Audio Source Separation based on Independent Component Analysis

Shoji Makino and **Hiroshi Sawada**

NTT Communication Science Laboratories, Kyoto, Japan

NTT

ICA2003

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

April 1-4, 2003 Nara

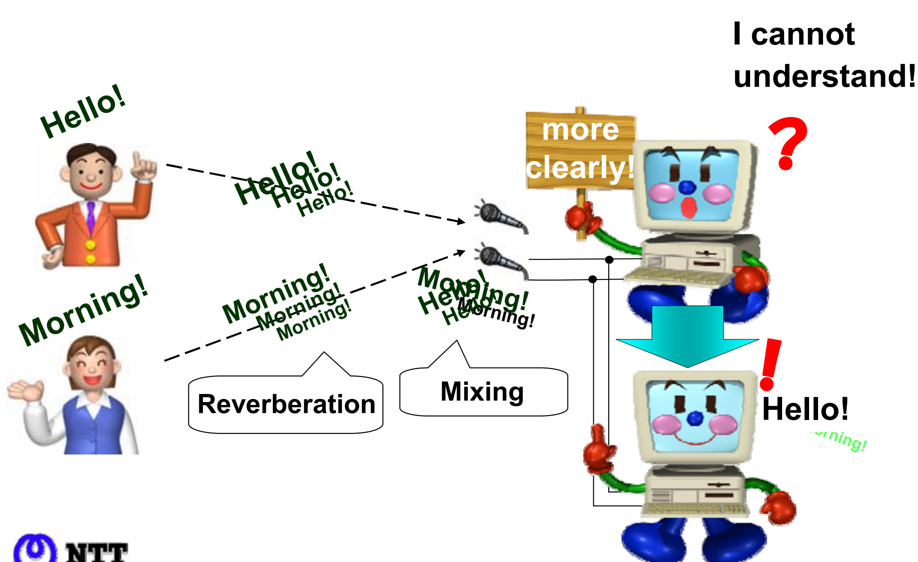
General Chair: Shun-ichi Amari(RIKEN)
Organizing Chair: Shoji Makino(NTT)
Program Chair: Andrzej Cichocki(RIKEN)
Noboru Murata(Waseda)

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

NTT

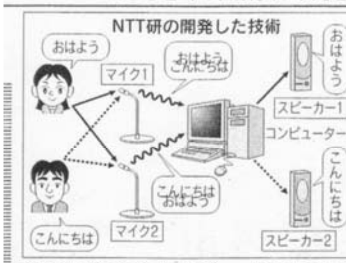
2001年（平成13年）7月17日（火曜日）

NTT

2人の声 同時に聞き分け

会議録の自動編集に応用

NTTコミュニケーション科学基礎研究所は、同時に話す二人の声をコンピューターで聞き分ける技術を開発した。重なった二人の声を話し手ごとに分離し、別々にスピーカーから出力する。今後は聖徳太子の耳のように複数人の声が聞き分けられるよう改良を進め、会議の議事録を自動編集するシステムなどへの応用を目指す。



開発した技術は、部屋の中で話し手がいる位置や雑音の状況などに関係なく使えるのが特徴。二つのマイクがそれぞれとらえた二人の重なった音声をコンピューターで処理し、一人ずつの声に分けてスピーカーから再生する。

二つのマイクがとらえた信号には、二人の位置とマイクの位置が違うために、音声がシステムに到達する時間や音の響き具合（残響）に微妙なずれがある。脳波の分析に使われる独立成分分析という手法を応用、ある変数をかけて二人の話者の信号を比較し、一人の声を大きく、もう一人の声が小さくなるように処理する。

