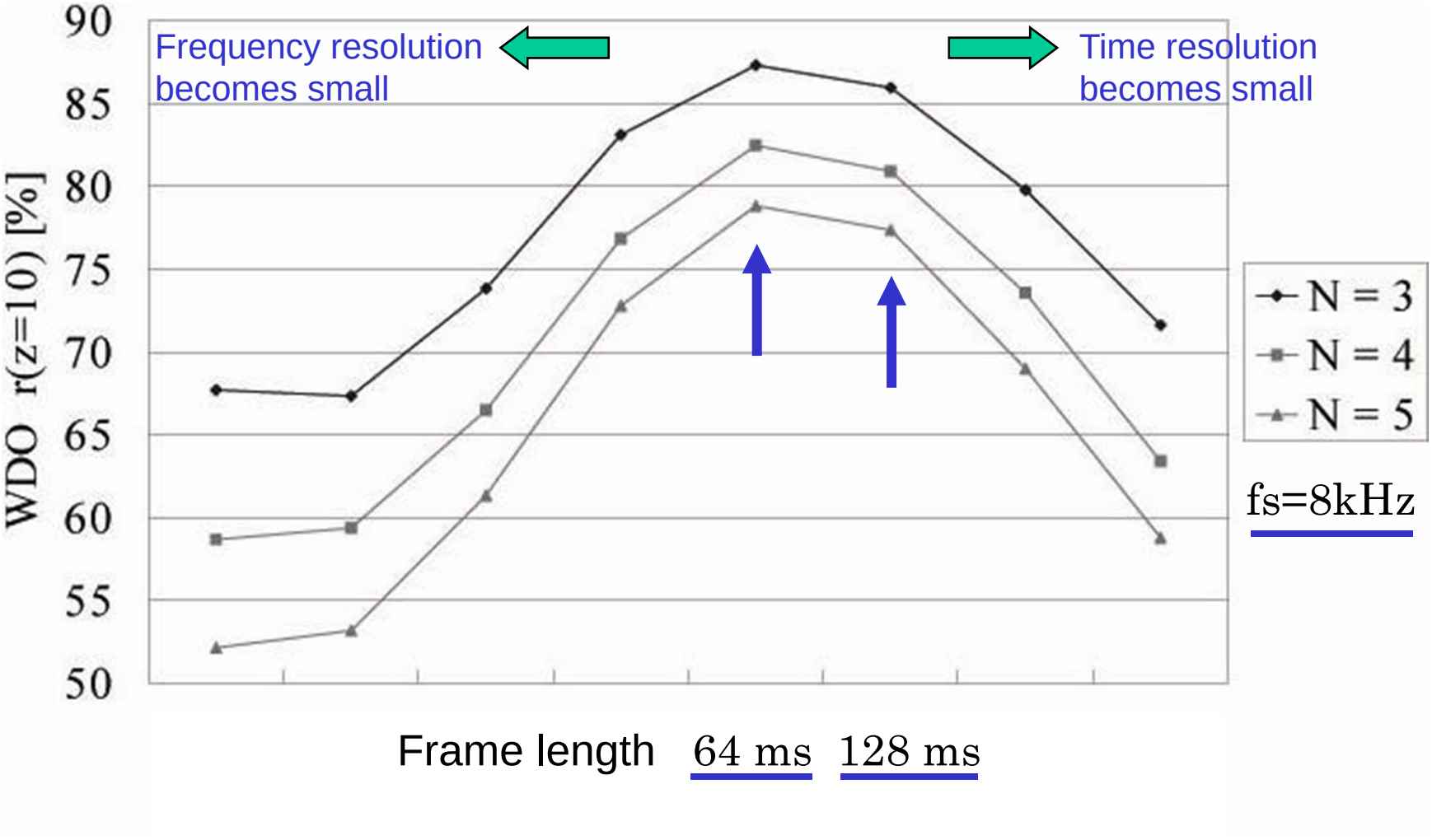


Time-frequency domain



Sparsest Frame Size

A source dominates the other sources by 10 dB



Sparse Source Separation

1. Binary Mask Approach

   Hard mask $(M(f, t) = 0 \text{ or } 1)$

   Winner takes all.

2. L1-norm Minimization Approach

   Soft mask $(0 \leq M(f,t) \leq 1)$

   Mixing matrix estimation +

   N-M source removal + Separation

Interspeech 2011 Florence

Tutorial M5

(August 27th, 2011)

Blind Speech Separation based on Independent Component Analysis and Sparse Component Analysis

Part II

Hiroshi Sawada

(NTT Communication Science Laboratories)

Tutorial structure

- Main topics of Part I
  - Basic concepts of BSS, ICA, SCA
  - Convolutive BSS
  - Frequency-domain approach
  - BSS and adaptive beamformer
- Main topics of Part II
  - Detailed procedures for FD approach
  - Frequency bin-wise separation (ICA, SCA)
  - Permutation alignment (activity, TDOA)

Outline of Part II

1. Overview of frequency domain approach
2. Frequency bin-wise separation
    1. ICA: independent component analysis
    2. SCA: sparse component analysis
3. Permutation alignment
    1. Activity sequence clustering
    2. TDOA, DOA clustering
4. Concluding remarks

Approaches to convolutive BSS

- Time-domain approach
  - Directly calculates separation filters $w_{ij}(l)$

$$y_i(t) = \sum_{j=1}^{J} \sum_{l=0}^{L-1} w_{ij}(l) x_j(t-l)$$

t Time

l Filter tap

  - Theoretically sound (no approximation)
- Frequency-domain approach
  - Approximated with instantaneous mixture model in each frequency bin

$$y_i(n,f) = \sum_{j=1}^{J} w_{ij}(f) x_j(n,f)$$

n Time frame index

f Frequency

Notations

s: sources
x: mixtures
y: separated signals
z: normalized signals

W: separation matrix
V: whitening matrix
U: unitary matrix

H: true mixing matrix
A: estimated mixing matrix

I: number of sources
J: number of microphones

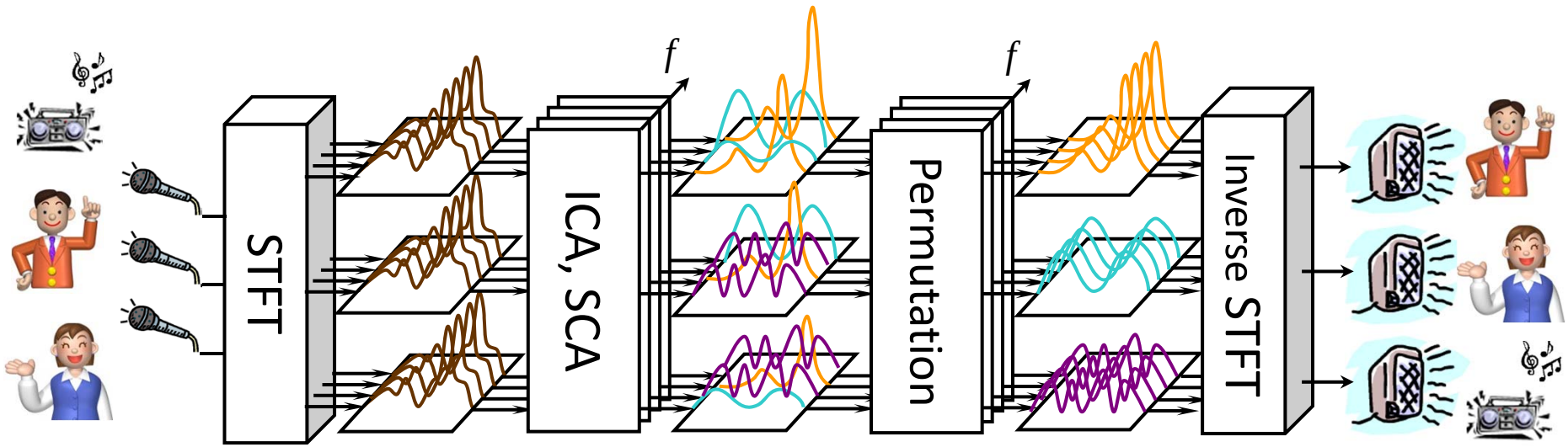
i: ICA/SCA output index
j: microphone index
k: permutation aligned output index

f: frequency
n: frame number

C: class

Flow of frequency-domain BSS

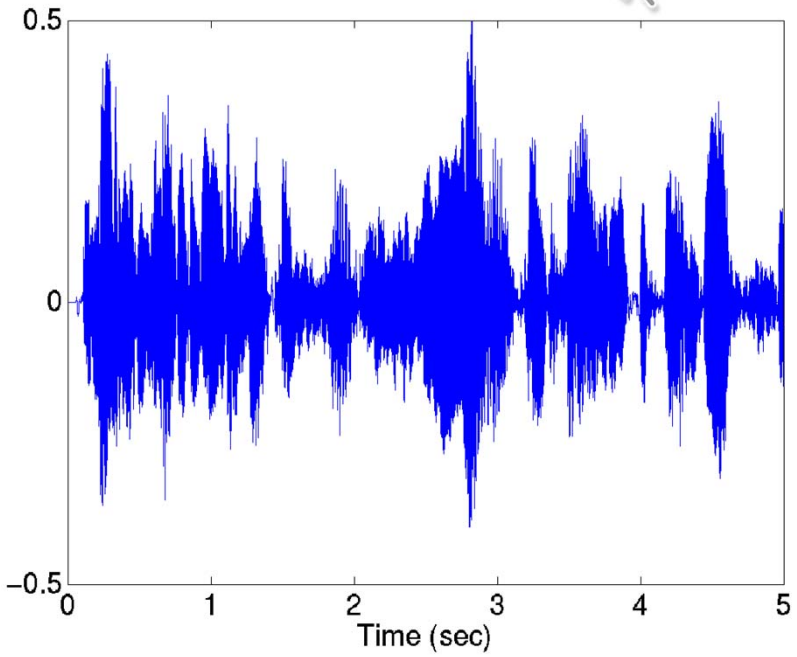
1. Time domain → Frequency domain
2. Separation of frequency bin-wise mixtures
3. Permutation alignment
4. Frequency domain → Time domain



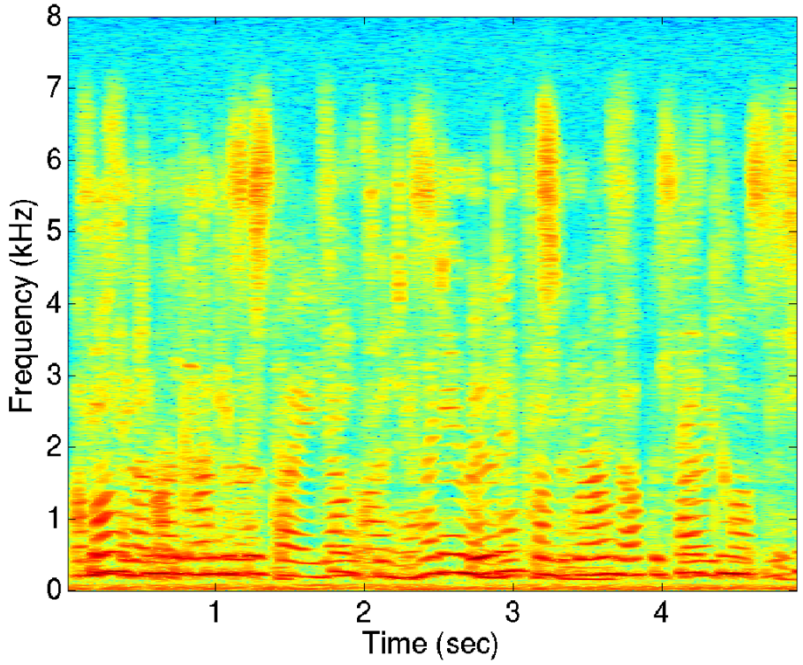
STFT: short-time Fourier transform

- From a time-domain real-valued signal
- To a time-frequency-domain complex-valued signal

$$x(t) \in \mathbb{R} \quad \xrightarrow{\text{STFT}} \quad x(n, f) \in \mathbb{C}$$



Spectrogram: only amplitudes are displayed



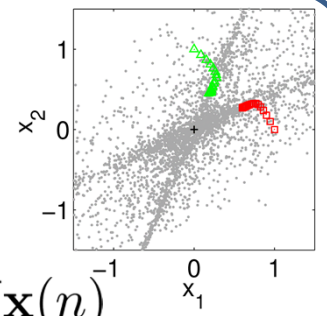
Do separation in frequency domain

STFT

$\mathbf{x}$

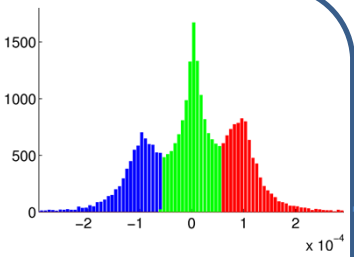
Frequency bin-wise separation

Sec. 2.1
ICA



$$\mathbf{y}(n) = \mathbf{W}\mathbf{x}(n)$$

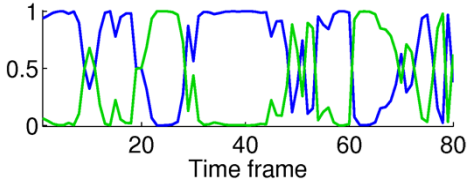
Sec. 2.2
SCA



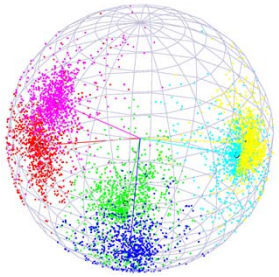
$$p(\mathbf{x}|\theta) = \sum_{i=1}^{I} \alpha_i p(\mathbf{x}|\mathbf{a}_i, \sigma_i)$$

Permutation alignment

Sec. 3.1
Activity sequence clustering



Sec. 3.2
TDOA, DOA clustering



Inverse STFT

$\mathbf{y}$

Inverse STFT

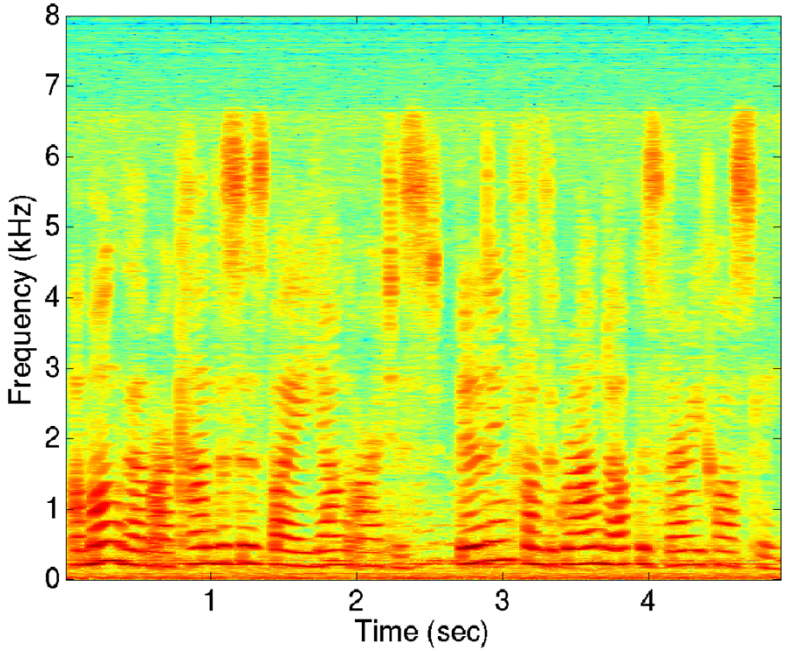
- From a time-frequency-domain complex-valued signal
- To a time-domain real-valued signal

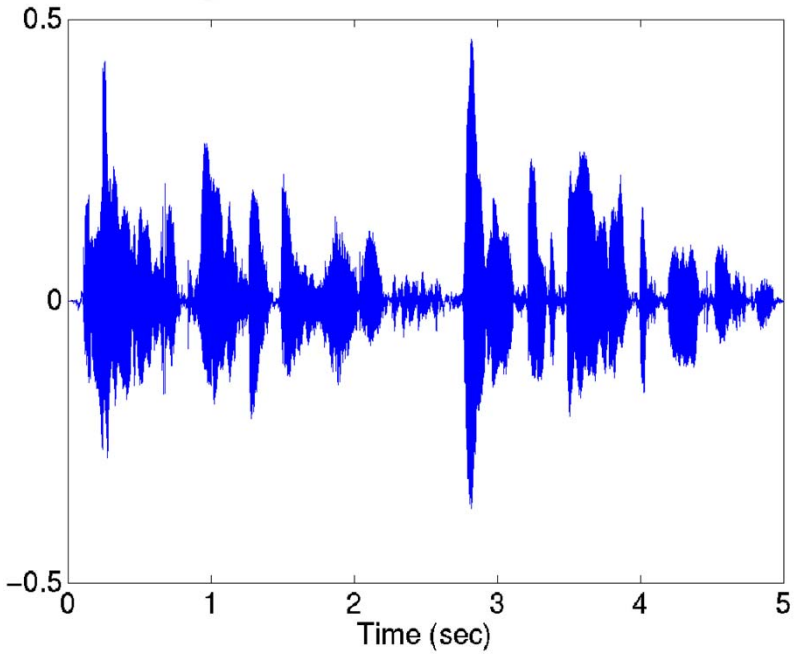
$$y(n, f) \in \mathbb{C}$$

Inverse STFT

$$y(t) \in \mathbb{R}$$

Spectrogram: only amplitudes are displayed



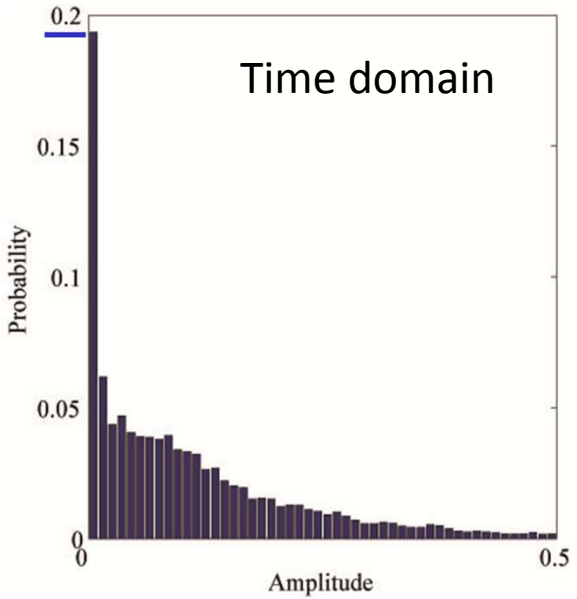


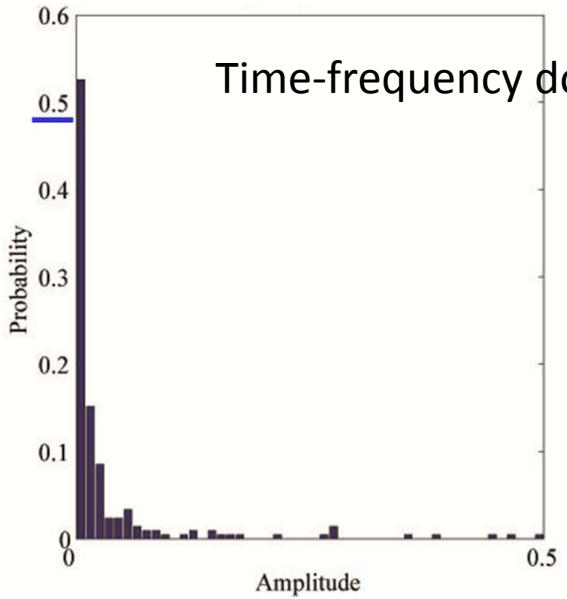
Why in frequency domain?

- Efficiency
  - Convolution is approximated by complex multiplication

$$y_i(t) = \sum_{j=1}^{J} \sum_{l=0}^{L-1} w_{ij}(l) x_j(t-l) \qquad y_i(n,f) = \sum_{j=1}^{J} w_{ij}(f) x_j(n,f)$$

- Sparser representation
  - Sources rarely overlap





Permutation ambiguity

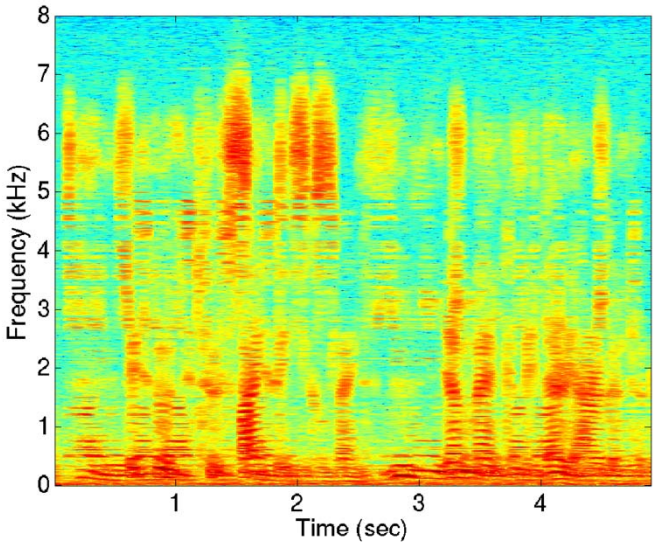
- If **y** is an ICA / SCA solution
- Then $\mathbf{y}(n) \leftarrow \mathbf{P}\mathbf{y}(n)$ is also a solution for any permutation matrix **P**

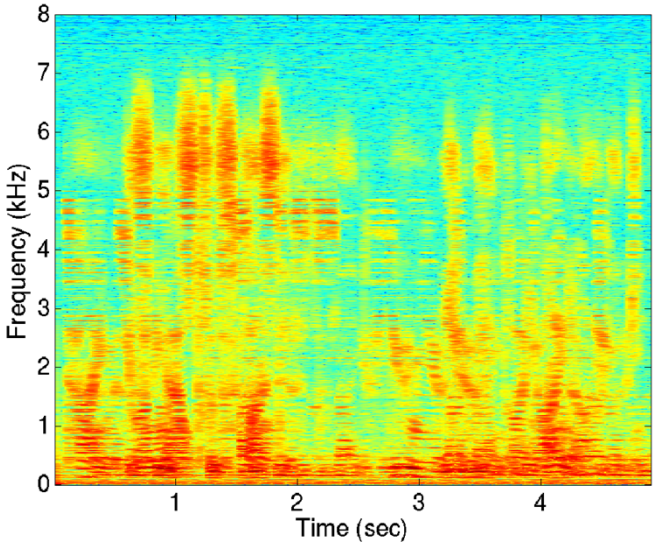
$$\begin{bmatrix} y_1(n) \\ y_2(n) \\ y_3(n) \end{bmatrix} \leftarrow \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} y_1(n) \\ y_2(n) \\ y_3(n) \end{bmatrix}$$

  - Independence of y_1, y_2, y_3 does not change
- Leads to a big problem of frequency-domain BSS

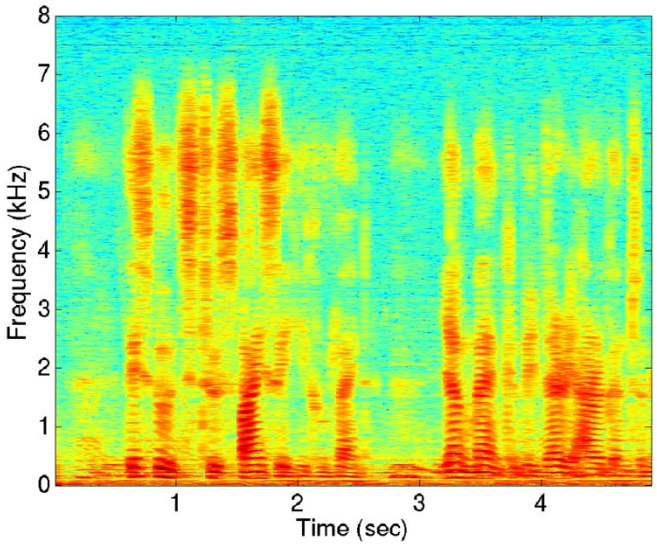
Big problem: permutation alignment

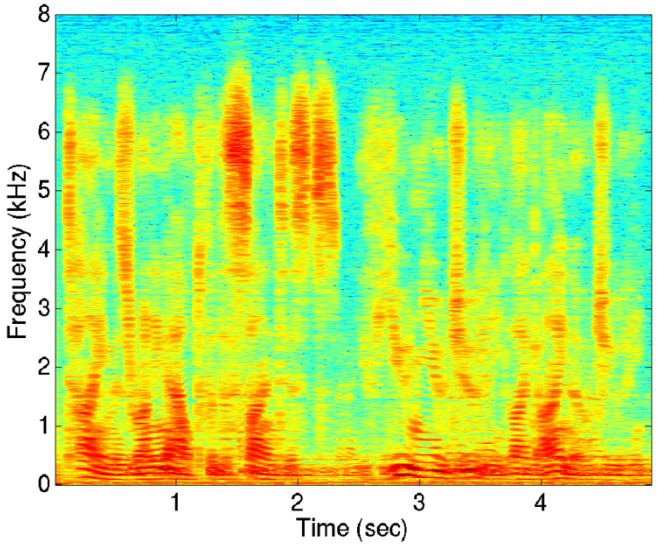
Raw ICA/SCA outputs





Permutation aligned



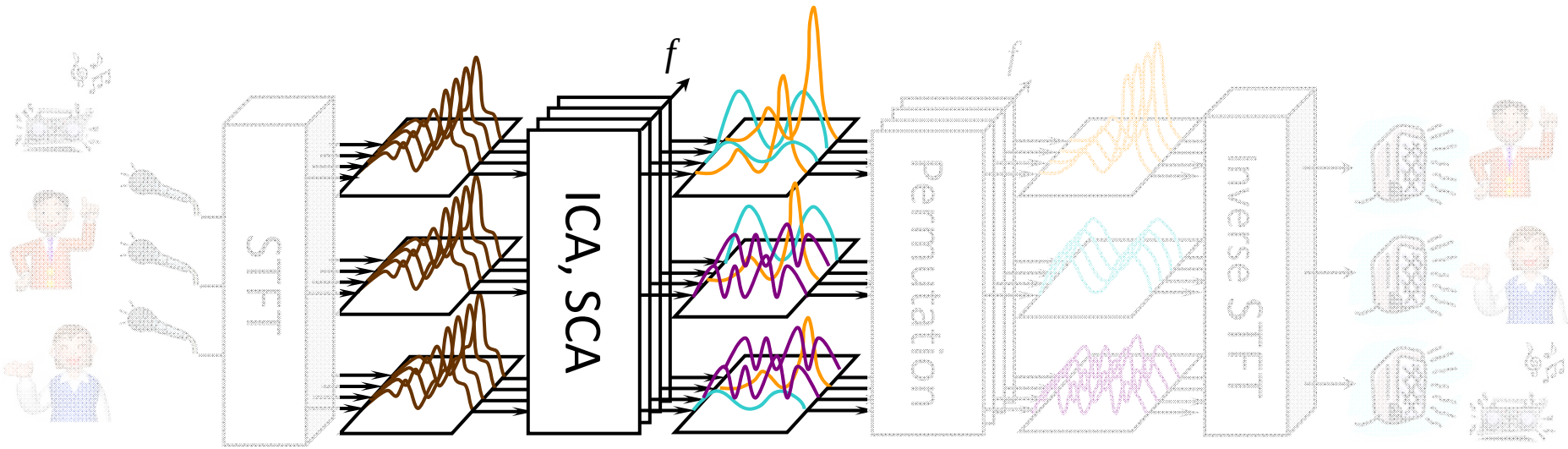


Outline of Part II

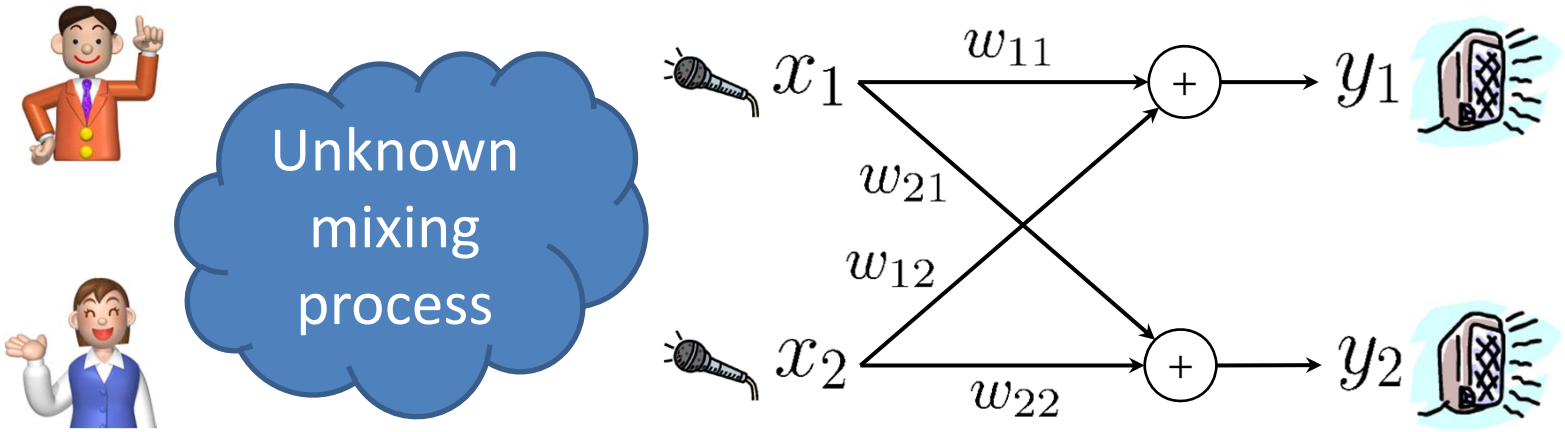
1. Overview of frequency domain approach
2. Frequency bin-wise separation
    1. ICA: independent component analysis
    2. SCA: sparse component analysis
3. Permutation alignment
    1. Activity sequence clustering
    2. TDOA, DOA clustering
4. Concluding remarks

Flow of frequency-domain BSS

1. Time domain → Frequency domain
2. Separation of frequency bin-wise mixtures
3. Permutation alignment
4. Frequency domain → Time domain



Independent component analysis



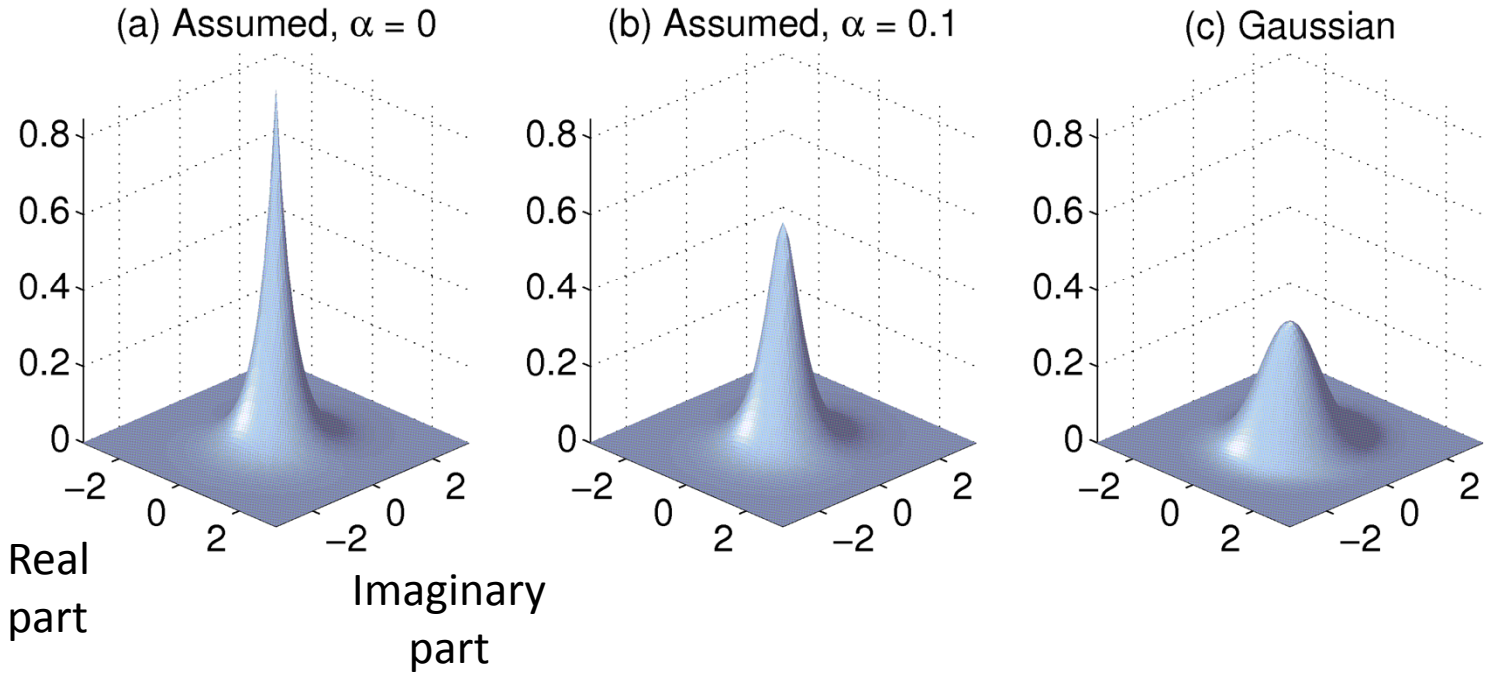
- Linear operation $\mathbf{y}(n) = \mathbf{W}\mathbf{x}(n)$
- Output independence $p(\mathbf{y}) = \prod_{i=1}^{I} p(y_i)$
- Non-gaussianity $p(y_i) \neq \frac{1}{\pi\sigma^2} \exp\left(-\frac{|y_i|^2}{\sigma^2}\right)$

Complex-valued source model

- Super-Gaussian distribution

$$p(y_i) \propto \exp\left(-\frac{\sqrt{|y_i|^2 + \alpha}}{b}\right) \qquad y_i \in \mathbb{C}$$

  - More peaky than Gaussian



Likelihood of separation matrix **W**

- Likelihood of **W** for the whole observations

$$p(\mathcal{X}|\mathbf{W}) = \prod_{n=1}^{N} p(\mathbf{x}(n)|\mathbf{W}) \qquad \mathcal{X} = \{\mathbf{x}(1), \ldots, \mathbf{x}(N)\}$$

- p.d.f, linear transformation

$$p(\mathbf{x}|\mathbf{W}) = |\det \mathbf{W}|\, p(\mathbf{y}) \quad \Leftarrow \quad \mathbf{y}(n) = \mathbf{W}\mathbf{x}(n)$$

- Output independence

$$p(\mathbf{y}) = \prod_{i=1}^{I} p(y_i)$$

Log-likelihood function

$$\mathcal{L} = \frac{1}{N} \log p(\mathcal{X}|\mathbf{W})$$
$$= \log |\det \mathbf{W}| + \sum_{i=1}^{I} \mathrm{E}\{\log p(y_i)\}$$

Maximum likelihood estimation

- Gradient descent algorithm

$$\mathbf{W} \leftarrow \mathbf{W} + \eta \cdot \frac{\partial \mathcal{L}}{\partial \mathbf{W}^*}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^*} = (\mathbf{W}^{\mathsf{H}})^{-1} - \mathrm{E}\{\mathbf{\Phi}(\mathbf{y})\mathbf{x}^{\mathsf{H}}\}$$

  - Expensive matrix inversion
  - Slow convergence

$$\mathbf{\Phi}(\mathbf{y}) = \begin{bmatrix} \phi(y_1) \\ \vdots \\ \phi(y_I) \end{bmatrix}$$

$$\phi(y_i) = -\frac{\partial \log p(y_i)}{\partial y_i^*}$$

e.g.

$$\phi(y_i) = \frac{y_i}{\sqrt{|y_i|^2 + \alpha}}$$

- Two practical ways
  - Natural gradient
  - Pre-whitening + FastICA

Natural gradient

$$\mathbf{W} \leftarrow \mathbf{W} + \eta \cdot \frac{\partial \mathcal{L}}{\partial \mathbf{W}^*} \mathbf{W}^{\mathsf{H}} \mathbf{W}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^*} \mathbf{W}^{\mathsf{H}} \mathbf{W} = \left[\mathbf{I} - \mathrm{E}\{ \mathbf{\Phi}(\mathbf{y}) \mathbf{y}^{\mathsf{H}} \} \right] \mathbf{W}$$

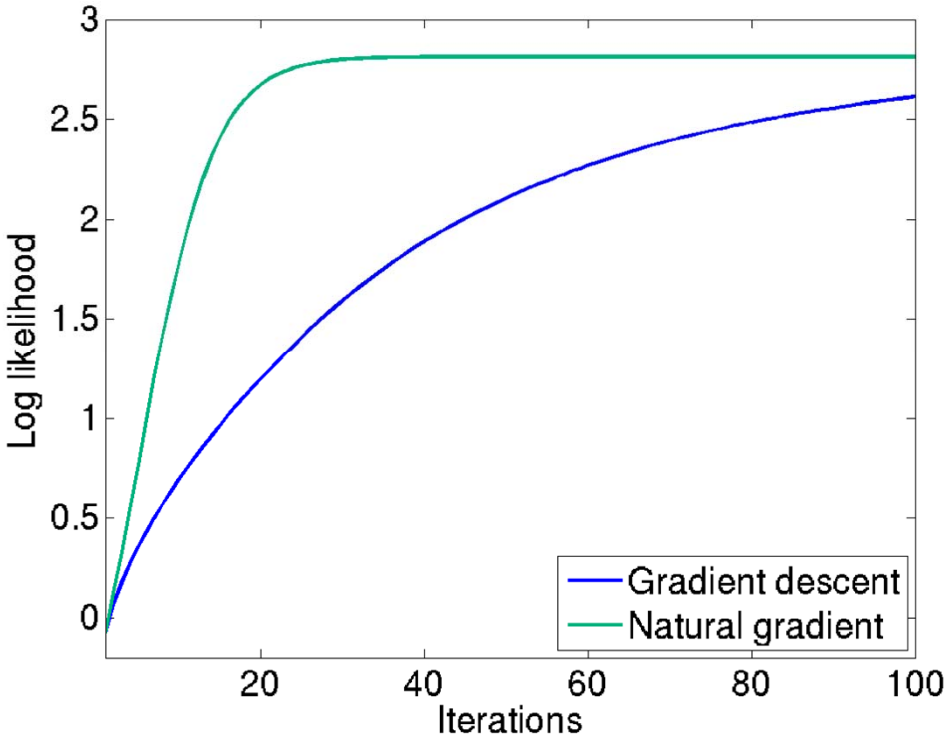
- No matrix inversion
  - Efficient computation
- Equivariance property
  - Free from the characteristics of mixing matrix (e.g. close to singular)

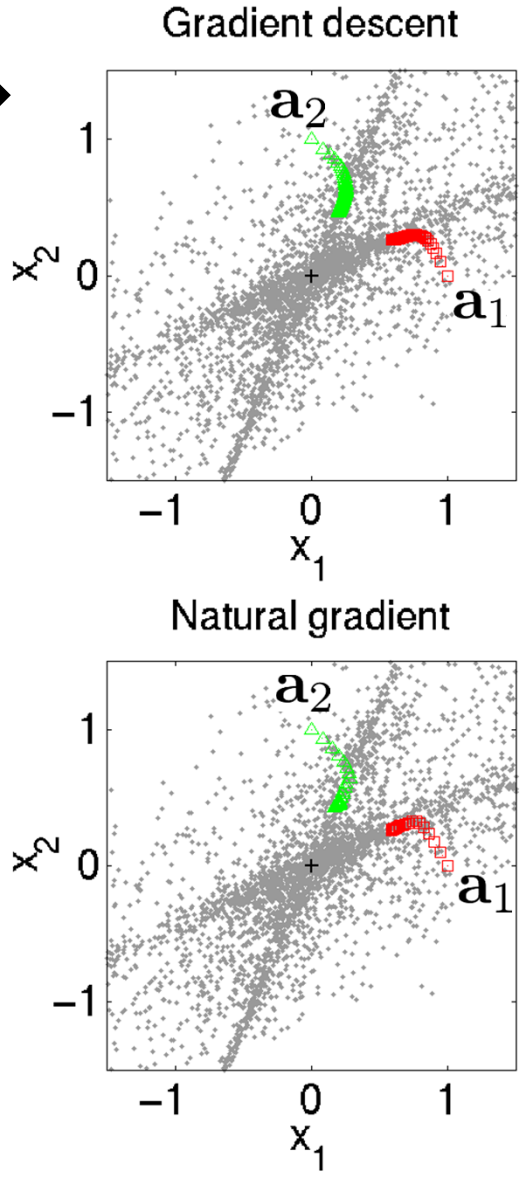
Comparison of convergence behavior

Estimated mixing matrix $\mathbf{A} = \mathbf{W}^{-1}$

Starting from $\mathbf{W} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

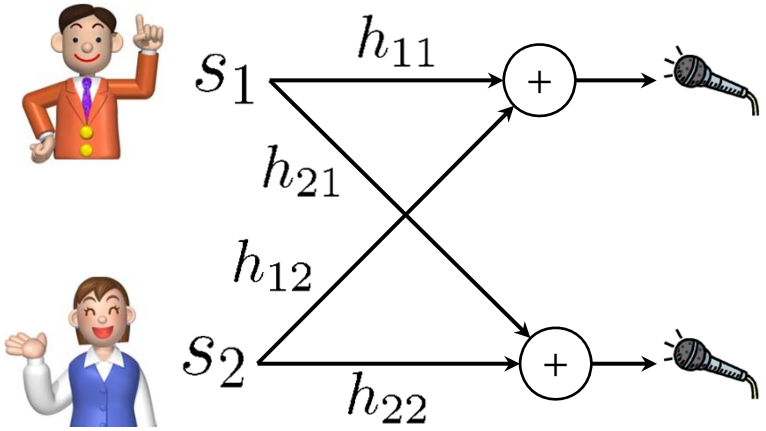
Step size: $\eta = 0.2$



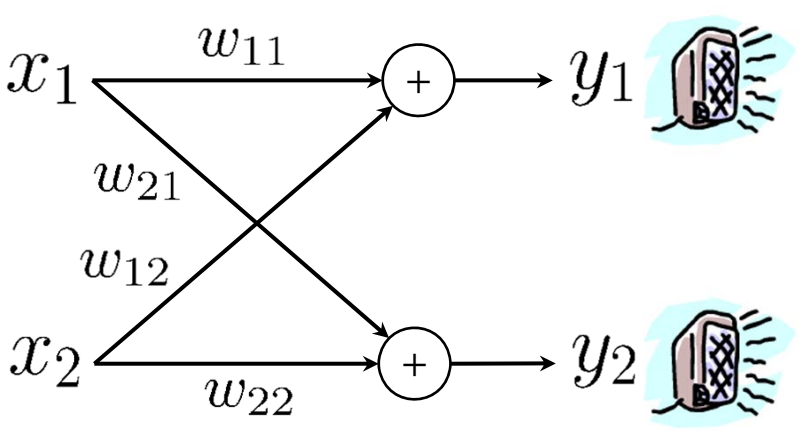


Mixing model and ICA solution

Mixing model



ICA solution

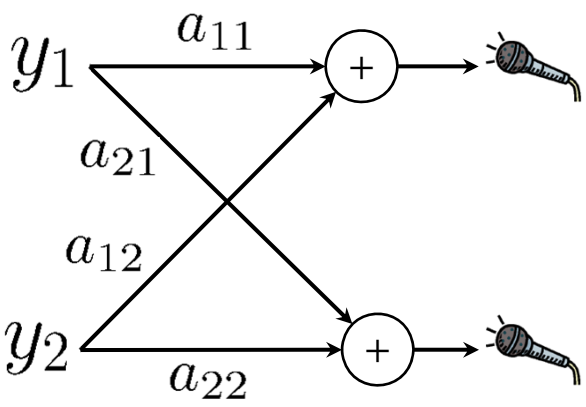


$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} \qquad \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

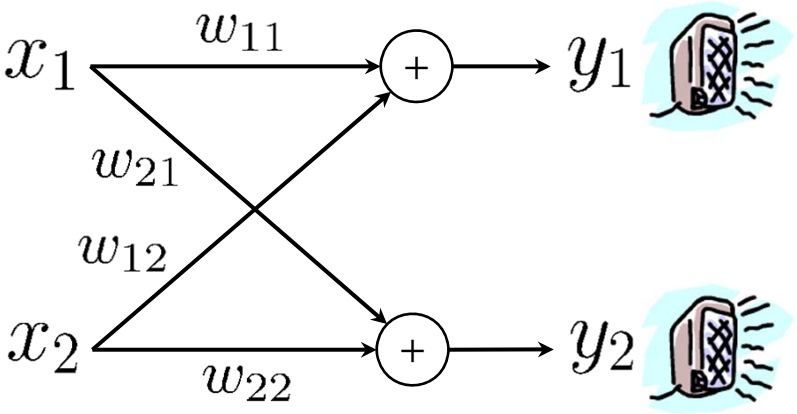
$$\mathbf{x} = \sum_{k=1}^{I} \mathbf{h}_k s_k \qquad \mathbf{y} = \mathbf{W}\mathbf{x}$$

Estimating mixing situation with ICA

Estimated mixing situation



ICA solution



$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \qquad \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\mathbf{x} = \sum_{i=1}^{I} \mathbf{a}_i y_i \quad \Leftarrow \quad \mathbf{y} = \mathbf{W}\mathbf{x}$$

Estimating mixing situation with ICA

Estimated mixing situation ⟵ ICA solution

$$\mathbf{x} = \sum_{i=1}^{I} \mathbf{a}_i y_i = \mathbf{A}\mathbf{y} \quad \longleftarrow \quad \mathbf{y} = \mathbf{W}\,\mathbf{x}$$

$$\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_I]$$

- Calculation of matrix **A**
  - If **W** has an inverse

    $$\mathbf{A} = \mathbf{W}^{-1}$$

  - Otherwise ($I < J$)

    $$\mathbf{A} = \mathrm{E}\{\mathbf{x}\mathbf{y}^H\}(\mathrm{E}\{\mathbf{y}\mathbf{y}^H\})^{-1}$$

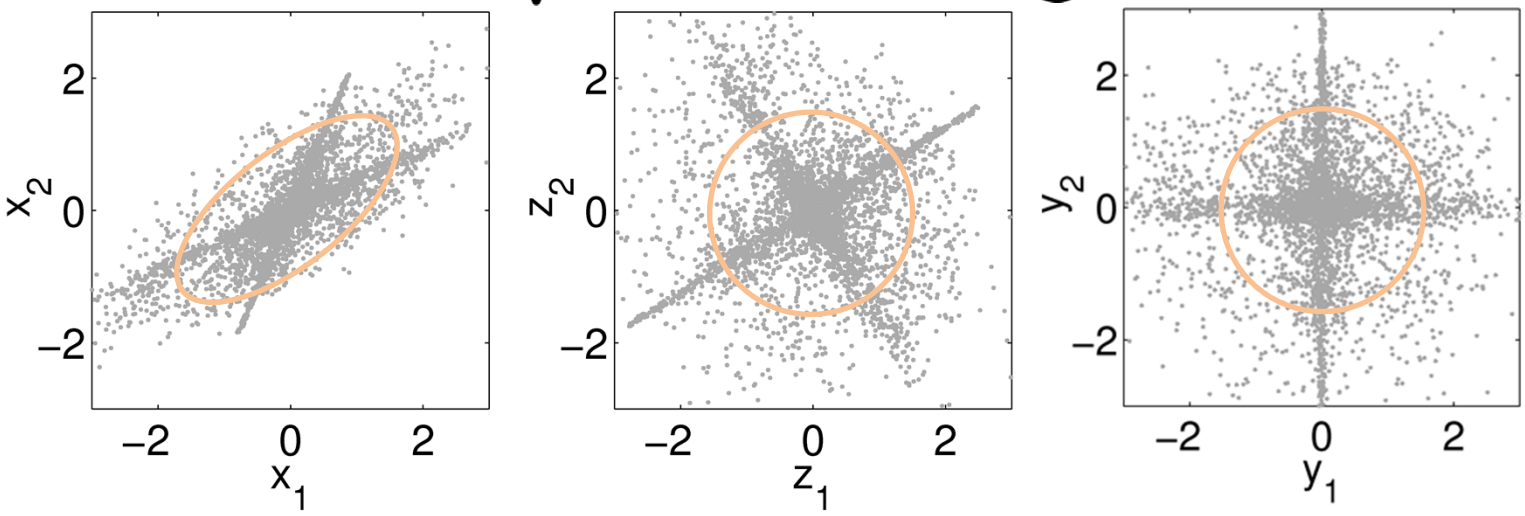
    $$\mathbf{A} = \mathbf{W}^{+}$$

- Least-mean-square estimator that minimizes $\mathrm{E}\{||\mathbf{x} - \mathbf{A}\mathbf{y}||^2\}$
- Moore-Penrose pseudo inverse

Pre-whitening + FastICA

Separation matrix of the form: $\mathbf{W} = \mathbf{UV}$

Mixtures $\xrightarrow{\mathbf{V}}$ Whitened $\xrightarrow{\mathbf{U}}$ Separated



Pre-whitening

$\mathbf{z}(n) = \mathbf{V}\mathbf{x}(n)$ s. t. $\mathrm{E}\{\mathbf{z}\mathbf{z}^{\mathsf{H}}\} = \mathbf{I}$

via eigenvalue decomposition of $\mathrm{E}\{\mathbf{x}\mathbf{x}^{\mathsf{H}}\}$

FastICA

$\mathbf{y}(n) = \mathbf{U}\mathbf{z}(n)$

Maximize the log-likelihood w.r.t. a unitary matrix **U**

FastICA

- Log-likelihood w.r.t a unitary matrix **U**

$$\mathcal{L} = \log|\det \mathbf{U}| + \sum_{i=1}^{I} \mathrm{E}\{\log p(y_i)\}$$

$$= \sum_{i=1}^{I} \mathrm{E}\{\log p(y_i)\} \quad \Leftarrow \quad \det \mathbf{U} = 1$$

- Let $G(y_i) = \log p(y_i) = -\sqrt{|y_i|^2 + \alpha}$ e.g. $\alpha = 0.1$
- Maximize $\mathrm{E}\{G(y_i)\}$ for $y_1, y_2, \ldots, y_I$
  - with unitary constraint

$$\mathbf{u}_i^{\mathsf{H}} \mathbf{u}_{i'} = \begin{cases} 1 & \text{if } i = i' \\ 0 & \text{if } i \neq i' \end{cases} \qquad \mathbf{U} = [\mathbf{u}_1 \cdots \mathbf{u}_I]^{\mathsf{H}}$$

FastICA algorithm

- For $i = 1, \ldots, I$ (sequentially)
  - Iterate the followings until convergence
    - $y_i = \mathbf{u}_i^{\mathsf{H}} \mathbf{z}$ Separated signal calculation
    - $\mathbf{u}_i \leftarrow \mathrm{E}\{G''(y_i)\}\mathbf{u}_i - \mathrm{E}\{G'(y_i)\mathbf{z}\}$

Optimization of G by Newton's method

$$G'(y_i) = \frac{\partial G}{\partial y_i^*} = -\frac{y_i}{2\sqrt{|y_i|^2 + \alpha}} \qquad G''(y_i) = \frac{\partial G'}{\partial y_i} = -\frac{1}{2\sqrt{|y_i|^2 + \alpha}} \left[1 - \frac{1}{2} \frac{|y_i|^2}{|y_i|^2 + \alpha}\right]$$

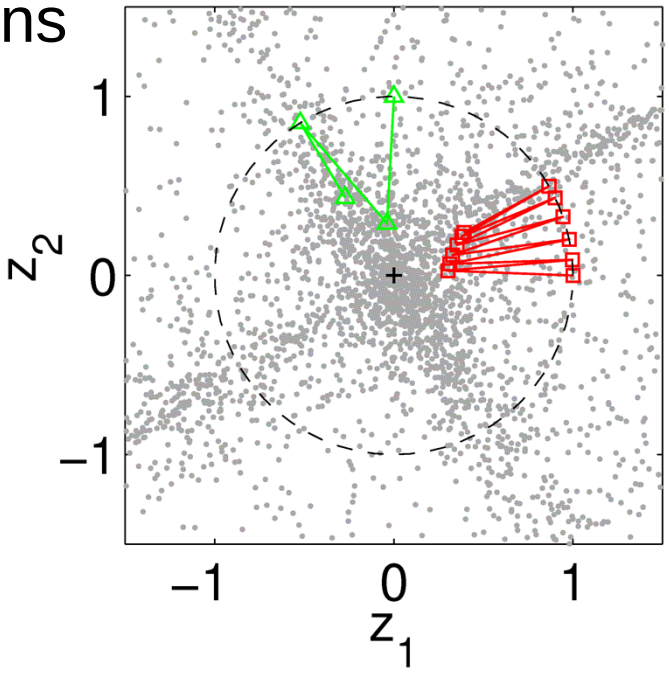
- $\mathbf{u}_i \leftarrow \mathbf{u}_i - \sum_{k=1}^{i-1} (\mathbf{u}_k^{\mathsf{H}} \mathbf{u}_i) \mathbf{u}_k$

Gram-schmidt orthogonalization

- $\mathbf{u}_i \leftarrow \frac{\mathbf{u}_i}{||\mathbf{u}_i||}$ Unit-norm normalization

FastICA convergence example

- Red (□)
  - Starting from $\mathbf{u}_1 = [1 \;\; 0]^T$
  - Likelihood maximization: points close to the origin
  - Unit-norm normalization: points on the unit sphere
  - Good solution only by 5 iterations
- Green (△)
  - Starting from $\mathbf{u}_2 = [0 \;\; 1]^T$
  - One-step solution by orthogonalization



Section 2.1 Summary

- Complex-valued ICA
  - Natural gradient
  - FastICA
- Applied to frequency bin-wise separation
  - Perform ICA for, e.g. 513, times separately
  - Permutation alignment is conducted afterwards
- #sources $\leq$ #microphones ($I \leq J$)
  - Otherwise, another method, e.g., sparse component analysis (SCA), should be employed