この操作を二回繰り返して、二人分の声をそれぞれ取り出す。

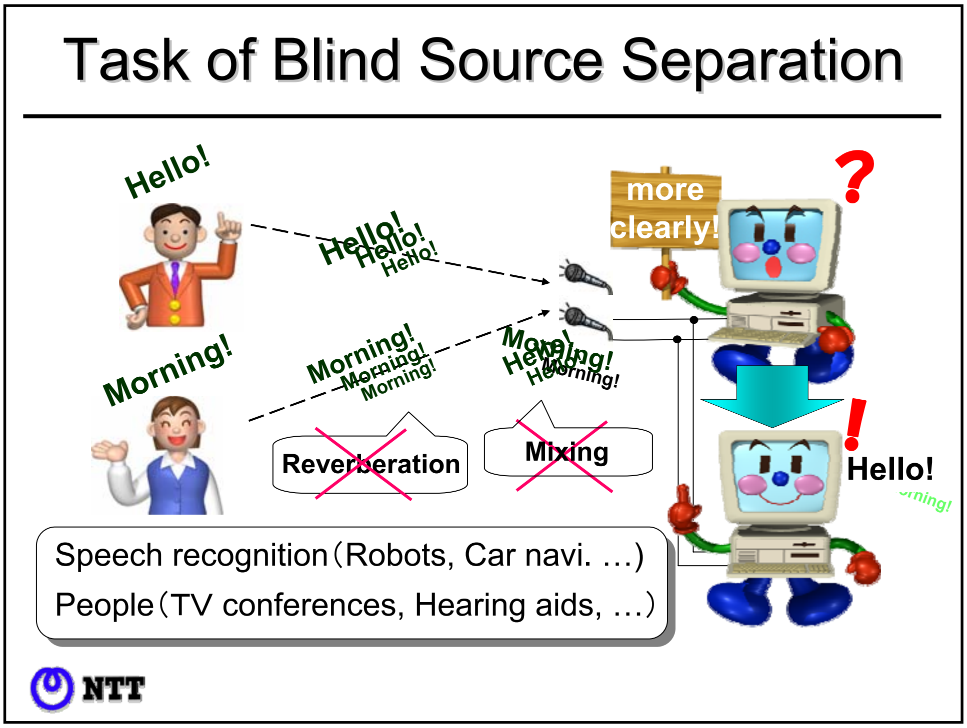
現在の技術は、音声を正確に分けるために話者の位置を固定したり、ヘッドセットマイクを使ったりして話者の声をしっかり取り込む必要があった。雑音が入るとうまく分離できないため、響かないような特殊な部屋でないとうまくいかなかった。新技術では、こうした制限が無く、通常の部