Outline of Part II

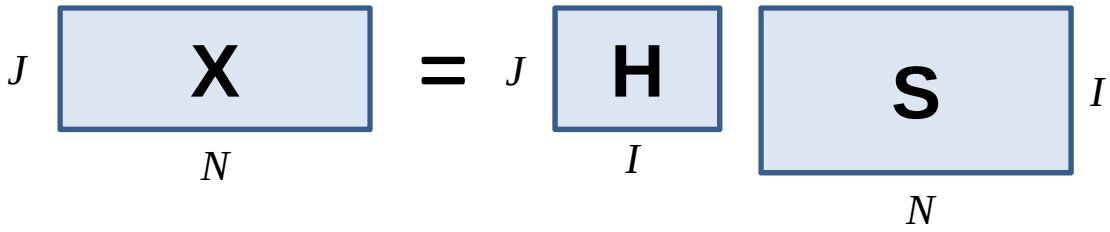
1. Overview of frequency domain approach
2. Frequency bin-wise separation
    1. ICA: independent component analysis
    2. SCA: sparse component analysis
3. Permutation alignment
    1. Activity sequence clustering
    2. TDOA, DOA clustering
4. Concluding remarks

Sparse Component Analysis

- Estimate sources **s** (and mixing matrix **H**)
  - From observation mixtures **x**

$$\mathbf{x}(n) = \mathbf{H}\mathbf{s}(n)$$

- Especially important when #sources > #microphones

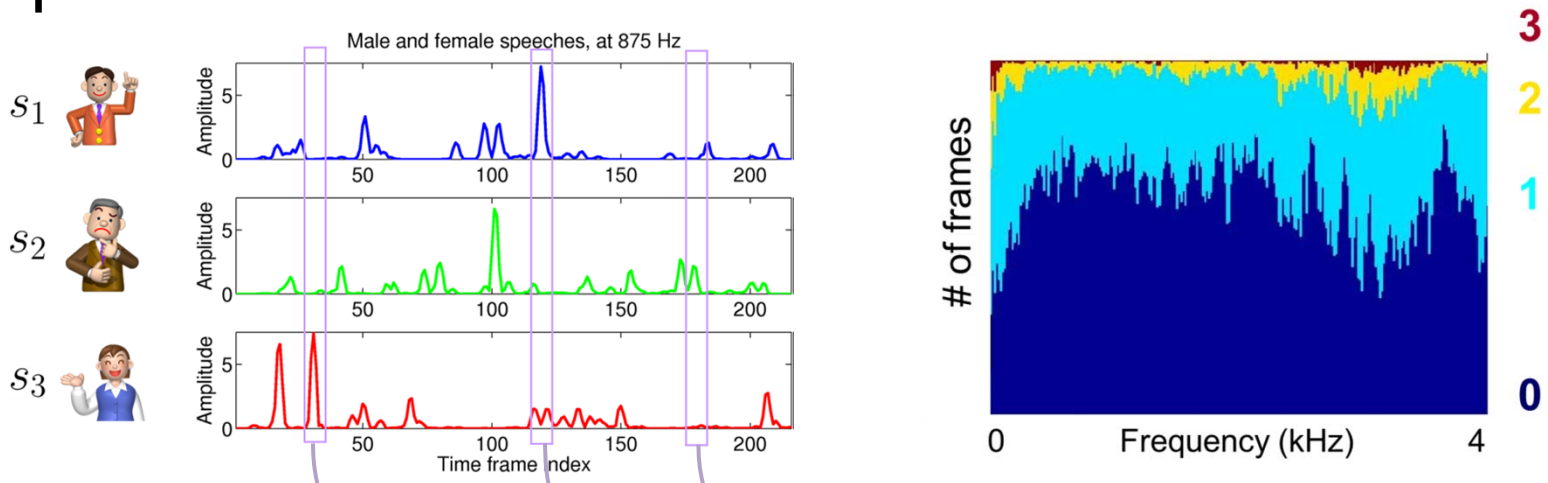


  - Otherwise ICA $\mathbf{y}(n) = \mathbf{W}\mathbf{x}(n)$ can be applied
    - Sources $\hat{\mathbf{s}}(n) = \mathbf{y}(n)$
    - Mixing matrix $\hat{\mathbf{H}} = \mathbf{W}^{-1}$

$(I < J)$ $\hat{\mathbf{H}} = \mathbf{W}^{+}$ Moore-Penrose pseudo inverse

SCA by clustering

- Sparseness of sources **s** is assumed



- Cluster observation vectors **x** to each source
  - $P(C_i|\mathbf{x}(n))$:posterior probability for class *i*

$P(C_1\|\mathbf{x})$	0.01	0.85	0.11
$P(C_2\|\mathbf{x})$	0.04	0.05	0.88
$P(C_3\|\mathbf{x})$	0.95	0.10	0.01

Source estimation methods in SCA

- L1-norm minimization
  - When mixing matrix **H** is also estimated

$$\mathbf{y}(n) = \hat{\mathbf{s}}(n) = \arg\min_{\mathbf{x}(n)=\hat{\mathbf{H}}\mathbf{s}(n)} \sum_{i=1}^{I} |s_i(n)|$$

- Binary masking
  - Masks are designed from the clustering results (posterior probabilities)

$$\mathcal{M}_i(n) = \begin{cases} 1 & \text{if } P(C_i|\mathbf{x}(n)) \geq P(C_{i'}|\mathbf{x}(n)), \quad \forall i' \neq i \\ 0 & \text{otherwise.} \end{cases}$$

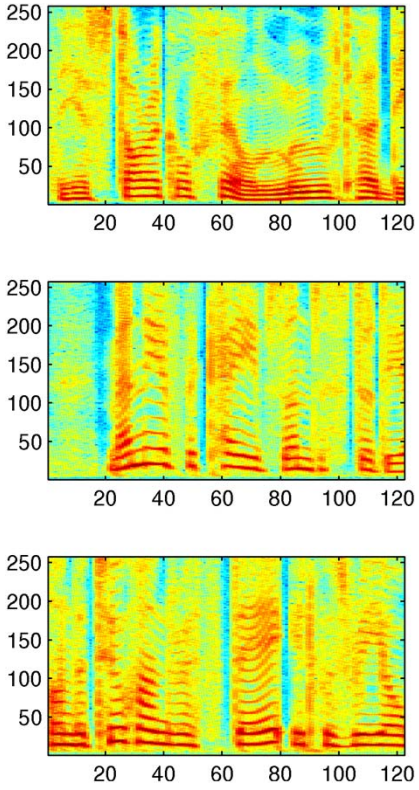
  - Sources are estimated simply by

$$y_i(n) = \hat{s}_i(n) = \mathcal{M}_i(n) \cdot x_1(n)$$

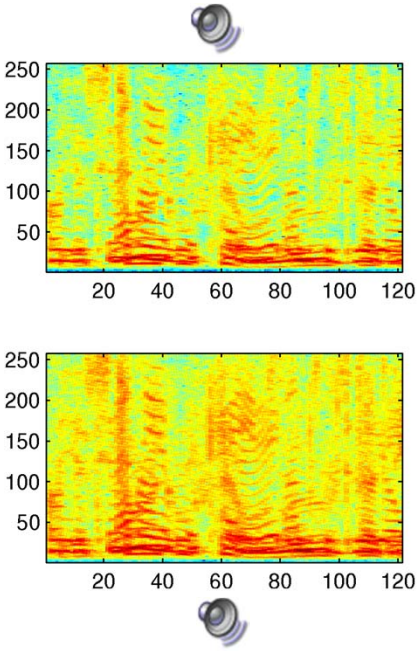
Time-frequency masking

- Binary masking applied to time-frequency representations

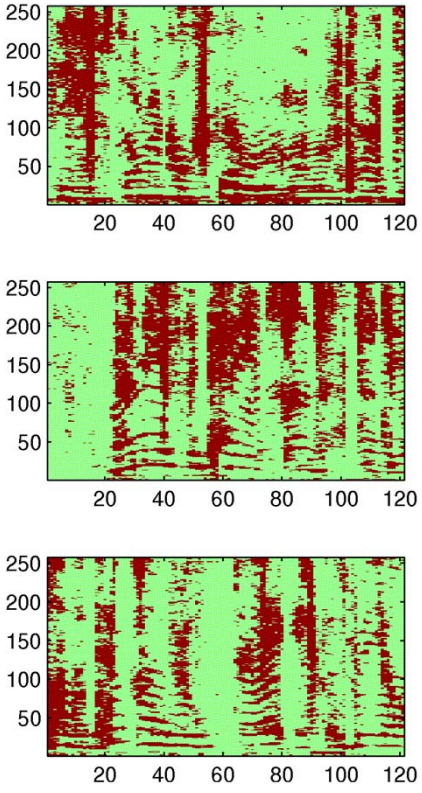
3 sources $s_k(n, f)$



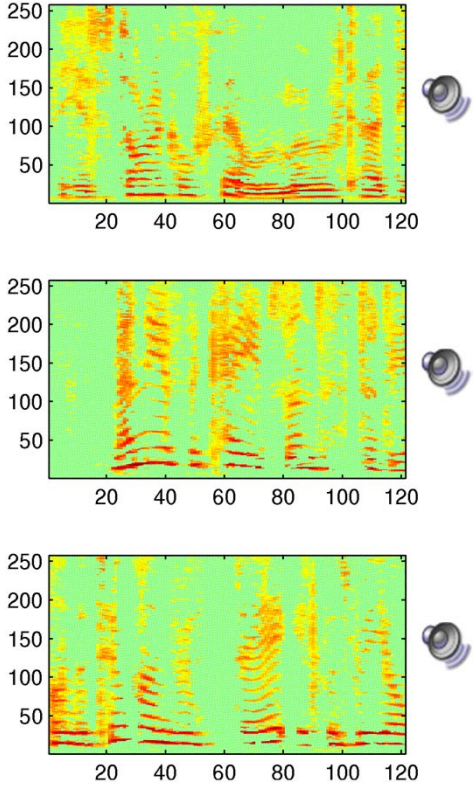
2 mixtures $x_j(n, f)$



Binary masks $\mathcal{M}_i(n, f)$



Separations $y_i(n, f)$



Simple method: Full-band TDOA clustering

STFT

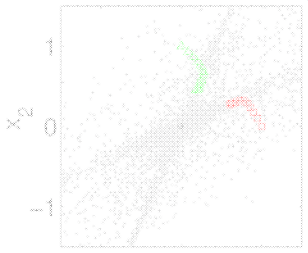
$\mathbf{x}$

SCA is performed in a full-band manner
(no need to align permutations)

Inverse STFT

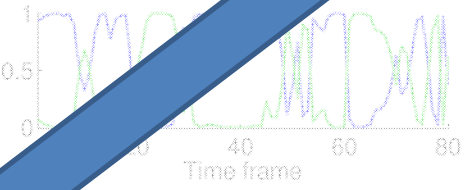
$\mathbf{y}$

Sec. 2.1
ICA

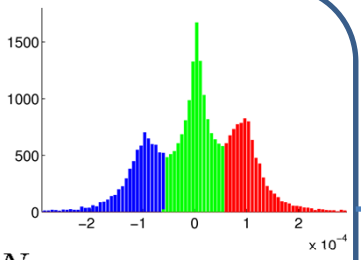


$$\mathbf{y}(\tau) = \mathbf{W}\,\mathbf{x}(\tau)$$

Sec. 3.1
Activity sequence clustering

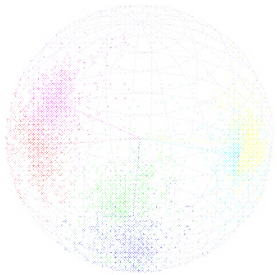


Sec. 2.2
SCA



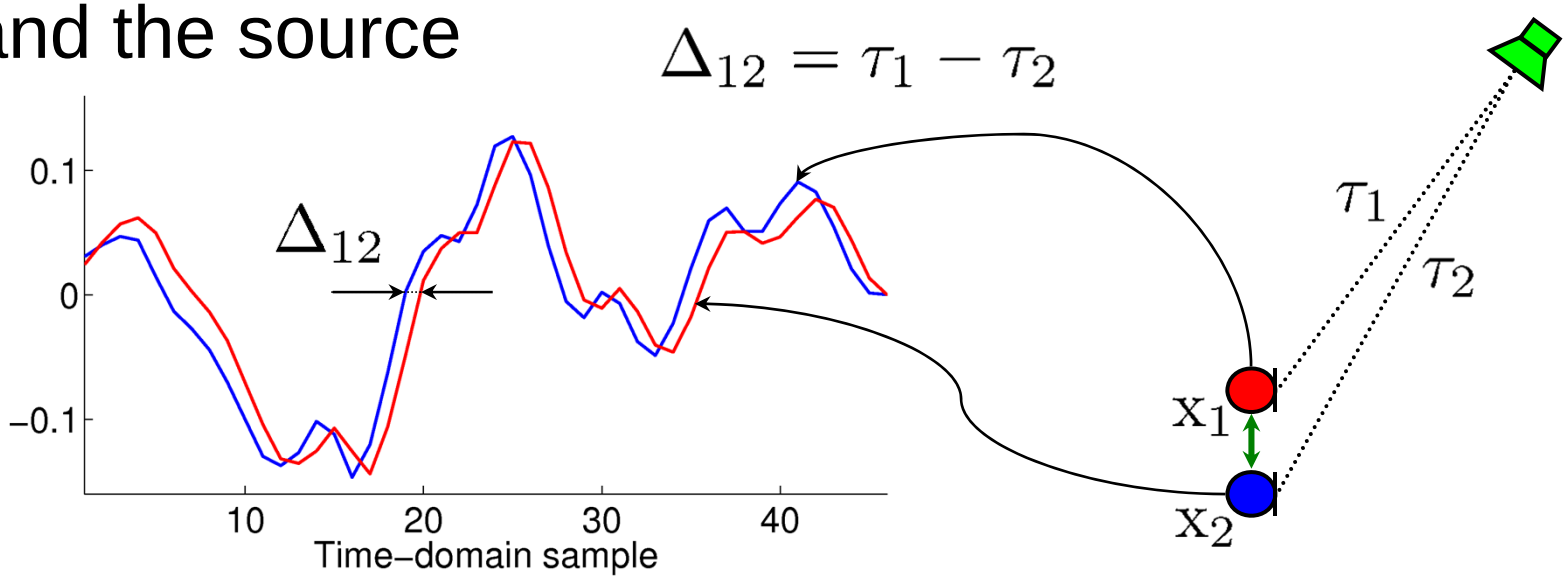
$$p(\mathbf{x}|\theta) = \sum_{i=1}^{N} \alpha_i\, p(\mathbf{x}|\mathbf{a}_i, \sigma_i)$$

Sec. 3.2
TDOA, DOA clustering



Time-difference-of-arrival (TDOA)

- Estimated for each source
- Caused by the positions of microphones and the source



- (Generalized) cross correlation

$$\Delta_{12} = \text{argmax}_\Delta \, x_1(t) \, x_2(t - \Delta)$$

$$\Delta_{12} = \text{argmax}_\Delta \sum_f \Phi(f) \, x_1(f) x_2^*(f) e^{\imath 2\pi f \Delta}$$

GCC-PHAT

$$\Phi(f) = \frac{1}{|x_1(f) x_2^*(f)|}$$

Frequency-dependent TDOA

- TDOA estimated at each time-frequency slot

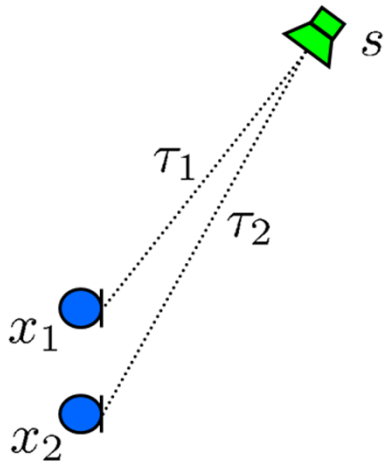
$$\Delta_{12}(n,f) = \frac{\arg[x_1(n,f)/x_2(n,f)]}{-2\pi f}$$

- Derivation of the formula
  - Frequency domain, single source $\begin{bmatrix} x_1(n,f) \\ x_2(n,f) \end{bmatrix} = \begin{bmatrix} h_1(f) \\ h_2(f) \end{bmatrix} s(n,f)$
  - Anechoic model $h_j(f) \approx \exp(-\imath 2\pi f \tau_j)$
  - We have

$$\frac{x_1(n,f)}{x_2(n,f)} = \frac{h_1(f)s(n,f)}{h_2(f)s(n,f)} \approx \exp\left[-\imath 2\pi f(\tau_2 - \tau_1)\right]$$

  - Taking the argument gives the formula

$$\Delta_{12}(n,f) = \tau_2 - \tau_1 = \frac{\arg[x_1(n,f)/x_2(n,f)]}{-2\pi f}$$



Valid frequency range

- The argument should be in the range

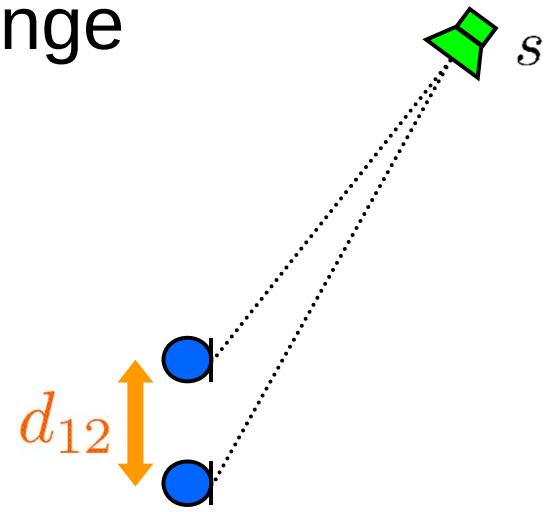
$$-\pi < \arg(\cdot) \leq \pi$$

$$-\pi < -2\pi f(\tau_2 - \tau_1) \leq \pi$$

- TDOA is bounded by

$$|\tau_2 - \tau_1| \leq d_{12}/v$$

velocity



- Frequency range for valid TDOA estimation

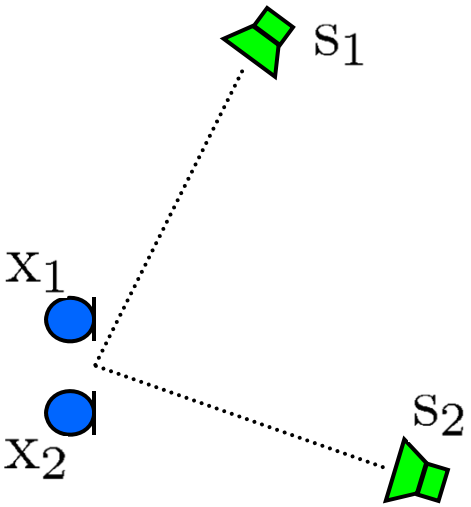
$$0 < f < \frac{v}{2\,d_{12}}$$

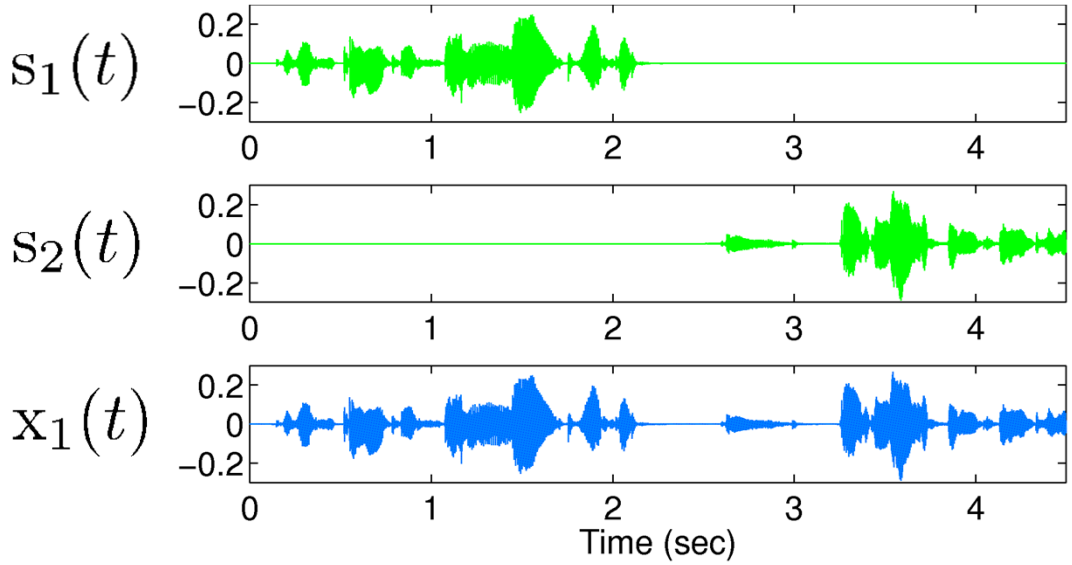
if $d_{12} = 4$ cm

then $0 < f < 4250$ Hz

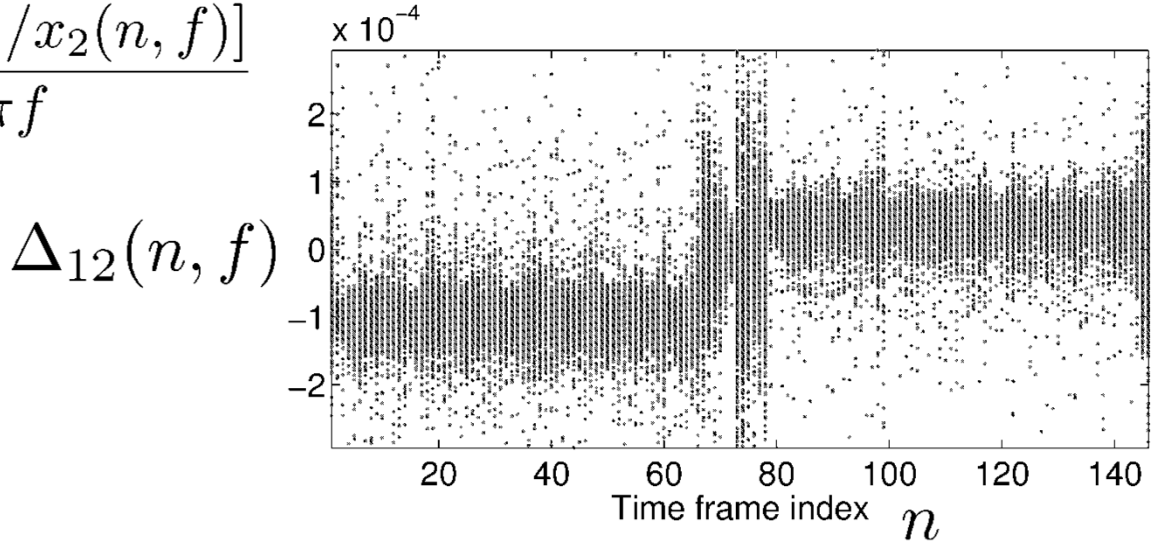
  - otherwise, spatial aliasing might occur

TDOAs for non-overlapped mixtures

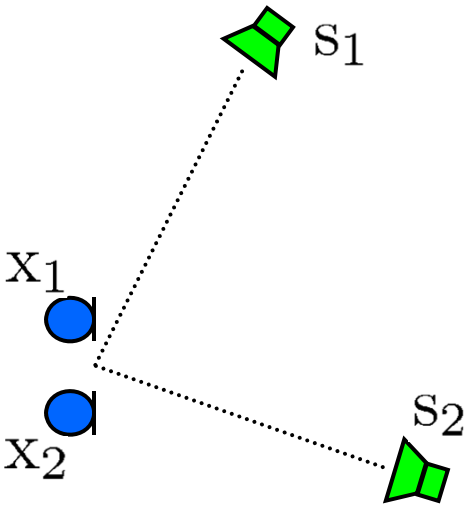


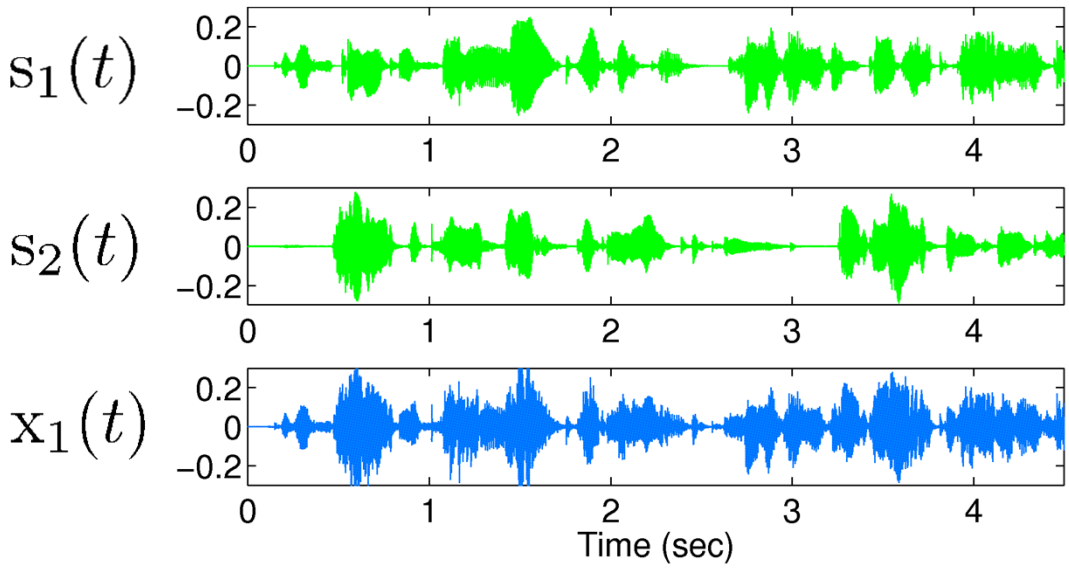


$$\Delta_{12}(n,f) = \frac{\arg[x_1(n,f)/x_2(n,f)]}{-2\pi f}$$

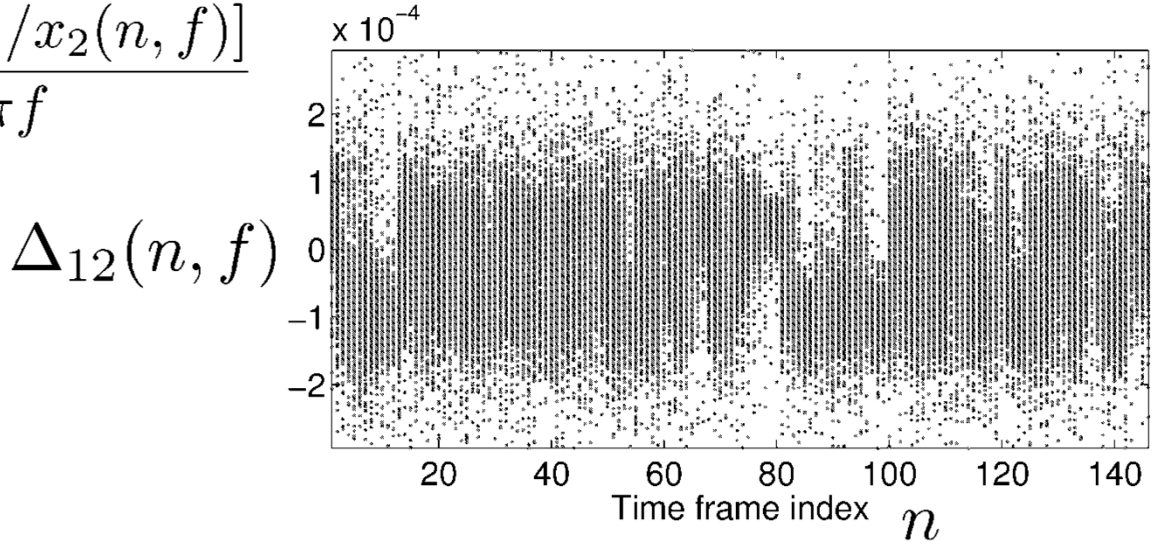


TDOAs for overlapped mixtures



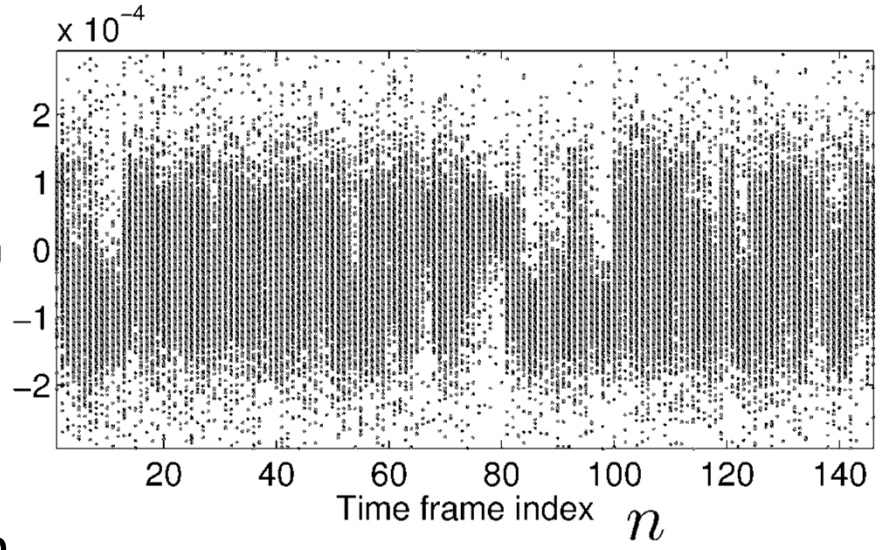


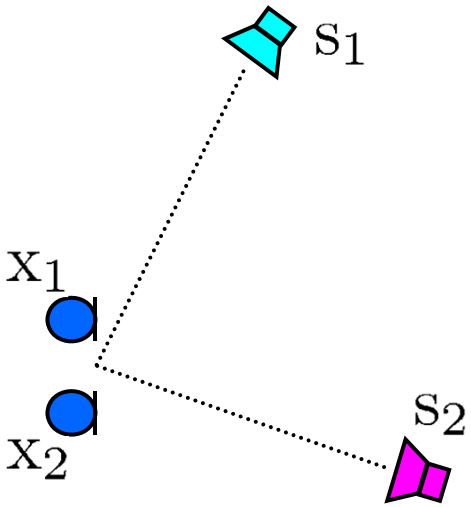
$$\Delta_{12}(n,f) = \frac{\arg[x_1(n,f)/x_2(n,f)]}{-2\pi f}$$