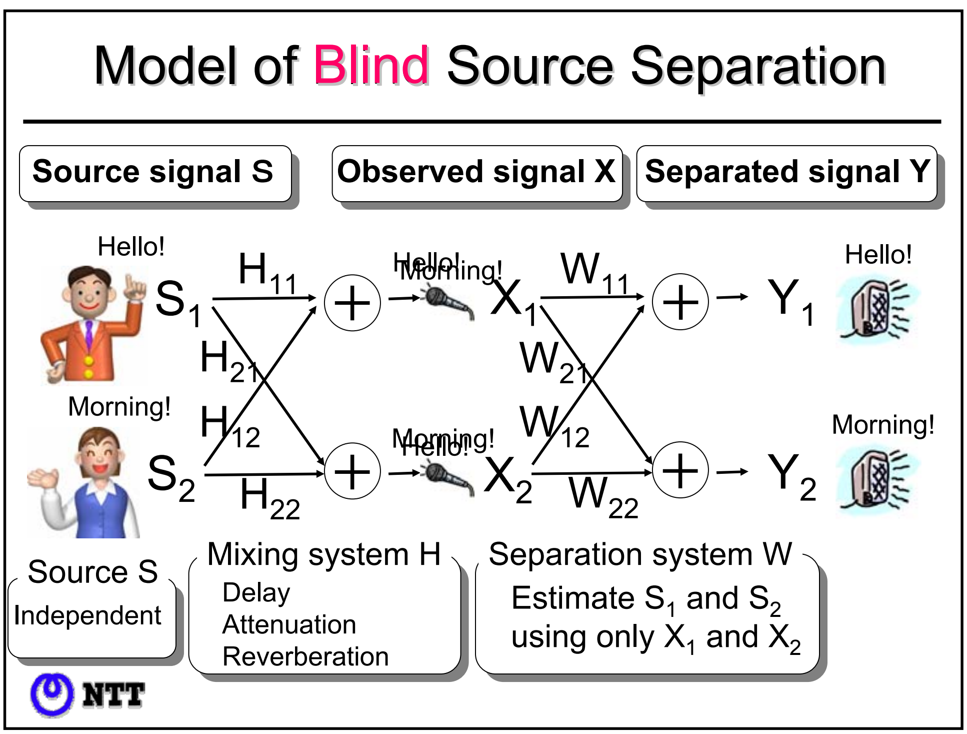
農業生物資源研究所 ㊦

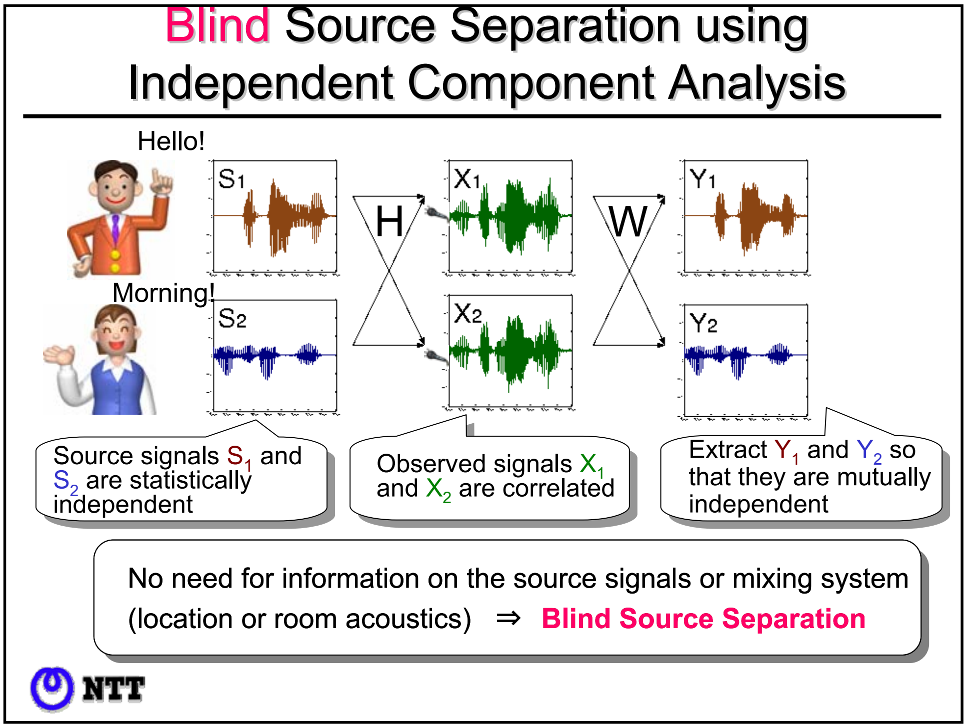
国研から脱皮 独立法人

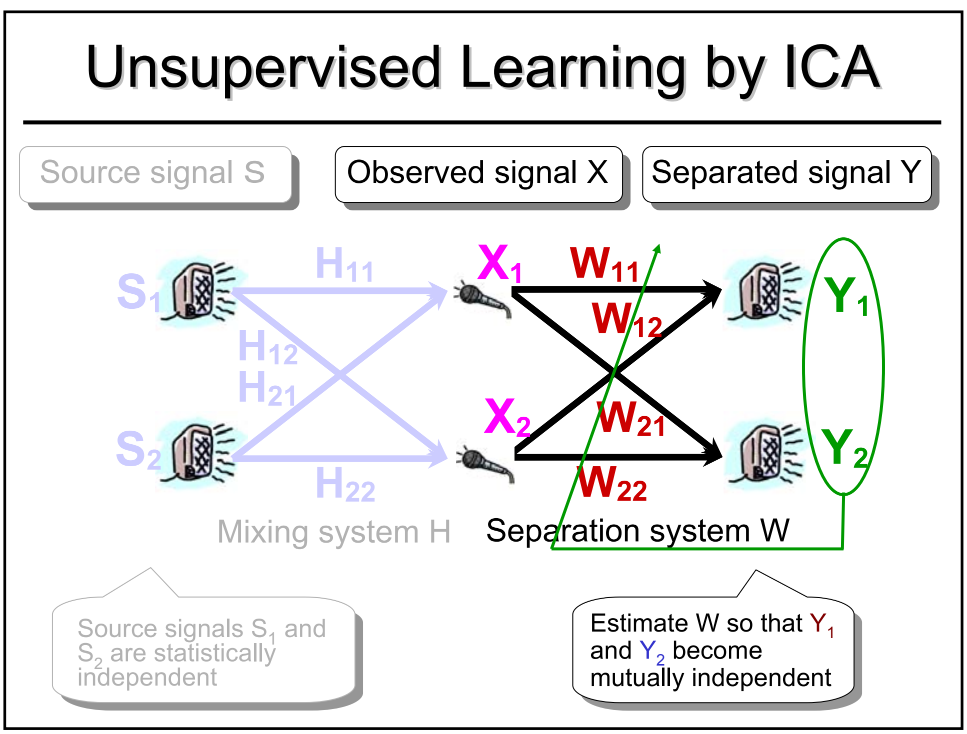
旧農生研時代
らイネゲノム
報）研究に取り
米と協調してい
たが、イネの
見が小麦やトウ
モロコシの改良
分かり一転、
析データはデ
ながらトップレ
るのがマネジメ
桂直樹理事長。
今年度からは
上でイネの品種

NTT


Task of Blind Source Separation
Hello!
Morning!
more clearly!
?
!
Hello!
Reverberation
Mixing
Speech recognition（Robots, Car navi. …)
People（TV conferences, Hearing aids, …)
NTT


Model of Blind Source Separation
Source signal S
Observed signal X
Separated signal Y
Hello!
Morning!
S_1
S_2
H_{11}
H_{21}
H_{12}
H_{22}
X_1
X_2
W_{11}
W_{21}
W_{12}
W_{22}
Y_1
Y_2
Hello!
Morning!
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
NTT


Blind Source Separation using Independent Component Analysis
Hello!
Morning!
S1
S2
H
X1
X2
W
Y1
Y2
Source signals S1 and S2 are statistically independent
Observed signals X1 and X2 are correlated