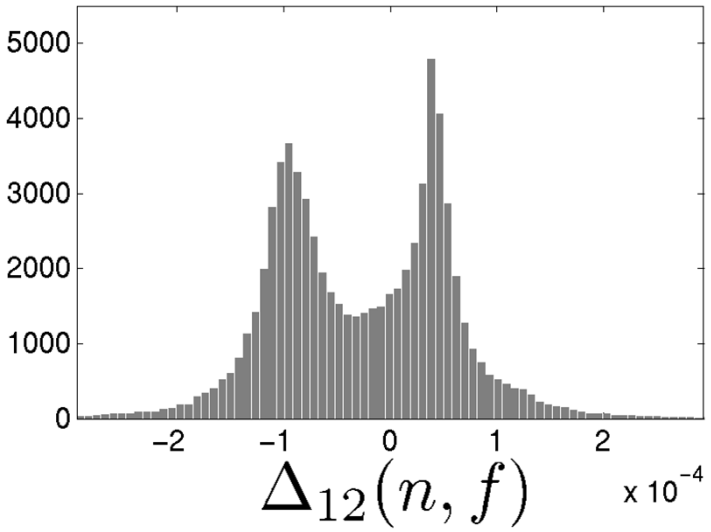
Clustering TDOAs

$\Delta_{12}(n, f)$





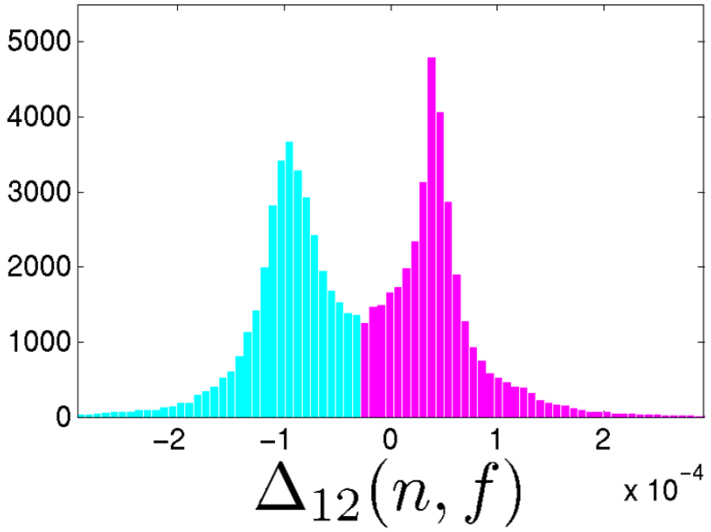
Histogram



Clustering

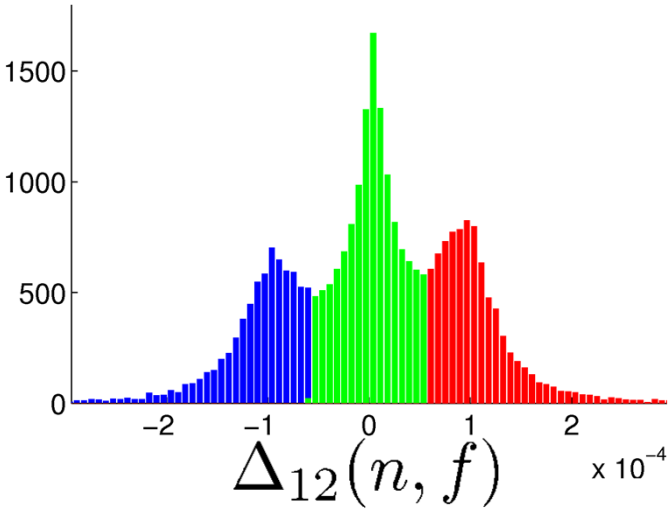
- K-means
- GMM

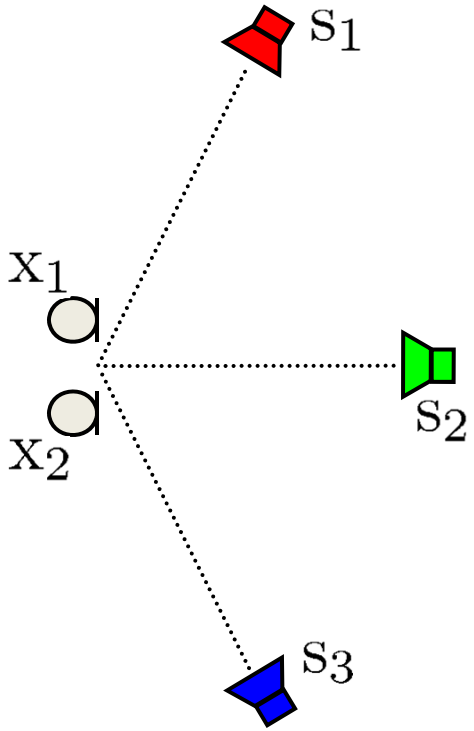
$P(C_k|\mathbf{x})$



T-F mask design for 3-source mixtures

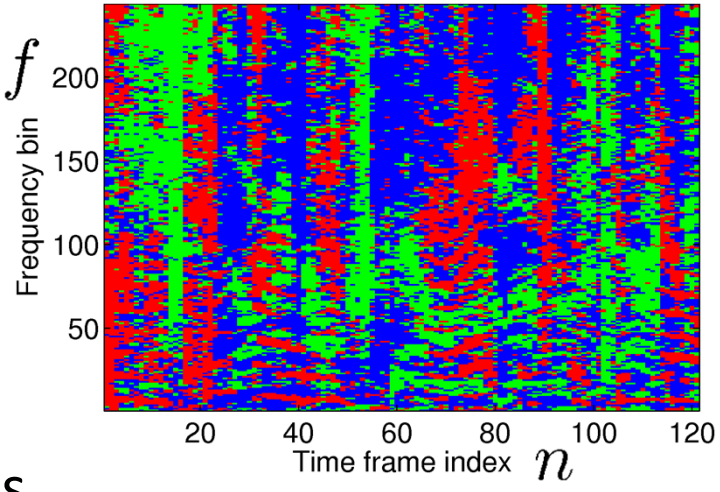
- Clustering TDOA





- Mask designed

$$\mathcal{M}_1(n,f) = 1$$
$$\mathcal{M}_2(n,f) = 1$$
$$\mathcal{M}_3(n,f) = 1$$



- Separated signals

$$y_i(n,f) = \mathcal{M}_i(n,f) \cdot x_1(n,f)$$

Summary: Full-band TDOA clustering

- Simple
  - TDOA calculation $\Delta_{12}(n,f) = \dfrac{\arg[x_1(n,f)/x_2(n,f)]}{-2\pi f}$
  - No need for permutation alignment
- Problems
  - Spatial aliasing
    - Valid frequency ranges are limited
  - Separation performance becomes worse as the reverberation time increases
    - Anechoic model does not hold

$$h_j(f) \approx \exp(-\imath 2\pi f \tau_j)$$

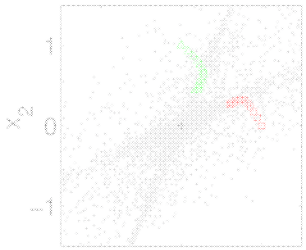
Two-stage method

STFT

$\mathbf{x}$

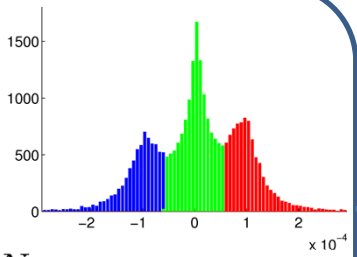
Frequency bin-wise separation

Sec. 2.1
ICA



$$\mathbf{y}(\tau) = \mathbf{W}\,\mathbf{x}(\tau)$$

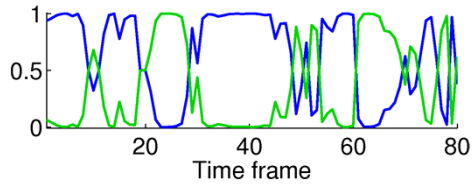
Sec. 2.2
SCA



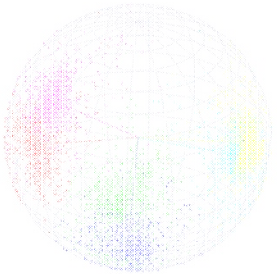
$$p(\mathbf{x}|\theta) = \sum_{i=1}^{N} \alpha_i\, p(\mathbf{x}|\mathbf{a}_i, \sigma_i)$$

Permutation alignment

Sec. 3.1
Activity sequence clustering



Sec. 3.2
TDOA, DOA clustering



Inverse STFT

$\mathbf{y}$

The problems shown in the previous slide can be solved.

Spectrogram examples

(a) Sources

$s_k(n,f)$

$k = 1,2,3$

(b) Mixtures

$\mathbf{x}(n,f) = \begin{bmatrix} x_1(n,f) \\ x_2(n,f) \end{bmatrix}$

(c) Bin-wise classification

$P(C_i|\mathbf{x})$

$i = 1,2,3$

(d) Permutation aligned classification

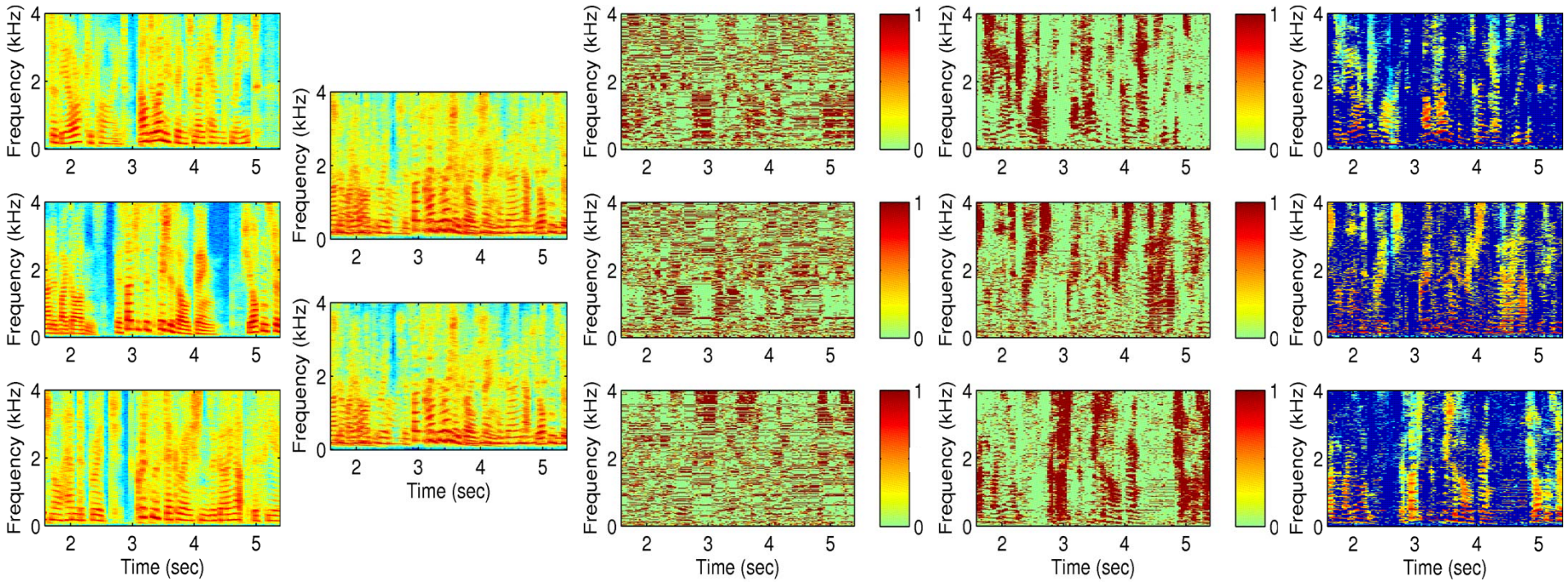
$P(C_k|\mathbf{x})$

$k = 1,2,3$

(e) Separated signals

$y_k(n,f)$

$k = 1,2,3$



Sec. 2.2

Sec. 3.1

Frequency bin-wise SCA

- Mixing model

$$\mathbf{x}(n,f) = \sum_{k=1}^{I} \mathbf{h}_k(f) s_k(n,f)$$

- With sparseness assumption
  - Only one dominant source at a time-frequency slot

$$\mathbf{x}(n,f) = \mathbf{h}_{i^\star}(f) s_{i^\star}(n,f)$$

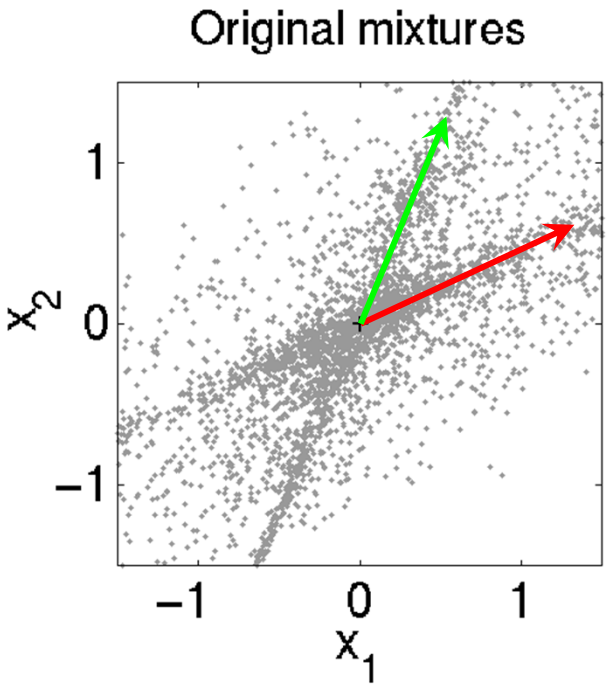
    - The dominant source index $i^\star \in \{1,\ldots,I\}$ depends on each time-frequency slot

- Binary mask design

$$\mathcal{M}_i(n,f) = \begin{cases} 1 & \text{if } i^\star(n,f) = i \\ 0 & \text{otherwise.} \end{cases}$$

Line orientation idea

- Mixtures of sparse components are distributed mainly along multiple lines
  - Each line represents a mixing vector $\mathbf{h}_k$
- Finding such lines leads to SCA



Modeling with Gaussian-like distribution

- Samples from one sparse component
  - modeled with Gaussian-like distribution

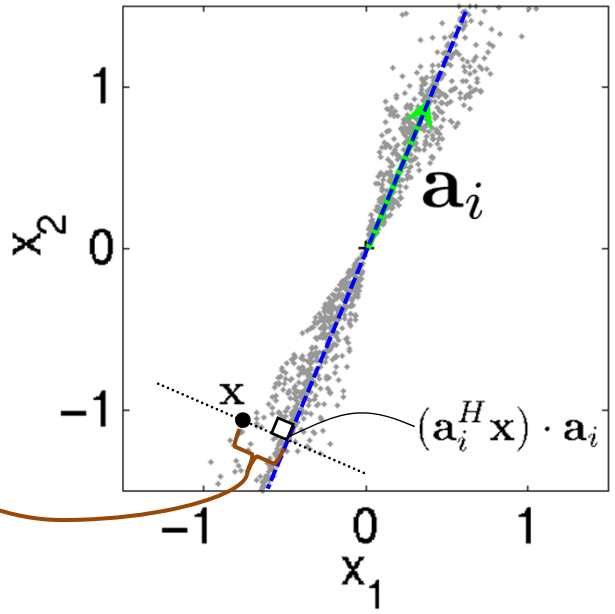
$$p(\mathbf{x}|\mathbf{a}_i, \sigma_i) = \frac{1}{(\pi\sigma_i^2)^{M-1}} \exp\left(-\frac{||\mathbf{x} - (\mathbf{a}_i^H\mathbf{x})\cdot\mathbf{a}_i||^2}{\sigma_i^2}\right)$$

$\mathbf{a}_i$: mean vector (unit-norm)

σ_i^2 : variance

$||\mathbf{x} - (\mathbf{a}_i^H\mathbf{x})\cdot\mathbf{a}_i||$

Distance to the subspace spanned by $\mathbf{a}_i$



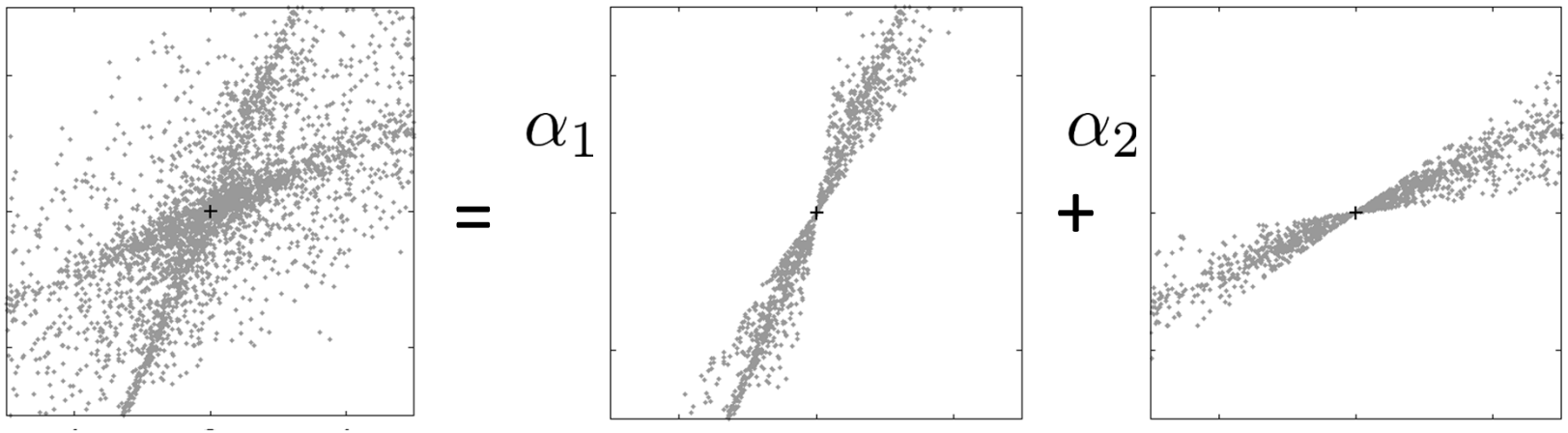
Gaussian mixture model

- Mixtures of several sparse components
  - Mixture model of Gaussians

$$p(\mathbf{x}|\theta) = \sum_{i=1}^{N} \alpha_i \, p(\mathbf{x}|\mathbf{a}_i, \sigma_i)$$

$$\theta = \{\mathbf{a}_1, \sigma_1, \alpha_1, \ldots, \mathbf{a}_N, \sigma_N, \alpha_N\}$$ Parameter set

$p(\mathbf{x}|\theta)$ $p(\mathbf{x}|\mathbf{a}_1, \sigma_1)$ $p(\mathbf{x}|\mathbf{a}_2, \sigma_2)$

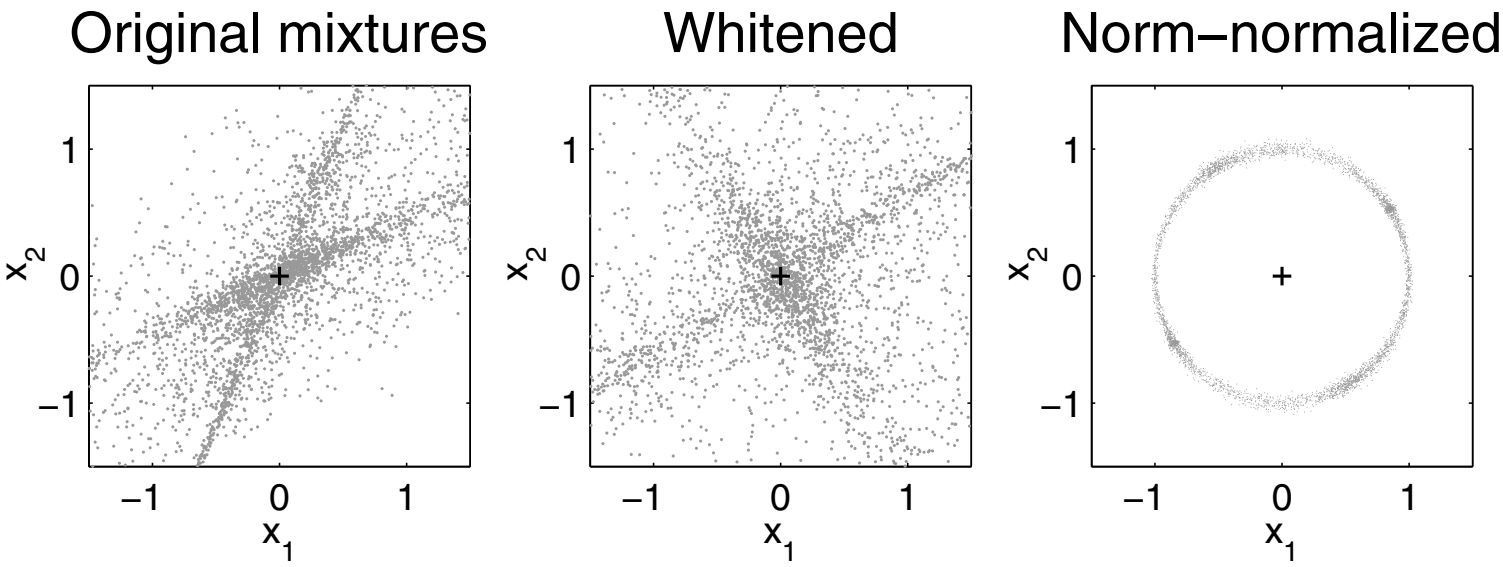


Some practical issues

- Pre-whitening
  - Effective for robust clustering
- Unit-norm normalization

$$\mathbf{x}(\tau) \leftarrow \frac{\mathbf{x}(\tau)}{||\mathbf{x}(\tau)||}$$

  - Otherwise, samples close to the origin have unreasonably large likelihoods



ML estimation by EM algorithm

- E-step (Posterior probability)

$$P(C_i|\mathbf{x}) = \frac{\alpha_i \, p(\mathbf{x}|\mathbf{a}_i, \sigma_i)}{\sum_{k=1}^{I} \alpha_k \, p(\mathbf{x}|\mathbf{a}_k, \sigma_k)}$$

- M-step (Parameter optimizations)
  - $\mathbf{a}_i$ is given by the eigenvector corresponding to the maximum eigenvalue of

$$\mathbf{R} = \sum_{n=1}^{N} P(C_i|\mathbf{x}(n)) \cdot \mathbf{x}(n)\mathbf{x}^H(n)$$

  - Other parameters updates

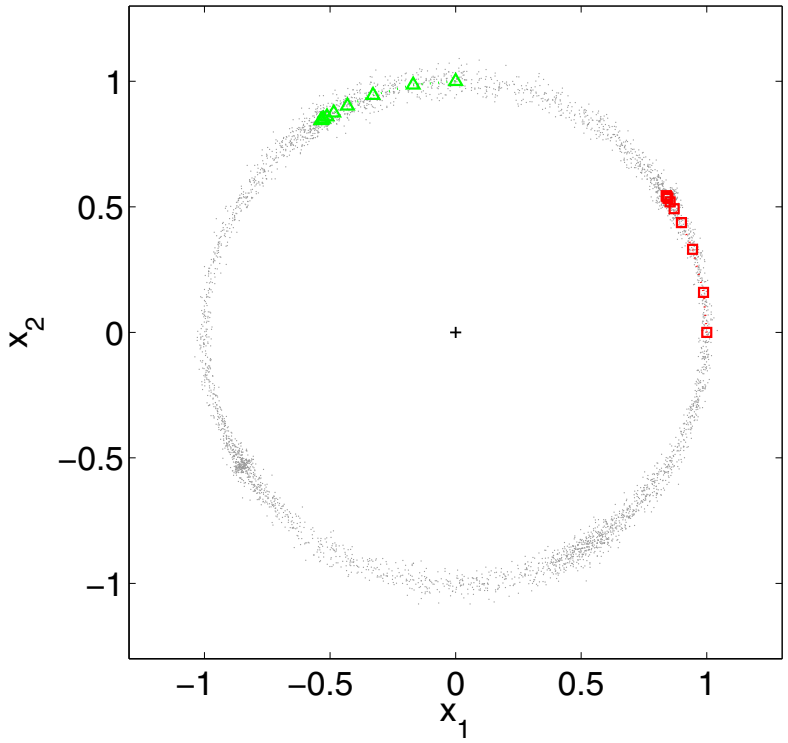
$$\sigma_i^2 = \frac{\sum_{n=1}^{N} P(C_i|\mathbf{x}(n)) \cdot ||\mathbf{x}(n) - (\mathbf{a}_i^H \mathbf{x}(n)) \cdot \mathbf{a}_i||^2}{(J-1) \cdot \sum_{n=1}^{N} P(C_i|\mathbf{x}(n))}$$

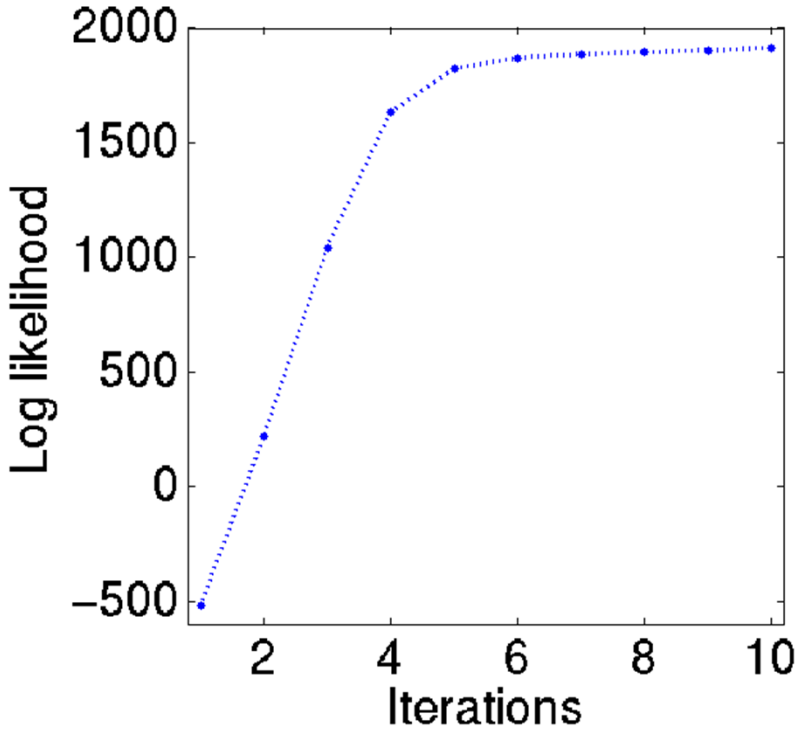
$$\alpha_i = \frac{1}{N} \sum_{n=1}^{N} P(C_i|\mathbf{x}(n))$$

GMM convergence example

Red (□) starting from $\mathbf{a}_1 = [1 \;\; 0]^T$

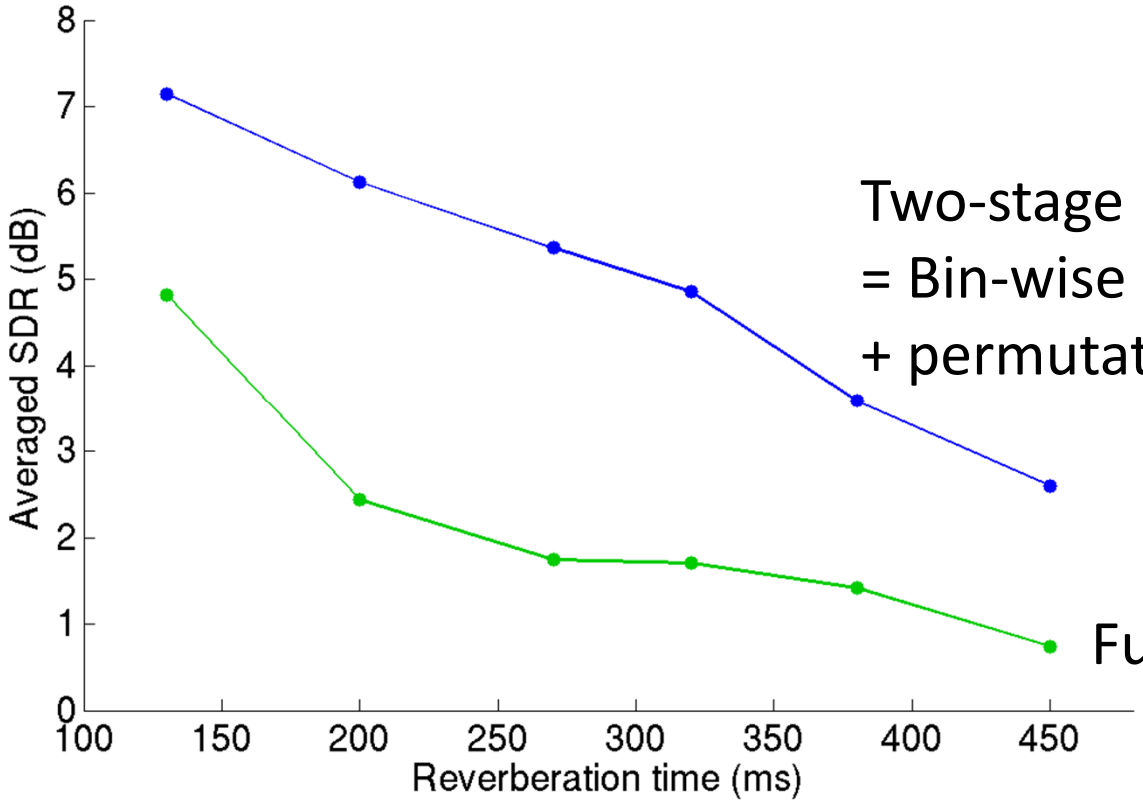
Green (△) starting from $\mathbf{a}_2 = [0 \;\; 1]^T$





Experimental comparison

- 3 microphones, 4 sources (underdetermined case)
- Source: speeches of 6 seconds
- Averaged SDR (signal-to-distortion ratio) over 8 combinations



Section 2.2 Summary

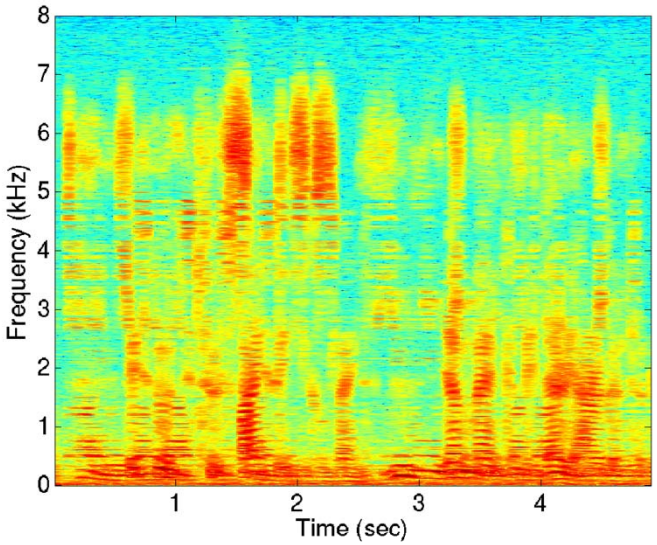
- SCA: sparse component analysis
- Full-band TDOA-based SCA
  - No need for permutation alignment
- Frequency bin-wise SCA
  - More precise frequency-dependent analysis
    - Effective in reverberant conditions
  - Need for permutation alignment
    - Perform SCA for, e.g. 513, times separately

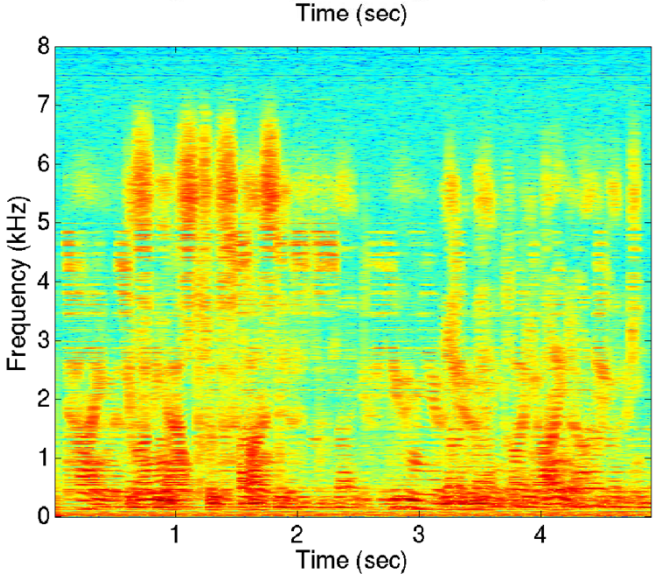
Outline of Part II

1. Overview of frequency domain approach
2. Frequency bin-wise separation
    1. ICA: independent component analysis
    2. SCA: sparse component analysis
3. Permutation alignment
    1. Activity sequence clustering
    2. TDOA, DOA clustering
4. Concluding remarks

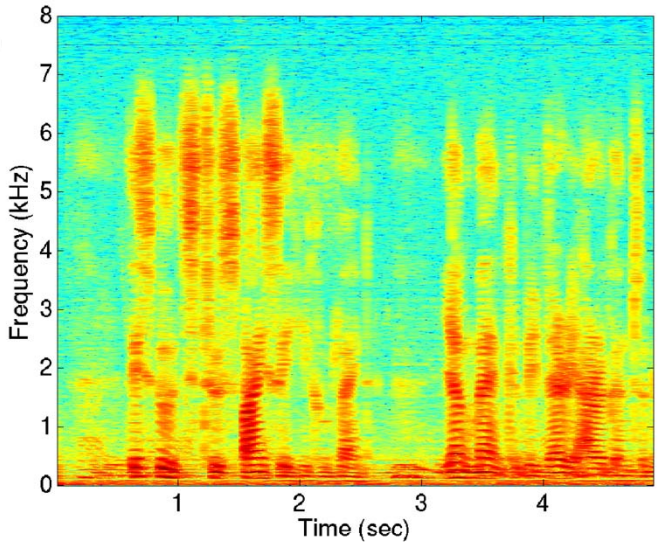
Big problem: permutation alignment

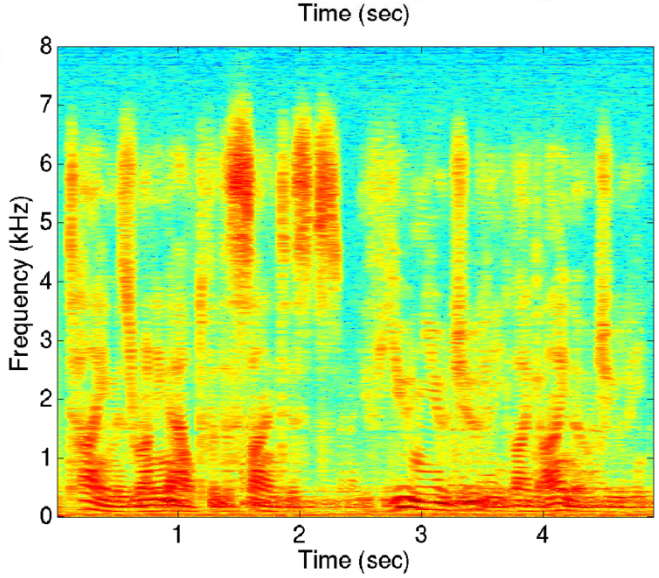
Raw ICA/SCA outputs





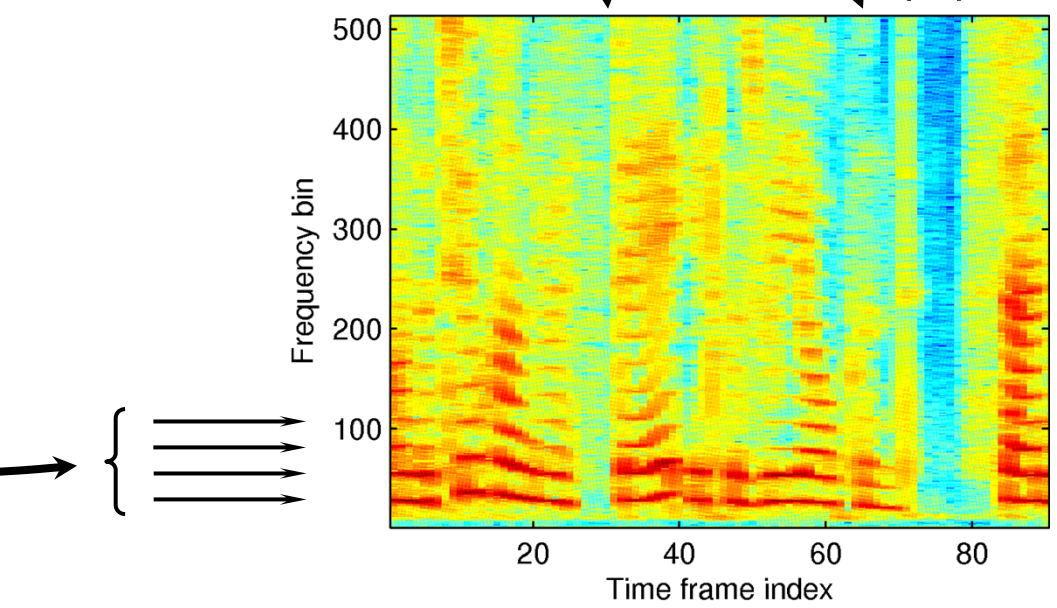
Permutation aligned





Dependence across frequencies

- Meaningful audio source has some structures
  - Common silence period
  - Common onset and offset
  - Harmonics



➡ Mutual dependence of separated signals across frequencies

Correlation coefficient

- Correlation coefficient between two sequences

$$\mathrm{cor}(v_i, v_j) = \frac{\mathrm{E}\{(v_i - \mu_i)(v_j - \mu_j)\}}{\sigma_i \sigma_j}$$

  - mean $\mu_i = \mathrm{E}\{v_i\}$
  - variance $\sigma_i^2 = \mathrm{E}\{v_i^2\} - \mu_i^2$

- Bounded by

$$-1 \leq \mathrm{cor}(v_i, v_j) \leq 1$$

  - becomes 1 if two sequences are identical

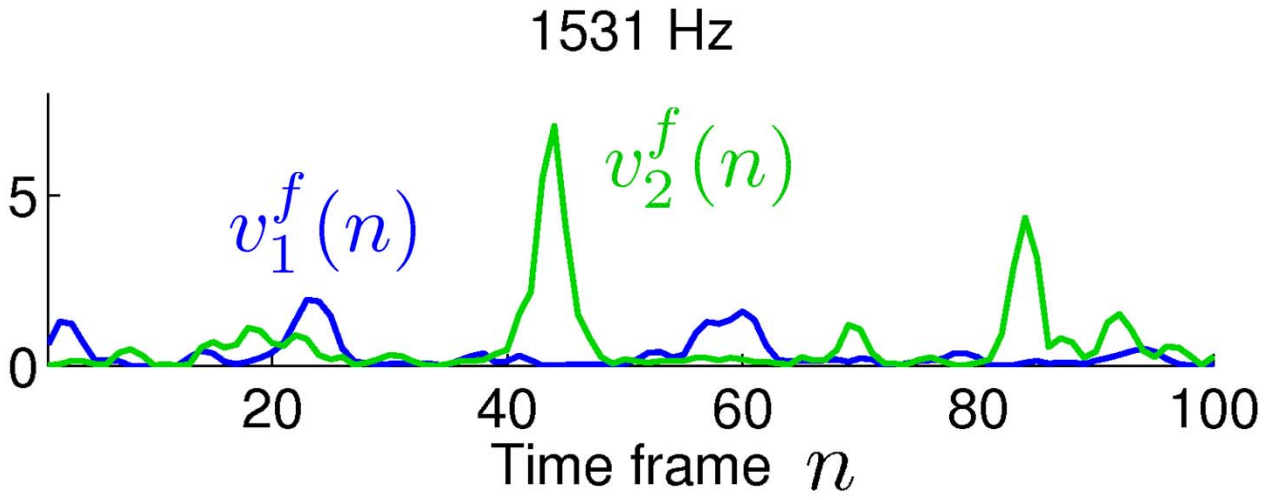
Envelopes of separated signals

- Envelope of bin-wise separated signal

$$v_i^f(n) = |y_i(n, f)|$$

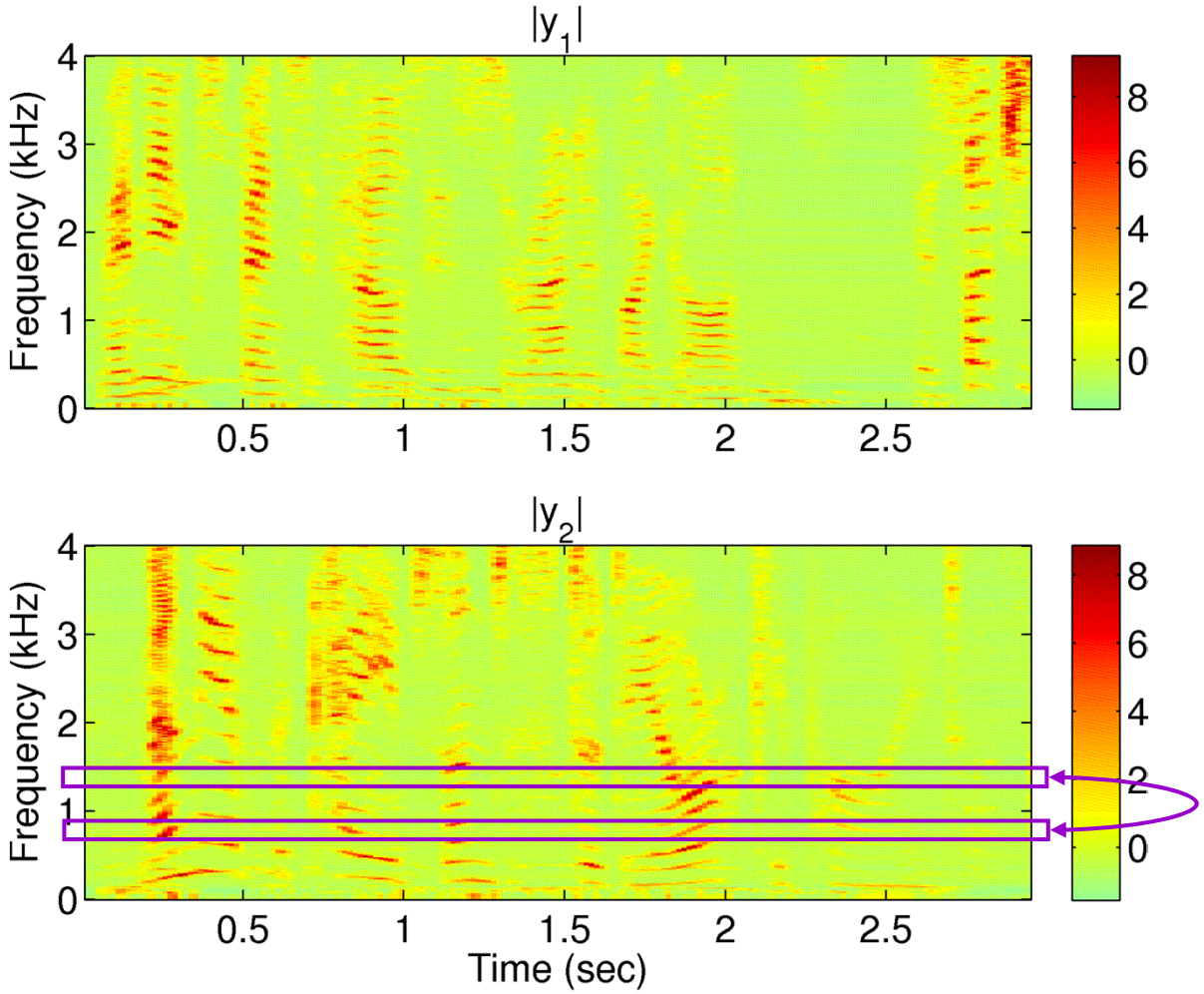
  - At frequency f and at output i

- Shows the signal activity at the frequency



Envelope examples

Two separated signals

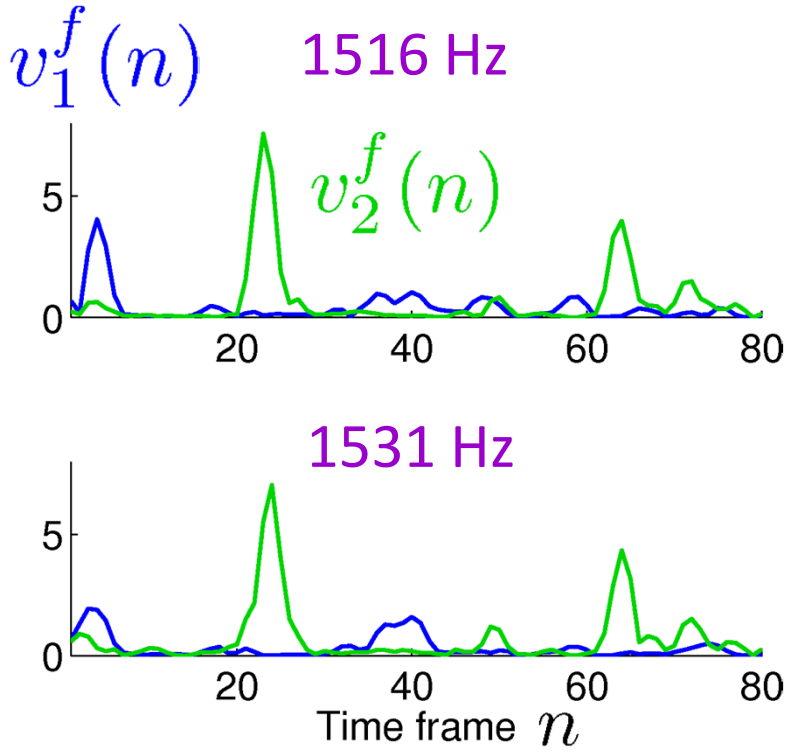


Normalized to zero-mean and unit-norm

High correlations are expected for the same source

Neighboring frequencies

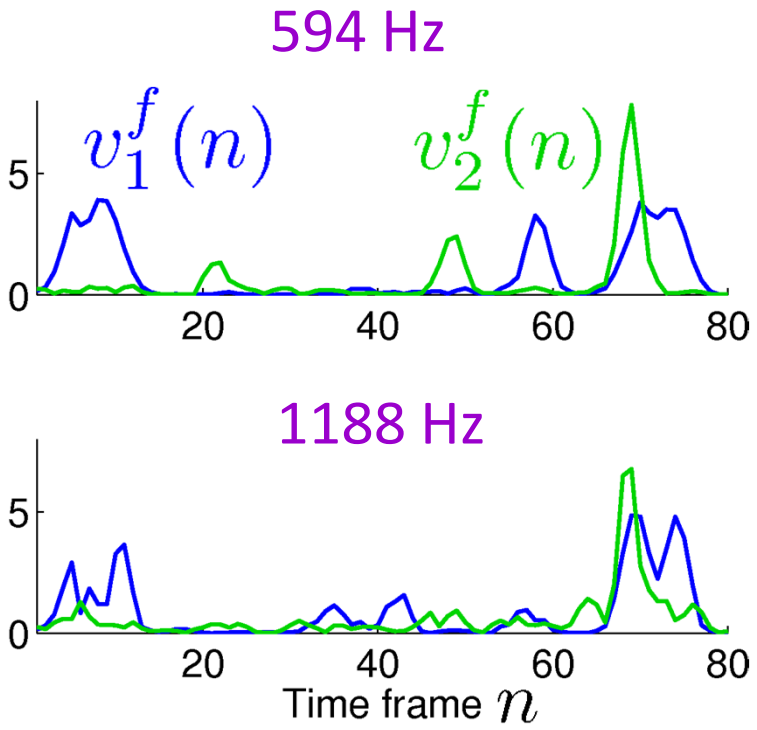
- Envelopes of neighboring frequencies are highly correlated
- A simple strategy for permutation alignment
  - Maximize correlation between neighbors
  - diagonalize



$$\begin{bmatrix} \mathrm{cor}(v_1^f, v_1^{f+1}) & \mathrm{cor}(v_1^f, v_2^{f+1}) \\ \mathrm{cor}(v_2^f, v_1^{f+1}) & \mathrm{cor}(v_2^f, v_2^{f+1}) \end{bmatrix} = \begin{bmatrix} 0.78 & -0.12 \\ -0.15 & 0.92 \end{bmatrix}$$

Harmonic frequencies

- High correlation among fundamental frequency f and its harmonics $2f, 3f, \ldots$
- Another strategy for permutation alignment
  - Maximize correlation among harmonics
  - diagonalize



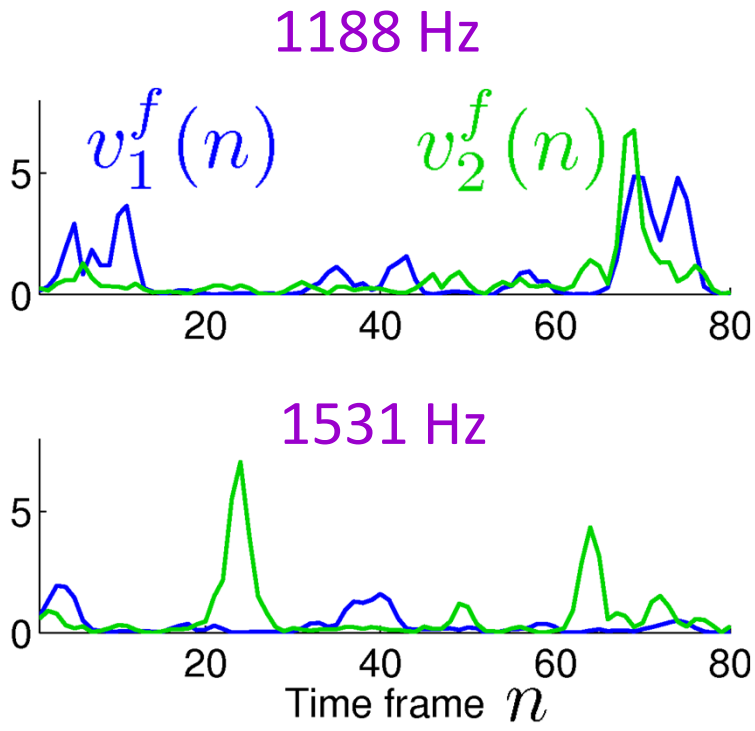
$$\begin{bmatrix} \mathrm{cor}(v_1^f, v_1^{2f}) & \mathrm{cor}(v_1^f, v_2^{2f}) \\ \mathrm{cor}(v_2^f, v_1^{2f}) & \mathrm{cor}(v_2^f, v_2^{2f}) \end{bmatrix} = \begin{bmatrix} 0.76 & 0.36 \\ 0.48 & 0.89 \end{bmatrix}$$