Extract Y1 and Y2 so that they are mutually independent
No need for information on the source signals or mixing system (location or room acoustics) ⇒ Blind Source Separation
NTT


Unsupervised Learning by ICA
Source signal S
Observed signal X
Separated signal Y
S1
S2
H11
H12
H21
H22
X1
X2
W11
W12
W21
W22
Y1
Y2
Mixing system H
Separation system W
Source signals S1 and S2 are statistically independent
Estimate W so that Y1 and Y2 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

NTT

Background Theory

- Minimization of Mutual Information

  (Minimization of Kullback-Leibler Divergence)

- Maximization of Non-Gaussianity

- Maximization of Likelihood

→ All solutions are identical

NTT

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

NTT

H(･): Entropy

Background Theory

- Maximization of Non-Gaussianity

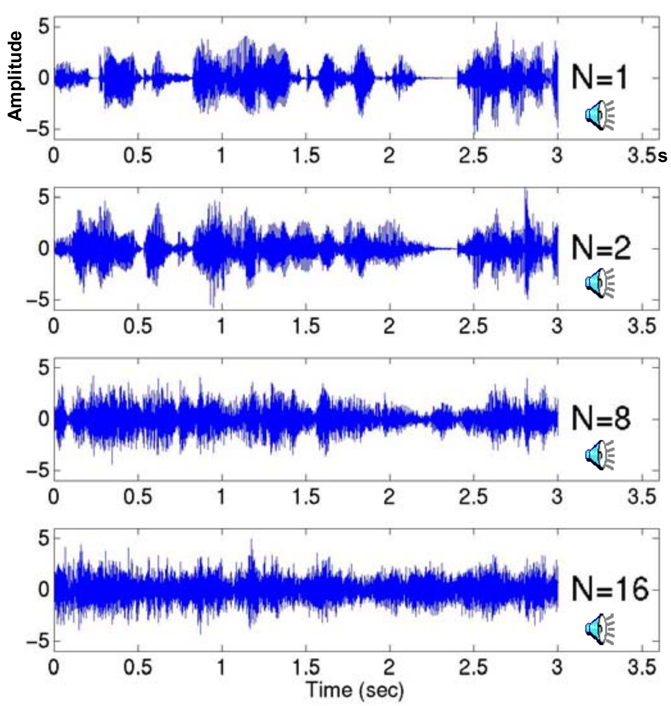
  - Make the output pdf away from Gaussian