Arbitrary pairs of frequencies

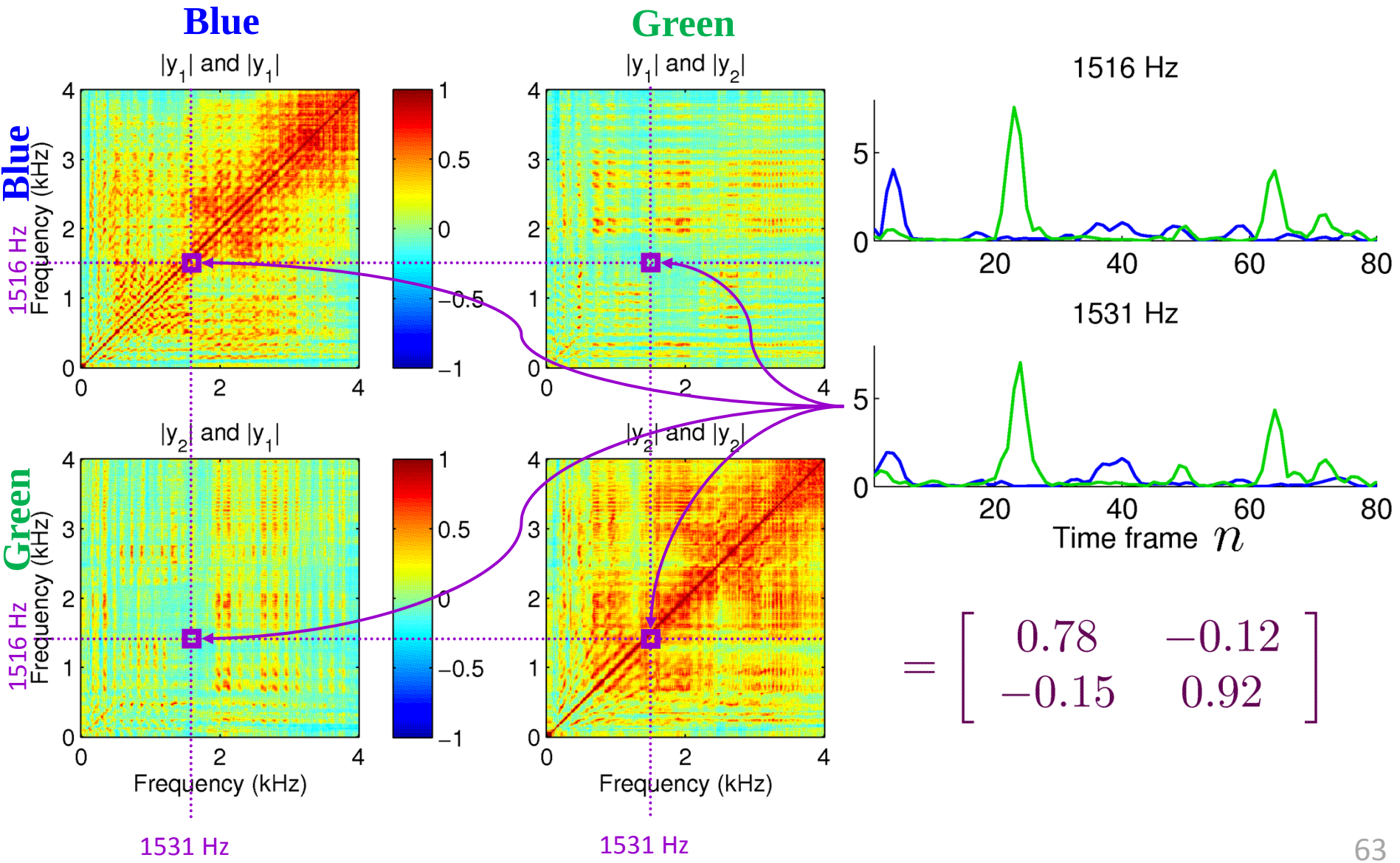
- Among frequencies that have no specific relation
  - May end up with almost zero correlation

$$\begin{bmatrix} \mathrm{cor}(v_1^f, v_1^g) & \mathrm{cor}(v_1^f, v_2^g) \\ \mathrm{cor}(v_2^f, v_1^g) & \mathrm{cor}(v_2^f, v_2^g) \end{bmatrix}$$

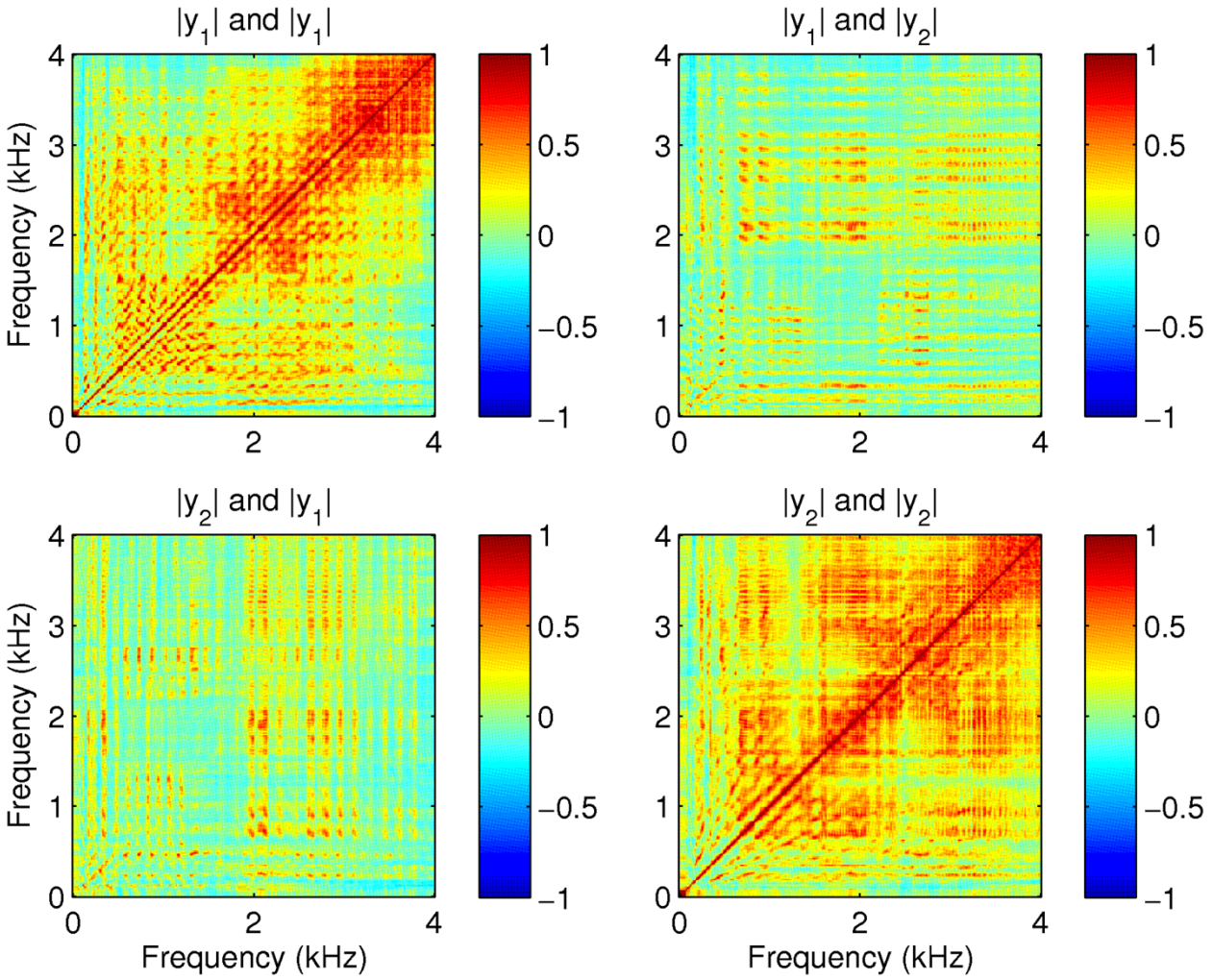
$$= \begin{bmatrix} 0.10 & -0.14 \\ -0.11 & 0.06 \end{bmatrix}$$



Correlation of envelopes: global view



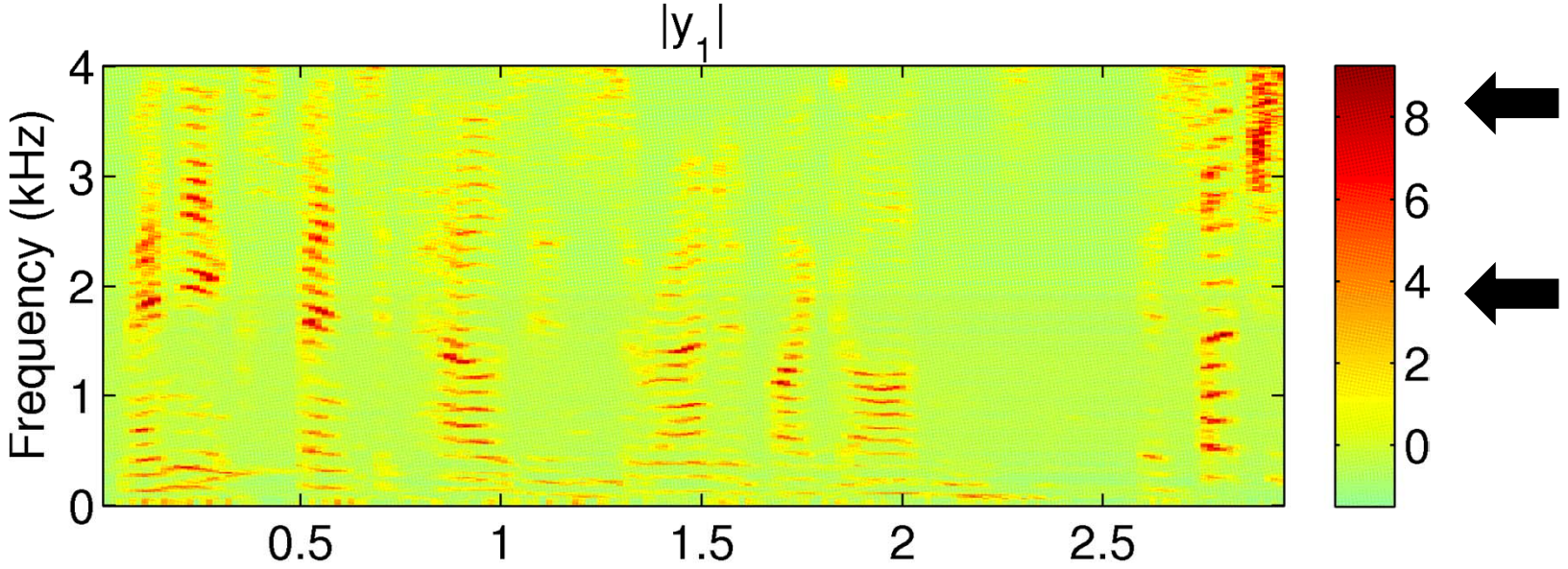
Correlation of envelopes: global view



1. High correlation only with adjacent or harmonic frequencies for the same source
2. Mostly zero correlation with different sources

Why high correlations only within limited pairs?

- Envelopes have a wide dynamic range even if they are normalized to zero-mean and unit-norm
  - Active signals are represented with various values



Dominance measure

- Another measure for signal activity

$$0 \leq v_i^f(n) = domMeasure_i(n, f) \leq 1$$

inactive active

  - High correlations expected among many frequencies

- More specifically

  - ICA: power ratio

$$v_i^f(n) = powRatio_i(n, f) = \frac{||\mathbf{a}_i(f) y_i(n, f)||^2}{\sum_{k=1}^{I} ||\mathbf{a}_k(f) y_k(n, f)||^2}$$

  - SCA: posterior probability

$$v_i^f(n) = P(C_i|\mathbf{x}(n, f)) = \frac{\alpha_i \, p(\mathbf{x}(n, f)|\mathbf{a}_i, \sigma_i)}{\sum_{k=1}^{I} \alpha_k \, p(\mathbf{x}(n, f)|\mathbf{a}_k, \sigma_k)}$$

Power ratio (ICA)

- Power ratio of i-th separated components to all the separated components at a time-frequency slot

$$v_i^f(n) = powRatio_i(n,f) = \frac{||\mathbf{a}_i(f)y_i(n,f)||^2}{\sum_{k=1}^{I}||\mathbf{a}_k(f)y_k(n,f)||^2}$$

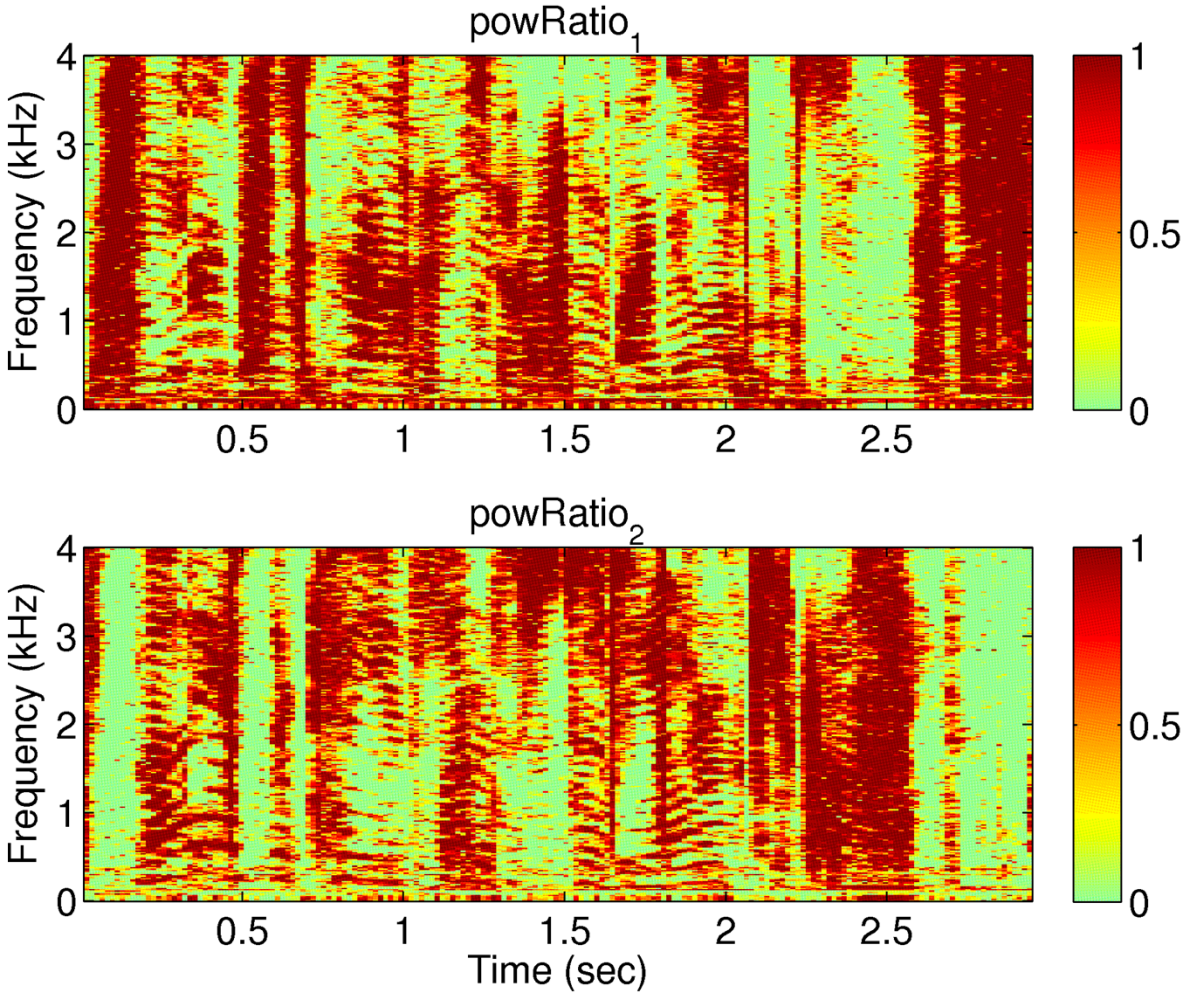
- Mixing system estimation should be conducted after ICA

$$\mathbf{x} = \sum_{i=1}^{I}\mathbf{a}_i y_i \quad \Leftarrow \quad \mathbf{y} = \mathbf{W}\mathbf{x}$$

Estimated mixing situation ICA solution

powRatio value example

Two separated signals (permutations are aligned)



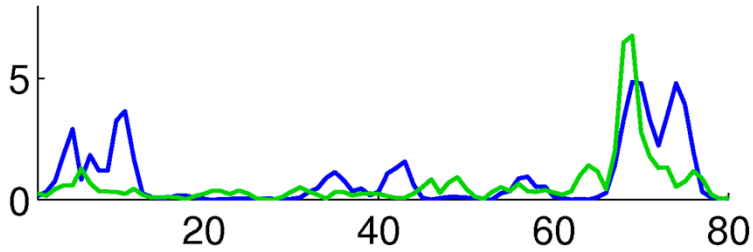
1. Active signal uniformly close to 1
2. Exclusive: if one is close to 1, then the other is close to 0

Comparison: envelope and powRatio

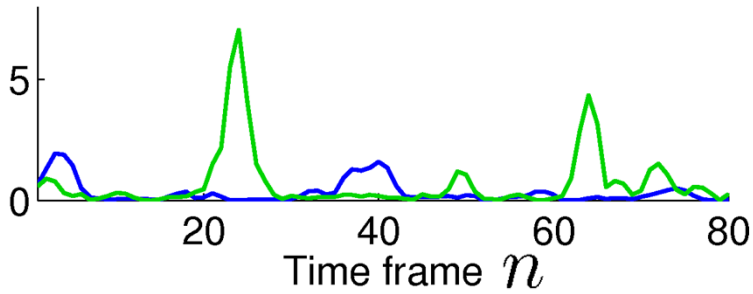
Envelope

$$|y_i(n, f)|$$

1188 Hz



1531 Hz

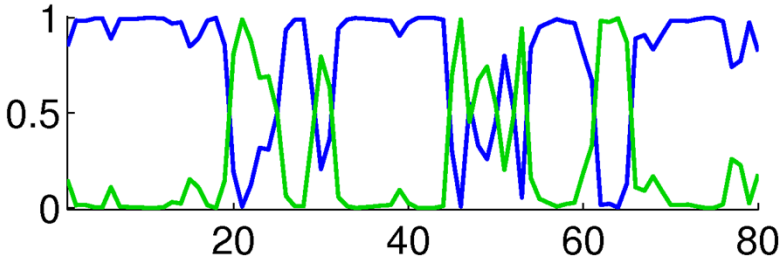


$$= \begin{bmatrix} 0.10 & -0.14 \\ -0.11 & 0.06 \end{bmatrix}$$

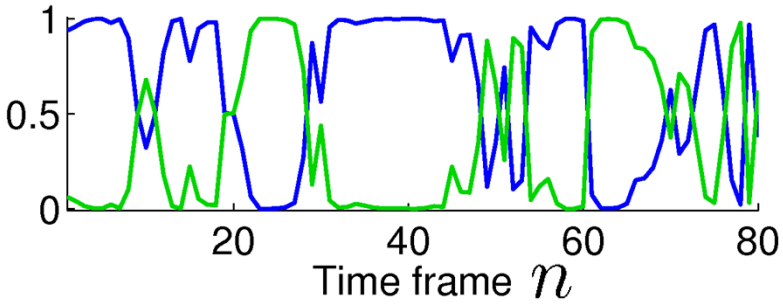
Power ratio

$$powRatio_i(n, f)$$

1188 Hz



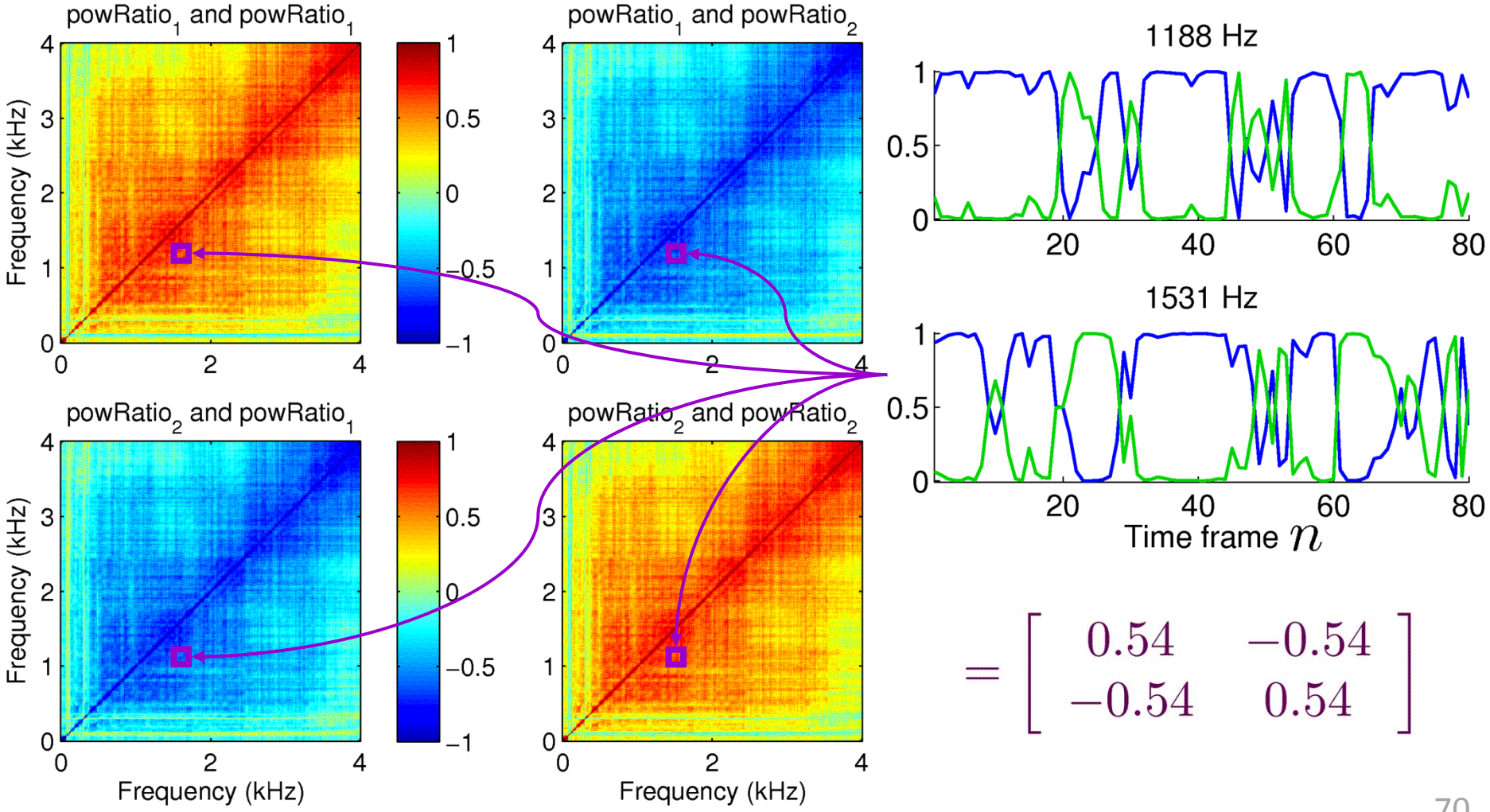
1531 Hz



$$= \begin{bmatrix} 0.54 & -0.54 \\ -0.54 & 0.54 \end{bmatrix}$$

Correlation of powRatio: global view

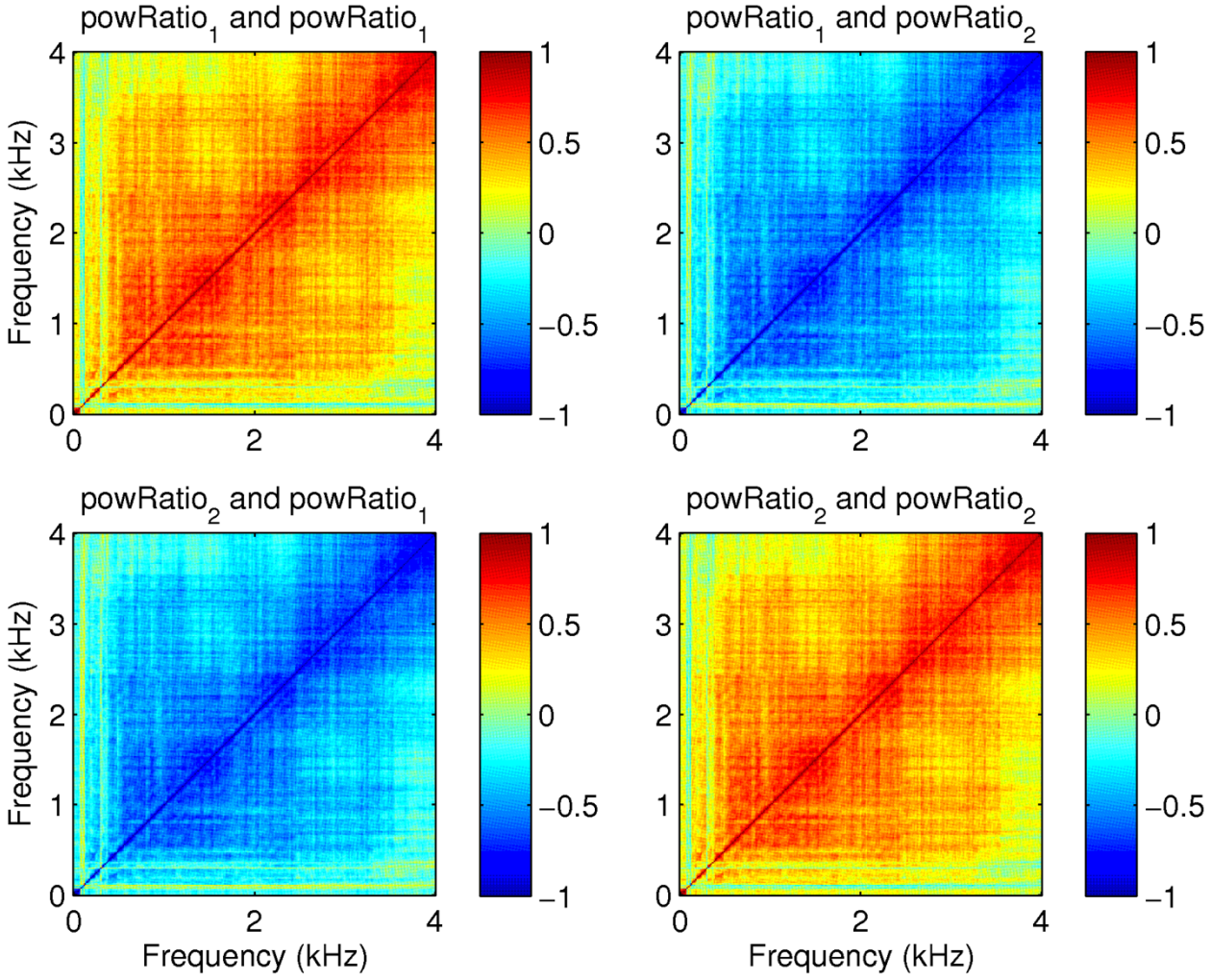
$$v_i^f(n) = powRatio_i(n, f)$$



$$= \begin{bmatrix} 0.54 & -0.54 \\ -0.54 & 0.54 \end{bmatrix}$$

Correlation of powRatio: global view

$$v_i^f(n) = powRatio_i(n, f)$$



1. High correlation between many frequencies for the same source
2. Negative correlation for different sources

Posterior probability (SCA)

- Posterior probability for i-th cluster at a time-frequency slot

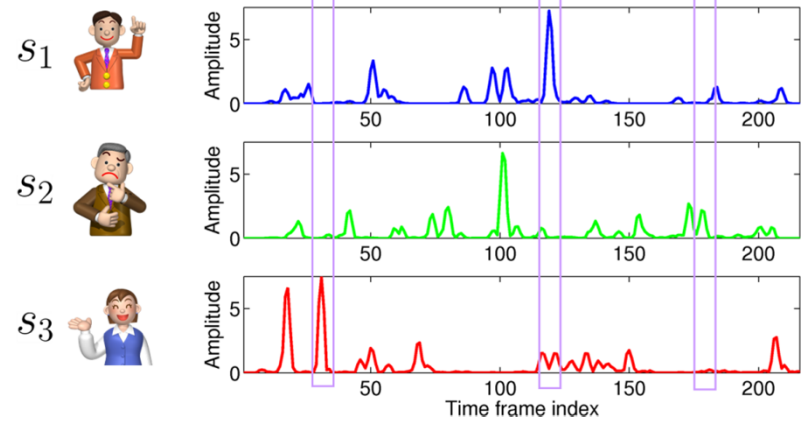
$$v_i^f(n) = P(C_i|\mathbf{x}(n,f)) = \frac{\alpha_i\, p(\mathbf{x}(n,f)|\mathbf{a}_i,\sigma_i)}{\sum_{k=1}^{I}\alpha_k\, p(\mathbf{x}(n,f)|\mathbf{a}_k,\sigma_k)}$$