NTT

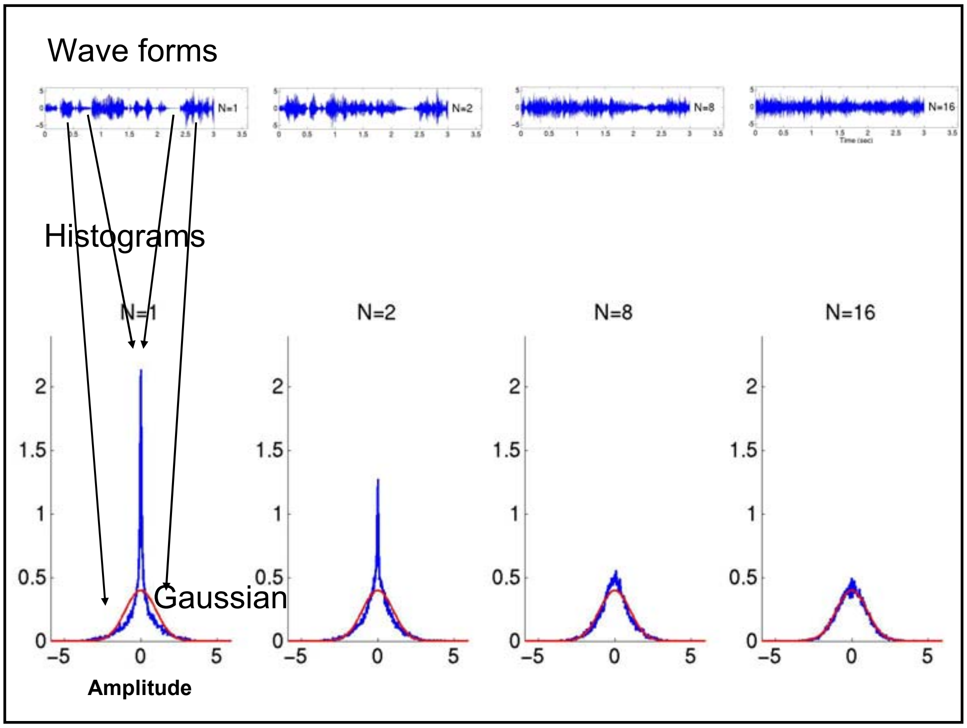
Central Limit Theorem

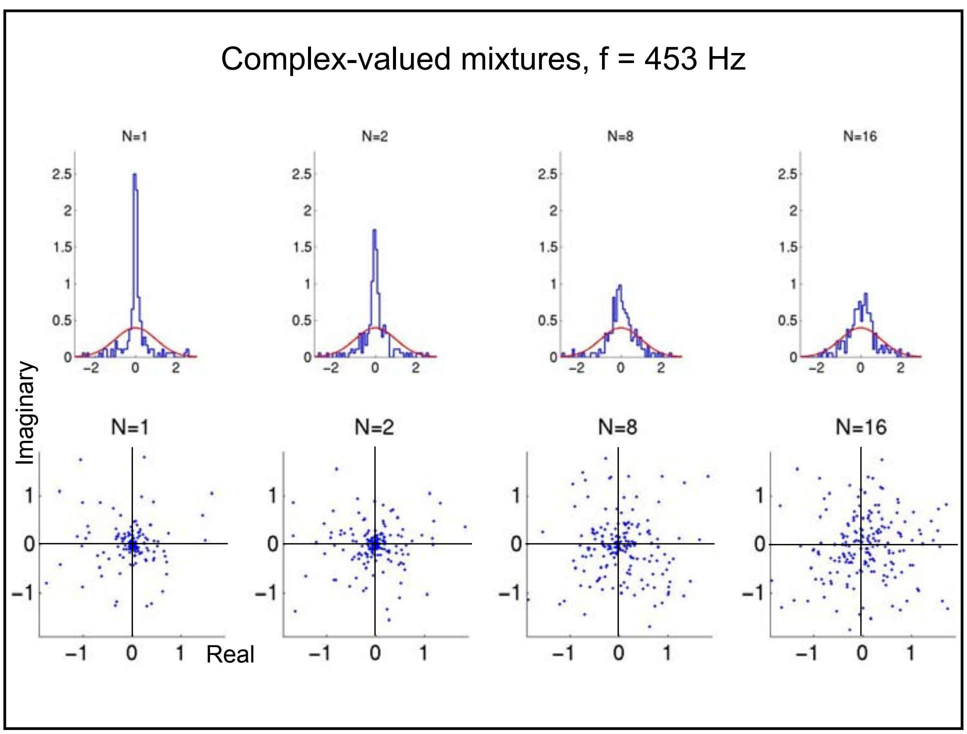
Mix independent components ⇒ Gaussian

NTT

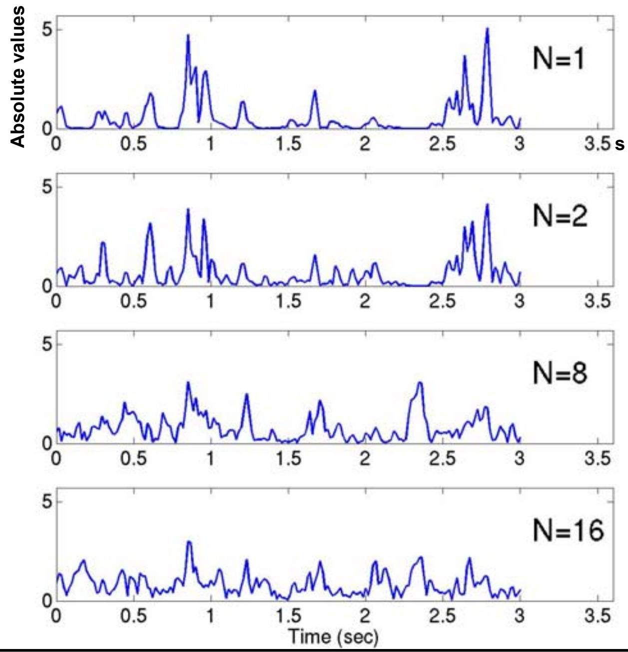


Wave forms

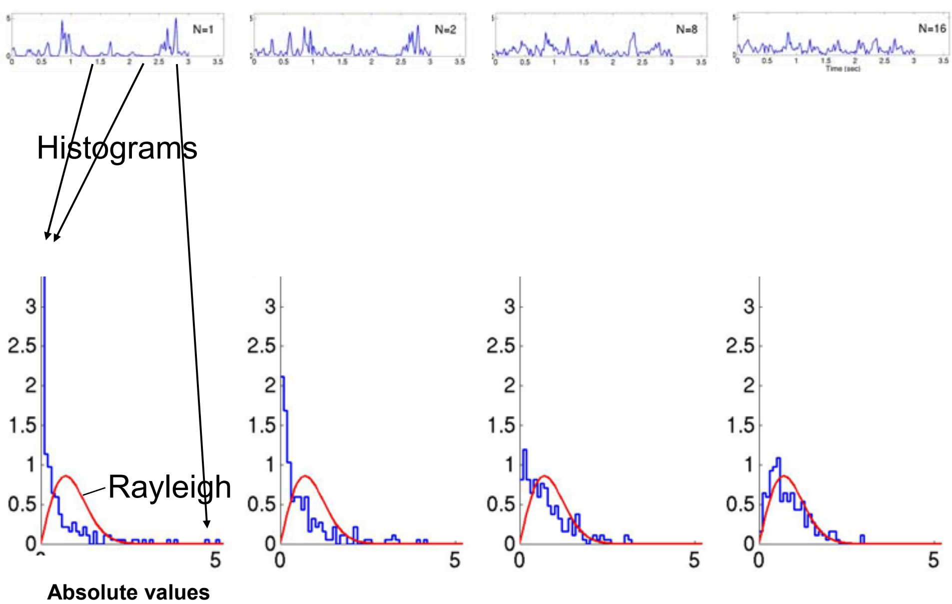

Wave forms
N=1
N=2
N=8
N=16
Time (sec)
Histograms
N=1
N=2
N=8
N=16
Gaussian
Amplitude


Complex-valued mixtures, f = 453 Hz
N=1
N=2
N=8
N=16
Imaginary
Real

Absolute values of complex-valued mixtures, f = 453 Hz



Absolute values of complex-valued mixtures, f = 453 Hz