  - Calculated in the E-step of the GMM-based clustering algorithm

$P(C_1|\mathbf{x})$ 0.01 0.85 0.11

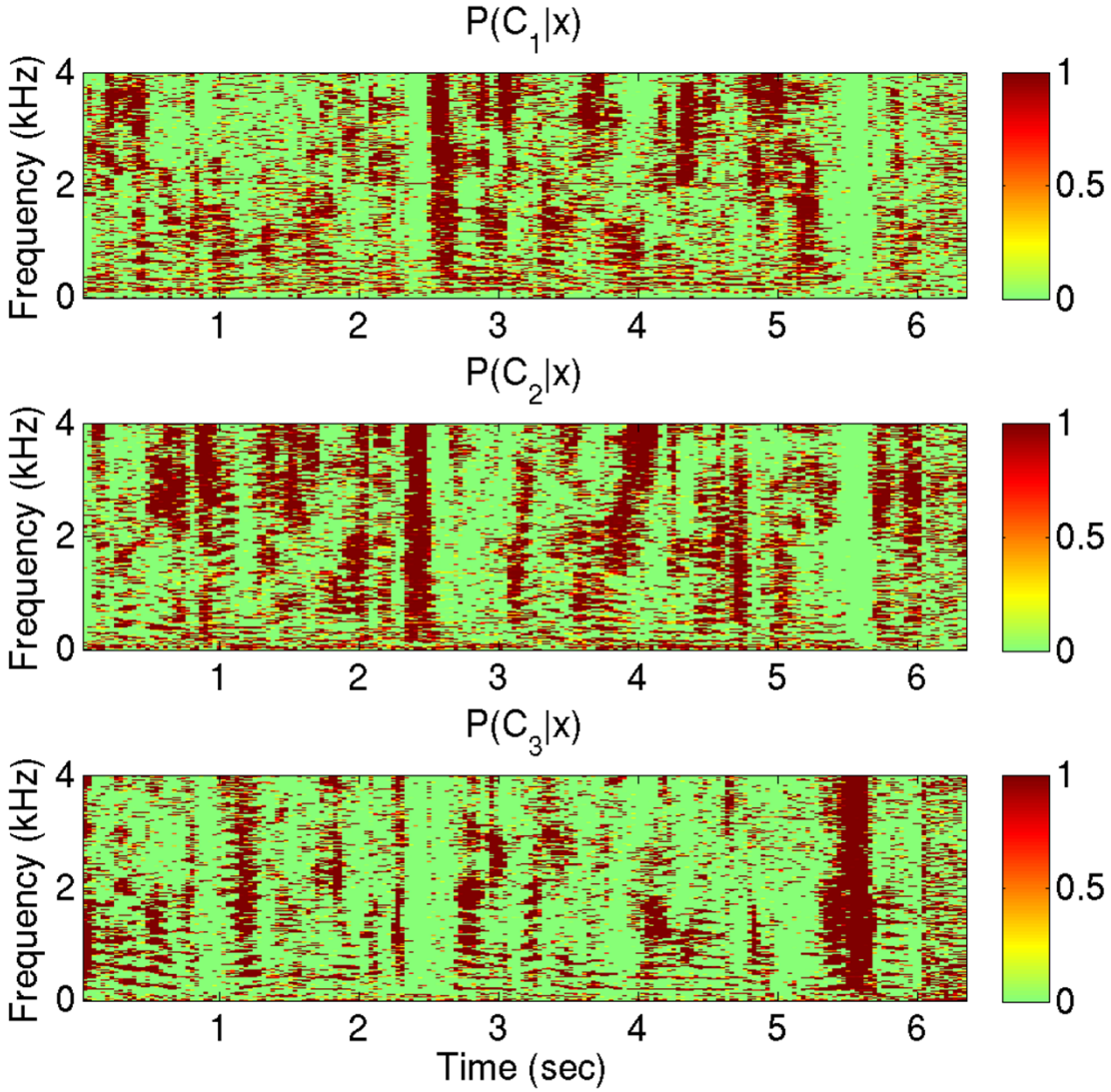
$P(C_2|\mathbf{x})$ 0.04 0.05 0.88

$P(C_3|\mathbf{x})$ 0.95 0.10 0.01



Posterior probability example

Three separated signals (permutations are aligned)

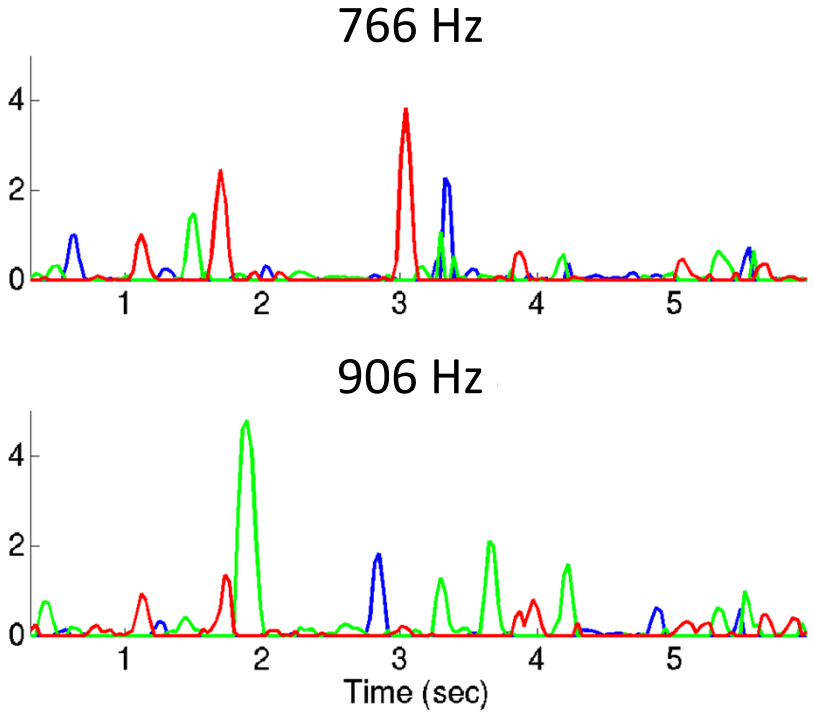


1. Active signal uniformly close to 1
2. Exclusive: if one is close to 1, then the others are close to 0

Comparison: envelope and posterior probability

Envelope

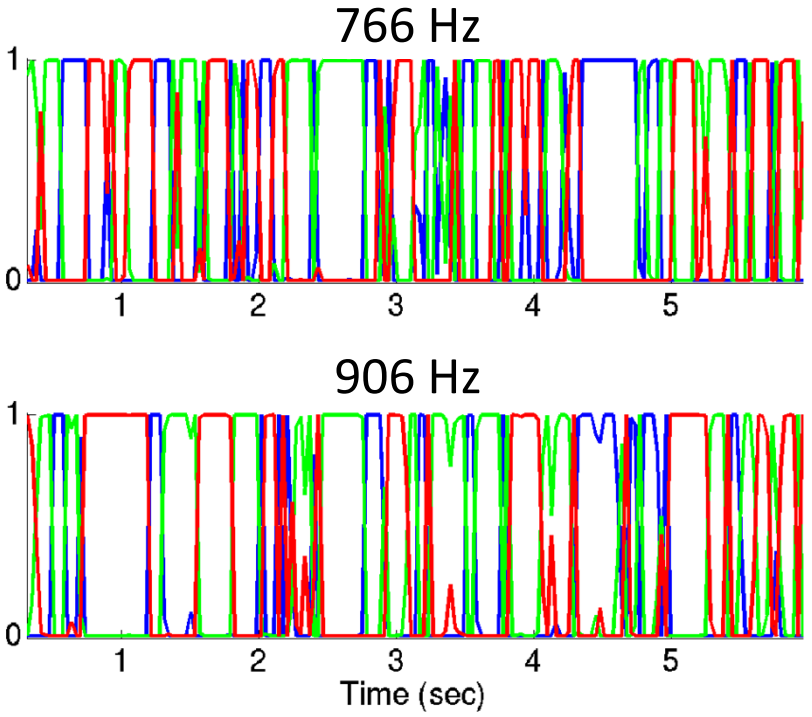
$$|y_i(n,f)|$$



$$= \begin{bmatrix} 0.01 & 0.07 & -0.11 \\ -0.09 & 0.11 & -0.15 \\ -0.07 & -0.07 & 0.44 \end{bmatrix}$$

Posterior probability

$$P(C_i|\mathbf{x}(n,f))$$



$$= \begin{bmatrix} 0.51 & -0.20 & -0.29 \\ -0.25 & 0.46 & -0.23 \\ -0.28 & -0.28 & 0.55 \end{bmatrix}$$

Strategies for permutation optimization

- Local optimization
  - Among neighboring or harmonic frequencies

$$\Pi_f = \arg\max_{\Pi} \sum_{k=1}^{I} \mathrm{cor}(v_i^f, v_k^{f+1})\big|_{i=\Pi(k)}$$

    - Effective for fine tuning
- Global optimization
  - Similar to k-means clustering, where centroids c_k are explicitly identified

$$\Pi_f = \arg\max_{\Pi} \sum_{k=1}^{I} \mathrm{cor}(v_i^f, c_k)\big|_{i=\Pi(k)}$$

    - Efficient and robust, but applicable only if high correlations within many frequency pairs

Global optimization

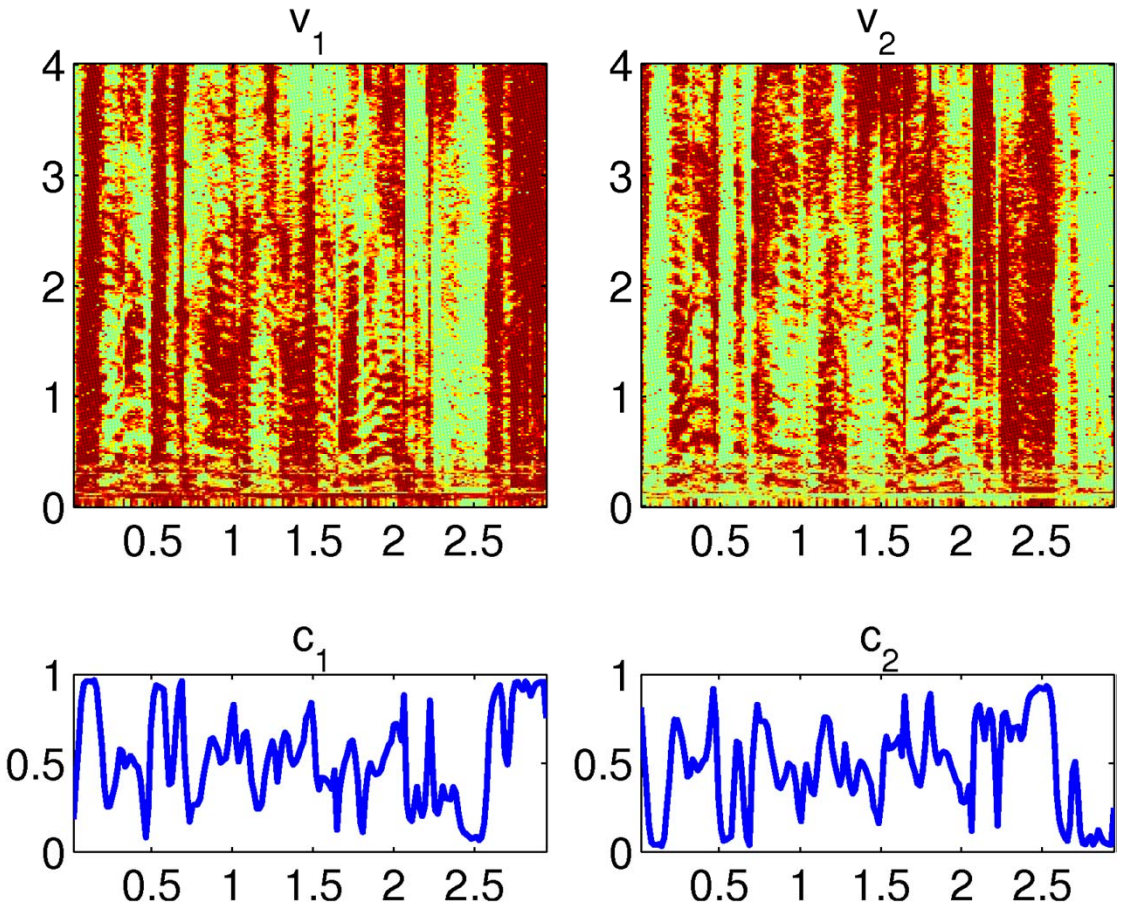
Optimization algorithm similar to k-means clustering

Permutation Π_f aligned to maximize

$$\sum_{k=1}^{I} \mathrm{cor}(v_i^f, c_k)\big|_{i=\Pi_f(k)}$$

Centroid $c_k(n)$ Member average

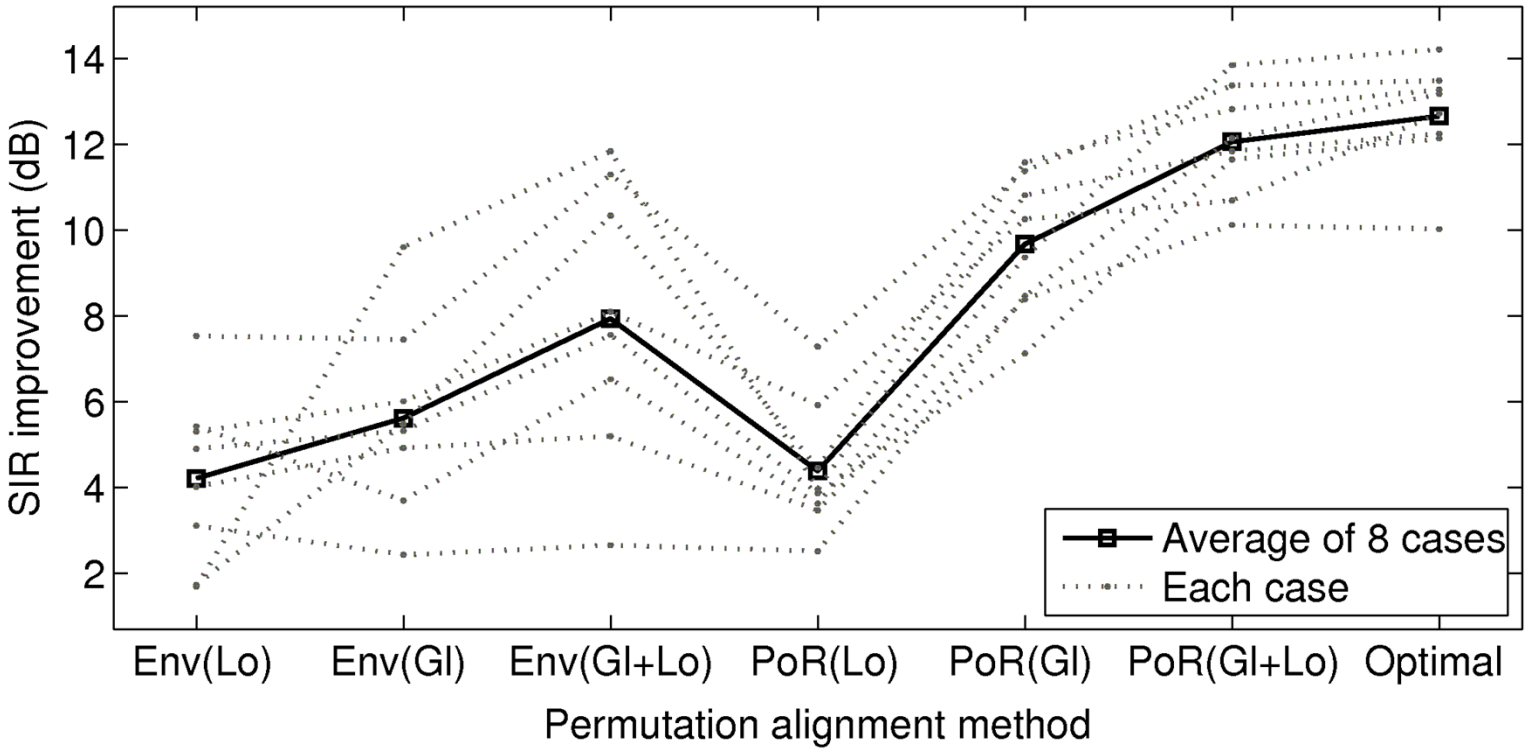
$$\frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} v_i^f(n)\big|_{i=\Pi_f(k)}$$



Represents the signal activity for each source

Experimental comparison (ICA)

3 sources, 3 microphones

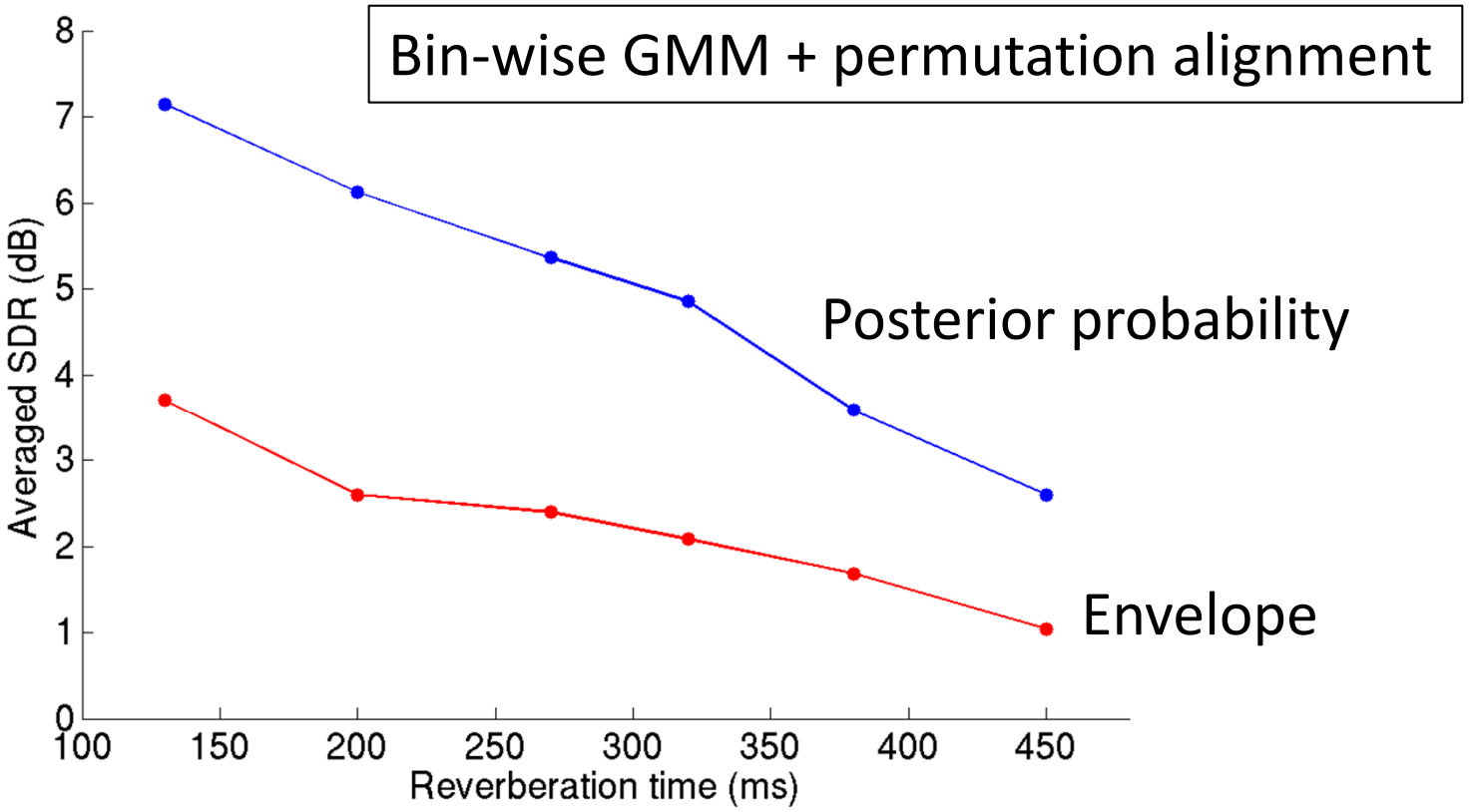


Env: envelope, PoR: powRatio, Lo: local, Gl: global

Global optimization with powRatio works well.
Subsequent local optimization improves the results further.

Experimental comparison (SCA)

- 3 microphones, 4 sources (underdetermined case)
- Source: speeches of 6 seconds
- Averaged SDR (signal-to-distortion ratio) over 8 combinations



Section 3.1 Summary

- Permutation alignment by clustering activity sequences
  - Similarity measure: correlation coefficient
    - (inner product after zero-mean and unit-norm normalization)
- Examined activity sequences
  - Envelope (ICA, SCA)
  - Power ratio (ICA)
  - Posterior probability (SCA)

Outline of Part II

1. Overview of frequency domain approach
2. Frequency bin-wise separation
    1. ICA: independent component analysis
    2. SCA: sparse component analysis
3. Permutation alignment
    1. Activity sequence clustering
    2. TDOA, DOA clustering
4. Concluding remarks

TDOA, DOA-based permutation alignment

- Very similar to Full-band TDOA-based SCA
- Clustering
  - TDOA: Time-Difference-Of Arrival
  - DOA: Direction-Of-Arrival
    - Estimated from ICA / SCA result
- Calculated from estimated mixing vector $\mathbf{a}_i$
  - ICA: $\mathbf{y} = \mathbf{W}\mathbf{x}$ inverse $\mathbf{x} = \sum_{i=1}^{I} \mathbf{a}_i y_i$
  - SCA (GMM): $p(\mathbf{x}|\theta) = \sum_{i=1}^{N} \alpha_i \, p(\mathbf{x}|\mathbf{a}_i, \sigma_i)$

$$\theta = \{\mathbf{a}_1, \sigma_1, \alpha_1, \ldots, \mathbf{a}_N, \sigma_N, \alpha_N\}$$

TDOA calculated from mixing vector

- Estimated mixing vector at frequency f for output i

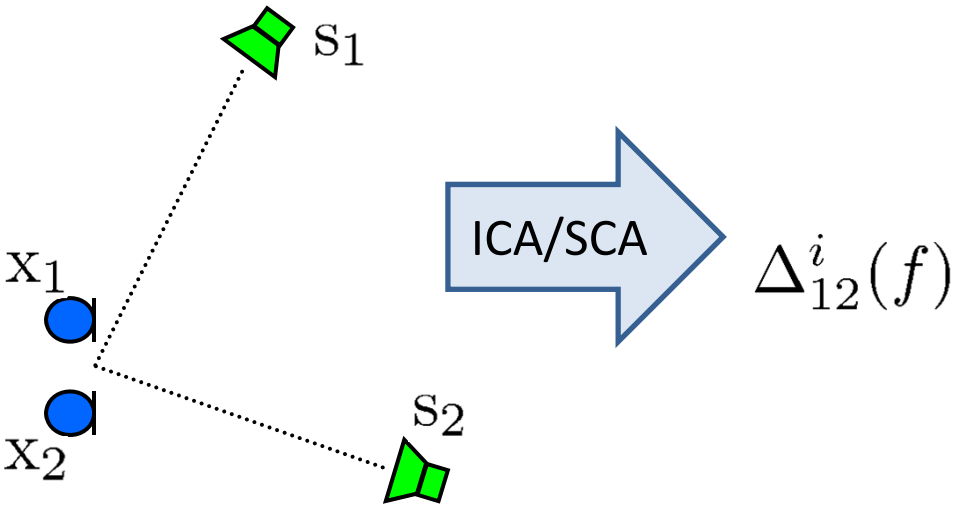
$$\mathbf{a}_i(f) = \begin{bmatrix} a_{1i}(f) \\ a_{2i}(f) \end{bmatrix}$$

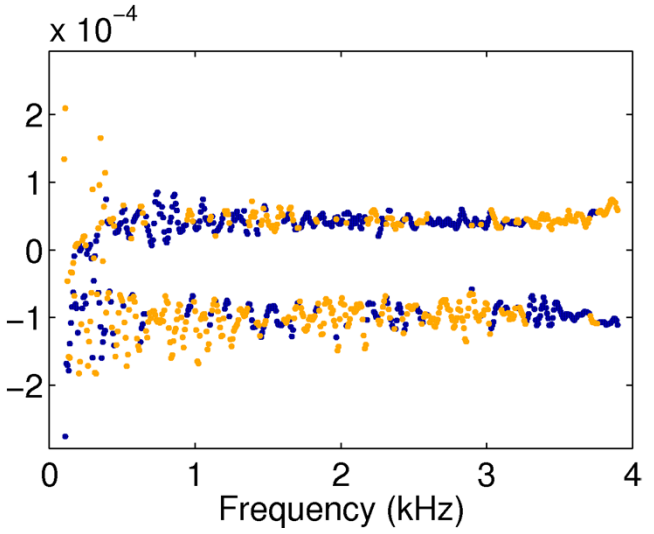
- Frequency dependent TDOA from mixing vector

$$\Delta_{12}^{i}(f) = \frac{\arg[a_{1i}(f)/a_{2i}(f)]}{-2\pi f}$$

Similar formula to TDOA estimation from time-frequency observations

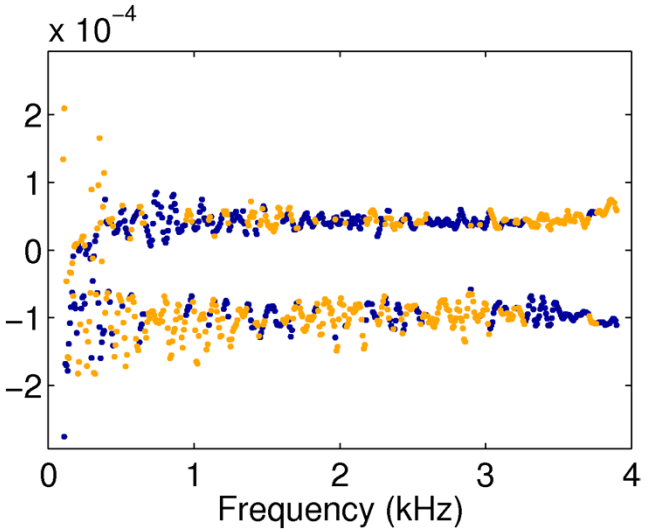
$$\Delta_{12}(n,f) = \frac{\arg[x_1(n,f)/x_2(n,f)]}{-2\pi f}$$



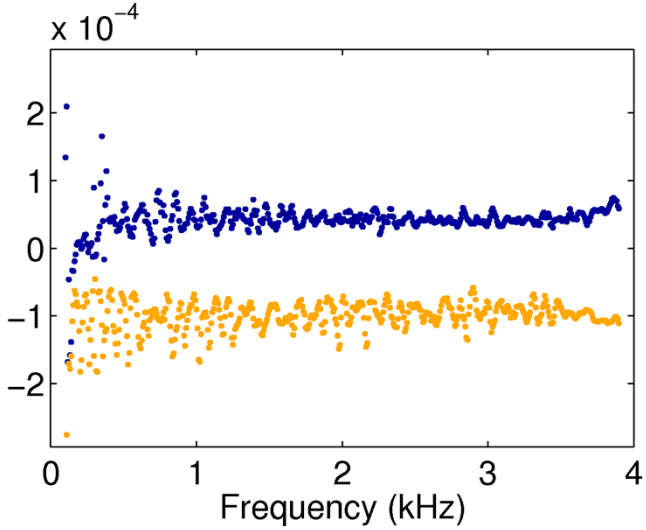


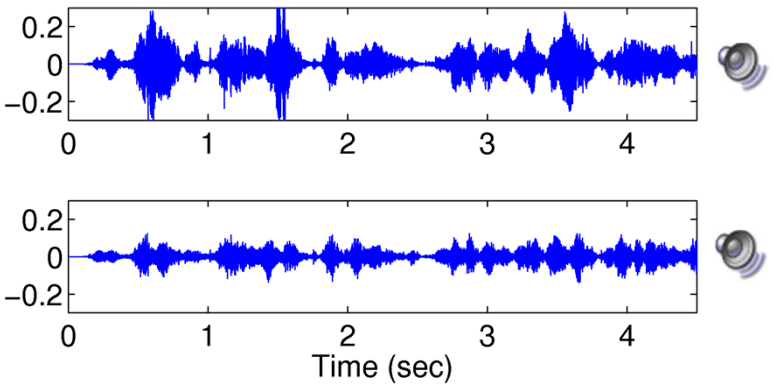
Permutation alignment

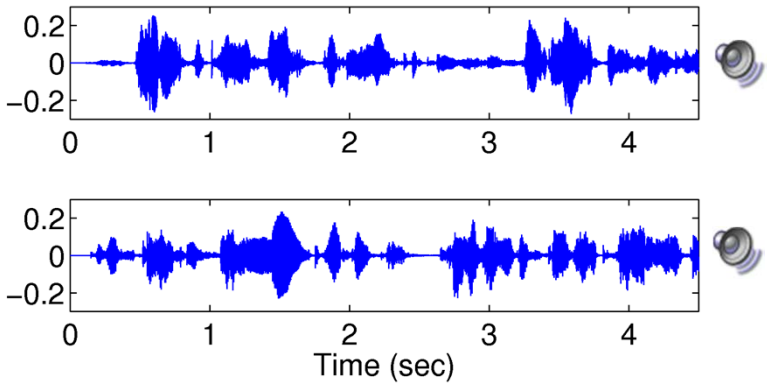
TDOA estimations calculated from estimated mixing vectors



Sorting
Clustering

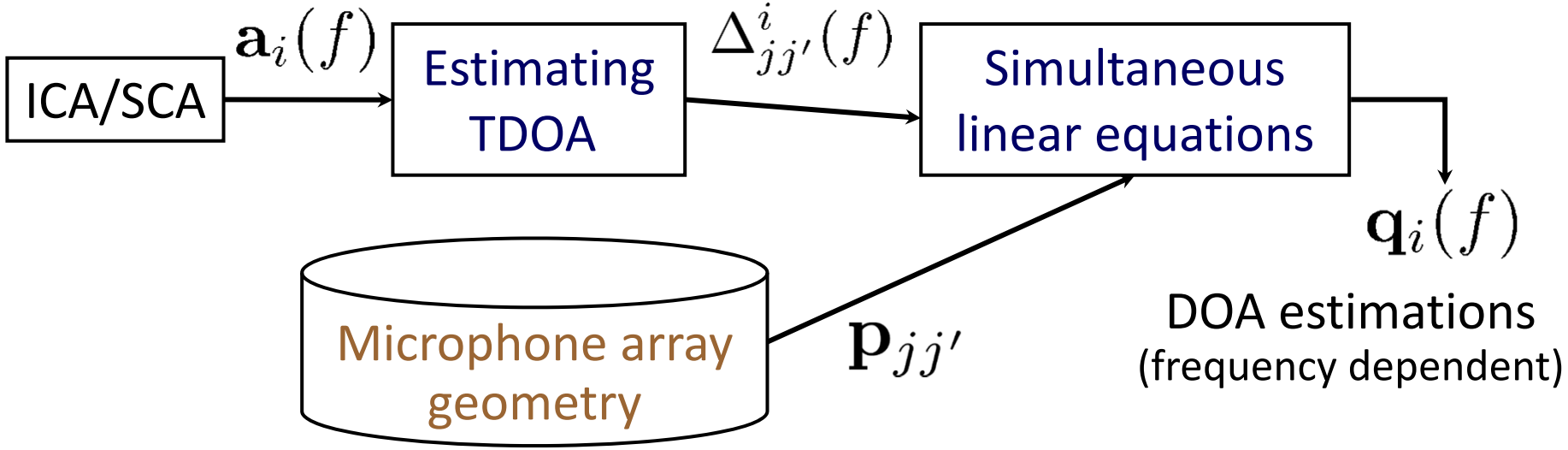






Estimating DOAs of sources

- DOA: Direction Of Arrival
- Useful for e.g. camera steering
- By additional operation after estimating TDOAs

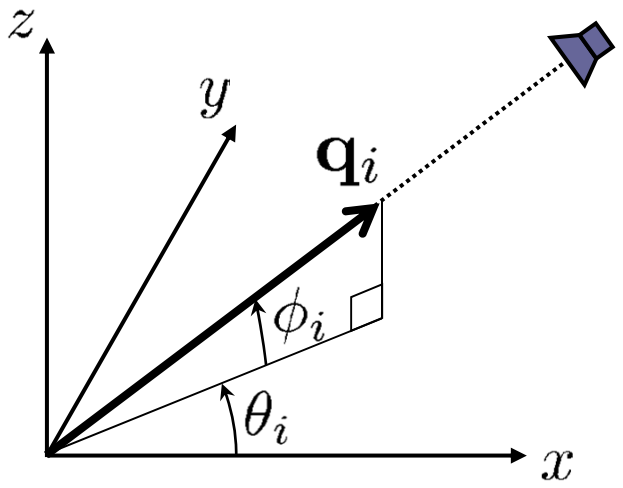


Needs to know: semi-blind

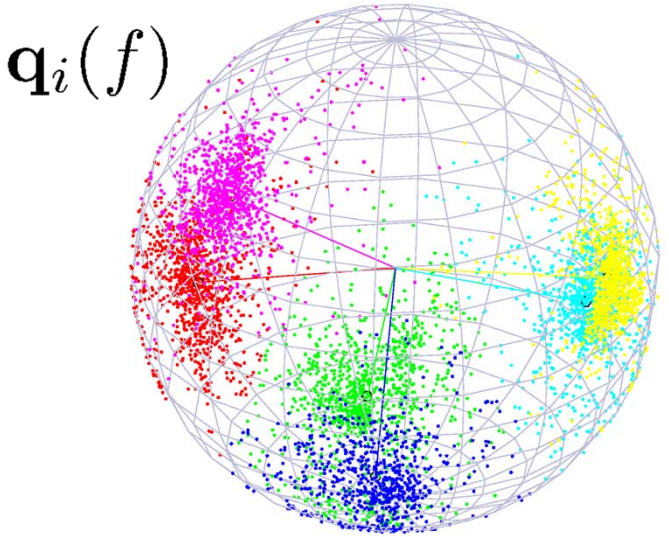
DOA: definition and example

3-dim unit-norm vector

$$\mathbf{q}_i = \begin{bmatrix} \cos\theta_i \cos\phi_i \\ \sin\theta_i \cos\phi_i \\ \sin\phi_i \end{bmatrix}$$



6 speakers and 8 microphones



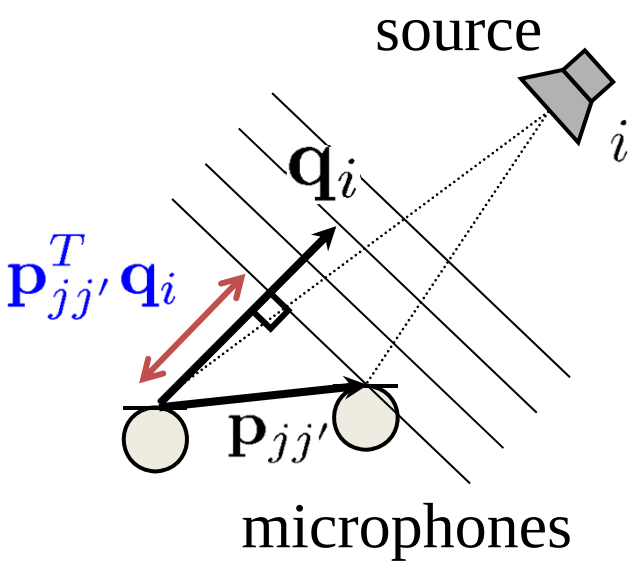
Linear equation for a pair

- Path difference

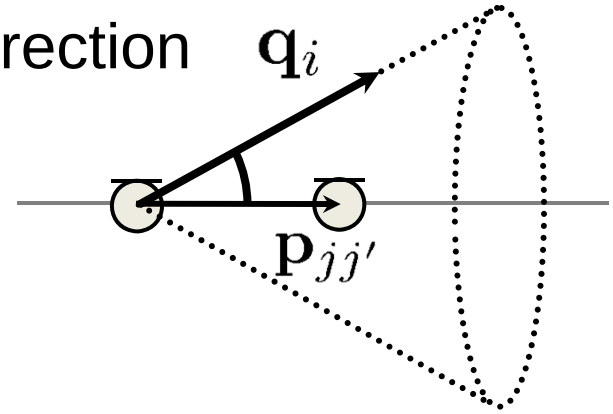
velocity

$$\mathbf{p}_{jj'}^T \mathbf{q}_i = \Delta_{jj'}^i \cdot v$$

TDOA



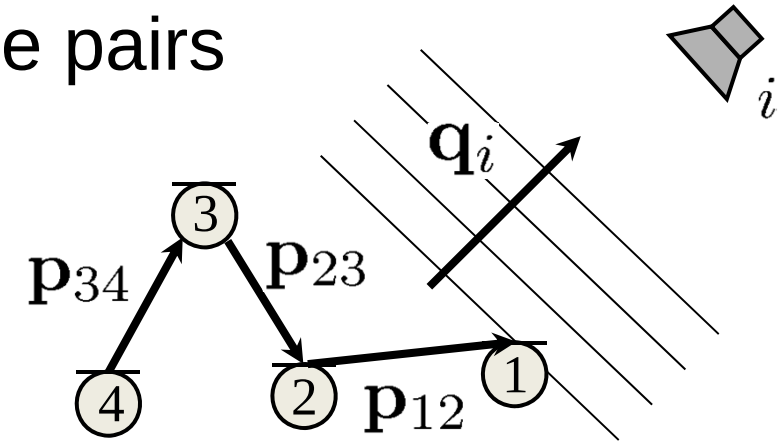
- Cone ambiguity
  - Need more pairs to specify a direction



Linear equations for multiple pairs

- Simultaneous linear equations
  - with multiple microphone pairs

$$\begin{bmatrix} \mathbf{p}_{12}^T \\ \mathbf{p}_{23}^T \\ \mathbf{p}_{34}^T \end{bmatrix} \mathbf{q}_i = \begin{bmatrix} \Delta_{12}^i \\ \Delta_{23}^i \\ \Delta_{34}^i \end{bmatrix} v$$



- DOA estimation $\mathbf{q}_i$
  - Least-squares solution using Moore-Penrose pseudoinverse

$$\mathbf{q}_i = \mathbf{D}^+ \begin{bmatrix} \Delta_{12}^i \\ \Delta_{23}^i \\ \Delta_{34}^i \end{bmatrix} v \qquad \text{with} \quad \mathbf{D} = \begin{bmatrix} \mathbf{p}_{12}^T \\ \mathbf{p}_{23}^T \\ \mathbf{p}_{34}^T \end{bmatrix}$$

Outline of Part II

1. Overview of frequency domain approach
2. Frequency bin-wise separation
    1. ICA: independent component analysis
    2. SCA: sparse component analysis
3. Permutation alignment
    1. Activity sequence clustering
    2. TDOA, DOA clustering
4. Concluding remarks

Concluding remarks

- BSS: blind speech/source separation
  - Convolutive mixtures (mixed in a real room)
  - Frequency-domain approach
    - ICA / SCA applied to each frequency
    - Permutations are aligned afterwards
- Can be solved if the situation is properly setup
  - Many challenges: moving sources, estimating the number of sources, short utterances, …
- Should integrated with other techniques
  - Dereverberation, noise reduction, speech recognition

SiSEC: Signal Separation Evaluation Campaign

http://sisec.wiki.irisa.fr/tiki-index.php

SiSEC 2011

Welcome to the main page for the third community-based Signal Separation Evaluation Campaign (SiSEC 2011).

SiSEC aims to be a large-scale regular campaign building upon the experience of previous evaluation campaigns (SiSEC2008, SiSEC2010) and first community-based Signal Separation Evaluation Campaign (SASSEC).
The unique aspect of this campaign is that, SiSEC is not a competition but a scientific evaluation from which we can draw rigorous scientific conclusions.

The 10th International Conference on Latent Variable Analysis and Signal Processing (LVA/ICA2012) will feature a special session on SiSEC 2011.

- You can download
  - several types of audio mixture data
  - some evaluation Matlab codes

Thank you very much for attending this tutorial !

Shoji Makino

Hiroshi Sawada

Selected references

ICA and BSS books

[Lee, 1998, Haykin, 2000, Hyvärinen et al., 2001, Cichocki and Amari, 2002, Makino et al., 2007]

ICA algorithms

- Information-maximization approach [Bell and Sejnowski, 1995]
- Maximum likelihood (ML) estimation [Cardoso, 1997]
- Natural gradient [Amari et al., 1996, Cichocki and Amari, 2002]
- Equivariance property [Cardoso and Souloumiac, 1996]
- FastICA [Hyvärinen et al., 2001]
- Complex valued ICA [Bingham and Hyvärinen, 2000, Sawada et al., 2003]

Time-domain approach to convolutive BSS

[Amari et al., 1997, Kawamoto et al., 1998, Matsuoka and Nakashima, 2001],
[Douglas and Sun, 2003, Buchner et al., 2004, Takatani et al., 2004, Douglas et al., 2005, Aichner et al., 2006]

Frequency-domain approach to convolutive BSS

[Smaragdis, 1998, Parra and Spence, 2000, Schobben and Sommen, 2002, Murata et al., 2001, Anemüller and Kollmeier, 2000, Mitianoudis and Davies, 2003, Asano et al., 2003, Saruwatari et al., 2003, Ikram and Morgan, 2005, Sawada et al., 2004, Mukai et al., 2006, Sawada et al., 2006, Hiroe, 2006, Kim et al., 2007, Lee et al., 2006, Sawada et al., 2007a, Sawada et al., 2011]

Approaches to permutation alignment

- Making separation matrices smooth in the frequency domain
  [Smaragdis, 1998, Parra and Spence, 2000, Schobben and Sommen, 2002, Buchner et al., 2004]
- Beamforming approach and estimating direction-of-arrival (DOA)
  [Saruwatari et al., 2003, Ikram and Morgan, 2005, Sawada et al., 2004, Mukai et al., 2006, Sawada et al., 2006]
- Correlation of envelopes [Murata et al., 2001, Anemüller and Kollmeier, 2000, Sawada et al., 2004]
- Nonstationary time-varying scale parameter [Mitianoudis and Davies, 2003]

- Multivariate density function [Hiroe, 2006, Kim et al., 2007, Lee et al., 2006]
- Dominance measure [Sawada et al., 2007a, Sawada et al., 2011]
- Simultaneous clustering [Araki et al., 2010]

Time-frequency masking approach to BSS

[Aoki et al., 2001, Rickard et al., 2001, Bofill, 2003, Yilmaz and Rickard, 2004, Roman et al., 2003],
[Araki et al., 2004, Araki et al., 2005], [Kolossa and Orglmeister, 2004, Sawada et al., 2006, Araki et al., 2007]

Line orientation idea for SCA

[O'Grady and Pearlmutter, 2004, O'Grady and Pearlmutter, 2008, Sawada et al., 2011]

Time difference of arrival (TDOA)

[Knapp and Carter, 1976, Omologo and Svaizer, 1997, DiBiase et al., 2001, Chen et al., 2004, Sawada et al., 2007b]

K-means clustering

[Duda et al., 2000]

References

[Aichner et al., 2006] Aichner, R., Buchner, H., Yan, F., and Kellermann, W. (2006). A real-time blind source separation scheme and its application to reverberant and noisy acoustic environments. *Signal Process.*, 86(6):1260–1277.

[Amari et al., 1996] Amari, S., Cichocki, A., and Yang, H. H. (1996). A new learning algorithm for blind signal separation. In *Advances in Neural Information Processing Systems*, volume 8, pages 757–763. The MIT Press.

[Amari et al., 1997] Amari, S., Douglas, S., Cichocki, A., and Yang, H. (1997). Multichannel blind deconvolution and equalization using the natural gradient. In *Proc. IEEE Workshop on Signal Processing Advances in Wireless Communications*, pages 101–104.

[Anemüller and Kollmeier, 2000] Anemüller, J. and Kollmeier, B. (2000). Amplitude modulation decorrelation for convolutive blind source separation. In *Proc. ICA 2000*, pages 215–220.

[Aoki et al., 2001] Aoki, M., Okamoto, M., Aoki, S., Matsui, H., Sakurai, T., and Kaneda, Y. (2001). Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones. *Acoustical Science and Technology*, 22(2):149–157.

[Araki et al., 2004] Araki, S., Makino, S., Blin, A., Mukai, R., and Sawada, H. (2004). Underdetermined blind separation for speech in real environments with sparseness and ICA. In *Proc. ICASSP 2004*, volume III, pages 881–884.

[Araki et al., 2005] Araki, S., Makino, S., Sawada, H., and Mukai, R. (2005). Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask. In *Proc. ICASSP 2005*.

[Araki et al., 2010] Araki, S., Nakatani, T., and Sawada, H. (2010). Simultaneous clustering of mixing and spectral model parameters for blind sparse source separation. pages 5–8.

[Araki et al., 2007] Araki, S., Sawada, H., Mukai, R., and Makino, S. (2007). Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors. *Signal Process.*, 87(8):1833–1847.

[Asano et al., 2003] Asano, F., Ikeda, S., Ogawa, M., Asoh, H., and Kitawaki, N. (2003). Combined approach of array processing and independent component analysis for blind separation of acoustic signals. *IEEE Trans. Speech Audio Processing*, 11(3):204–215.

[Bell and Sejnowski, 1995] Bell, A. and Sejnowski, T. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159.

[Bingham and Hyvärinen, 2000] Bingham, E. and Hyvärinen, A. (2000). A fast fixed-point algorithm for independent component analysis of complex valued signals. *International Journal of Neural Systems*, 10(1):1–8.

[Bofill, 2003] Bofill, P. (2003). Underdetermined blind separation of delayed sound sources in the frequency domain. *Neurocomputing*, 55:627–641.

[Buchner et al., 2004] Buchner, H., Aichner, R., and Kellermann, W. (2004). Blind source separation for convolutive mixtures: A unified treatment. In Huang, Y. and Benesty, J., editors, *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, pages 255–293. Kluwer Academic Publishers.

[Cardoso, 1997] Cardoso, J.-F. (1997). Infomax and maximum likelihood for blind source separation. *IEEE Signal Processing Letters*, 4(4):112–114.

[Cardoso and Souloumiac, 1996] Cardoso, J.-F. and Souloumiac, A. (1996). Jacobi angles for simultaneous diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 17(1):161–164.

[Chen et al., 2004] Chen, J., Huang, Y., and Benesty, J. (2004). Time delay estimation. In Huang, Y. and Benesty, J., editors, *Audio Signal Processing*, pages 197–227. Kluwer Academic Publishers.

[Cichocki and Amari, 2002] Cichocki, A. and Amari, S. (2002). *Adaptive Blind Signal and Image Processing*. John Wiley & Sons.

[DiBiase et al., 2001] DiBiase, J. H., Silverman, H. F., and Brandstein, M. S. (2001). Robust localization in reverberant rooms. In Brandstein, M. and Ward, D., editors, *Microphone Arrays*, pages 157–180. Springer.

[Douglas et al., 2005] Douglas, S. C., Sawada, H., and Makino, S. (2005). A spatio-temporal FastICA algorithm for separating convolutive mixtures. In *Proc. ICASSP 2005*, volume V, pages 165–168.

[Douglas and Sun, 2003] Douglas, S. C. and Sun, X. (2003). Convolutive blind separation of speech mixtures using the natural gradient. *Speech Communication*, 39:65–78.

[Duda et al., 2000] Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classification*. Wiley Interscience, 2nd edition.

[Haykin, 2000] Haykin, S., editor (2000). *Unsupervised Adaptive Filtering (Volume I: Blind Source Separation)*. John Wiley & Sons.

[Hiroe, 2006] Hiroe, A. (2006). Solution of permutation problem in frequency domain ICA using multivariate probability density functions. In *Proc. ICA 2006 (LNCS 3889)*, pages 601–608. Springer.

[Hyvärinen et al., 2001] Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. John Wiley & Sons.

[Ikram and Morgan, 2005] Ikram, M. Z. and Morgan, D. R. (2005). Permutation inconsistency in blind speech separation: Investigation and solutions. *IEEE Trans. Speech Audio Processing*, 13(1):1–13.

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

[Kim et al., 2007] Kim, T., Attias, H. T., Lee, S.-Y., and Lee, T.-W. (2007). Blind source separation exploiting higher-order frequency dependencies. *IEEE Trans. Audio, Speech and Language Processing*, pages 70–79.

[Knapp and Carter, 1976] Knapp, C. H. and Carter, G. C. (1976). The generalized correlation method for estimation of time delay. *IEEE Trans. Acoustic, Speech and Signal Processing*, 24(4):320–327.

[Kolossa and Orglmeister, 2004] Kolossa, D. and Orglmeister, R. (2004). Nonlinear postprocessing for blind speech separation. In *Proc. ICA 2004 (LNCS 3195)*, pages 832–839.

[Lee et al., 2006] Lee, I., Kim, T., and Lee, T.-W. (2006). Complex FastIVA: A robust maximum likelihood approach of MICA for convolutive BSS. In *Proc. ICA 2006 (LNCS 3889)*, pages 625–632. Springer.

[Lee, 1998] Lee, T. W. (1998). *Independent Component Analysis - Theory and Applications*. Kluwer Academic Publishers.

[Makino et al., 2007] Makino, S., Lee, T.-W., and Sawada, H., editors (2007). *Blind Speech Separation*. Springer.

[Matsuoka and Nakashima, 2001] Matsuoka, K. and Nakashima, S. (2001). Minimal distortion principle for blind source separation. In *Proc. ICA 2001*, pages 722–727.

[Mitianoudis and Davies, 2003] Mitianoudis, N. and Davies, M. (2003). Audio source separation of convolutive mixtures. *IEEE Trans. Speech and Audio Processing*, 11(5):489–497.

[Mukai et al., 2006] Mukai, R., Sawada, H., Araki, S., and Makino, S. (2006). Frequency-domain blind source separation of many speech signals using n ear-field and far-field models. *EURASIP Journal on Applied Signal Processing*, 2006:Article ID 83683, 13 pages.

[Murata et al., 2001] Murata, N., Ikeda, S., and Ziehe, A. (2001). An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing*, 41(1-4):1–24.

[O'Grady and Pearlmutter, 2004] O'Grady, P. D. and Pearlmutter, B. A. (2004). Soft-LOST: EM on a mixture of oriented lines. In *Proc. ICA 2004 (LNCS 3195)*, pages 430–436. Springer.

[O'Grady and Pearlmutter, 2008] O'Grady, P. D. and Pearlmutter, B. A. (2008). The LOST algorithm: Finding lines and separating

speech mixtures. *EURASIP Journal on Advances in Signal Processing*, pages Article ID 784296, 17 pages.

[Omologo and Svaizer, 1997] Omologo, M. and Svaizer, P. (1997). Use of the crosspower-spectrum phase in acoustic event location. *IEEE Trans. Speech Audio Processing*, 5(3):288–292.

[Parra and Spence, 2000] Parra, L. and Spence, C. (2000). Convolutive blind separation of non-stationary sources. *IEEE Trans. Speech Audio Processing*, 8(3):320–327.

[Rickard et al., 2001] Rickard, S., Balan, R., and Rosca, J. (2001). Real-time time-frequency based blind source separation. In *Proc. ICA2001*, pages 651–656.

[Roman et al., 2003] Roman, N., Wang, D., and Brown, G. J. (2003). Speech segregation based on sound localization. *Journal of Acousitical Society of America*, 114(4):2236–2252.

[Saruwatari et al., 2003] Saruwatari, H., Kurita, S., Takeda, K., Itakura, F., Nishikawa, T., and Shikano, K. (2003). Blind source separation combining independent component analysis and beamforming. *EURASIP Journal on Applied Signal Processing*, 2003(11):1135–1146.

[Sawada et al., 2007a] Sawada, H., Araki, S., and Makino, S. (2007a). Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS. In *Proc. ISCAS 2007*, pages 3247–3250.

[Sawada et al., 2011] Sawada, H., Araki, S., and Makino, S. (2011). Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. *IEEE Trans. Audio, Speech, and Language Processing*, 19(3):516–527.

[Sawada et al., 2006] Sawada, H., Araki, S., Mukai, R., and Makino, S. (2006). Blind extraction of dominant target sources using ICA and time-frequency masking. *IEEE Trans. Audio, Speech and Language Processing*, 14(6):2165–2173.

[Sawada et al., 2007b] Sawada, H., Araki, S., Mukai, R., and Makino, S. (2007b). Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation. *IEEE Trans. Audio, Speech, and Language Processing*, 15(5):1592–1604.

[Sawada et al., 2003] Sawada, H., Mukai, R., Araki, S., and Makino, S. (2003). Polar coordinate based nonlinear function for frequency domain blind source separation. *IEICE Trans. Fundamentals*, E86-A(3):590–596.

[Sawada et al., 2004] Sawada, H., Mukai, R., Araki, S., and Makino, S. (2004). A robust and precise method for solving the permutation problem of frequency-domain blind source separation. *IEEE Trans. Speech Audio Processing*, 12(5):530–538.

[Schobben and Sommen, 2002] Schobben, L. and Sommen, W. (2002). A frequency domain blind signal separation method based on decorrelation. *IEEE Trans. Signal Processing*, 50(8):1855–1865.

[Smaragdis, 1998] Smaragdis, P. (1998). Blind separation of convolved mixtures in the frequency domain. *Neurocomputing*, 22:21–34.

[Takatani et al., 2004] Takatani, T., Nishikawa, T., Saruwatari, H., and Shikano, K. (2004). High-fidelity blind separation of acoustic signals using SIMO-model-based independent component analysis. *IEICE Trans. Fundamentals*, E87-A(8):2063–2072.

[Yilmaz and Rickard, 2004] Yilmaz, O. and Rickard, S. (2004). Blind separation of speech mixtures via time-frequency masking. *IEEE Trans. Signal Processing*, 52(7):1830–1847.