Central Limit Theorem

Mix independent components ⇒ Gaussian

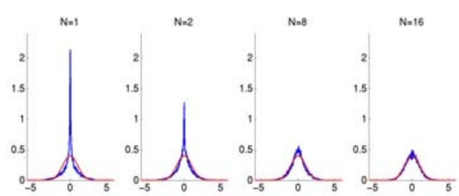
Find independent component ⇒ Non-Gaussian

Non-Gaussianity measures

- Negentropy
- Kurtosis

NTT

Maximization of Negentropies



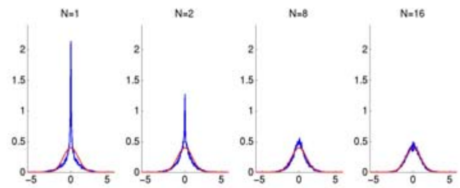
# sources N	1	2	8	16	Gaussian
Entropy *H*	1.19	1.33	1.39	1.40	1.41
Negentropy *N*	0.225	0.087	0.025	0.012	0

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

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

NTT

Maximization of Kurtosis



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

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

NTT

Background Theory

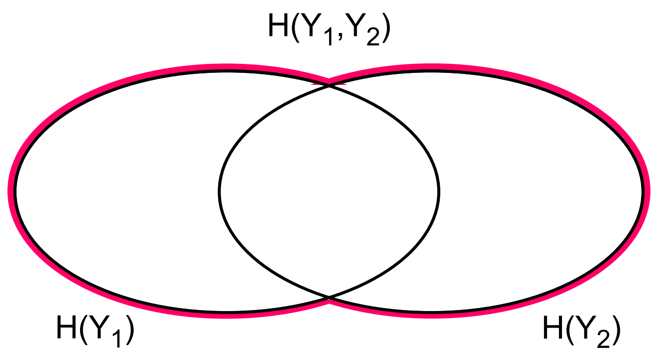
- Minimization of Mutual Information

  (Minimization of Kullback-Leibler Divergence)

  - <u>Make the output “decorrelated”</u>

NTT

Mutual Information

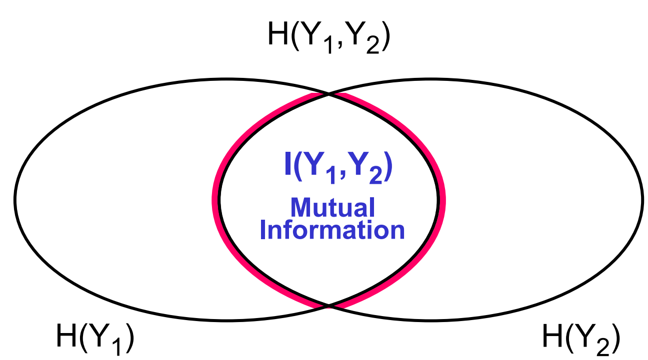


$$I(Y_1,Y_2)=\sum_{i=1}^{2}H(Y_i)-H(Y_1,Y_2)$$

Mutual Information | Marginal Entropy | Joint Entropy

H(・): Entropy

Mutual Information

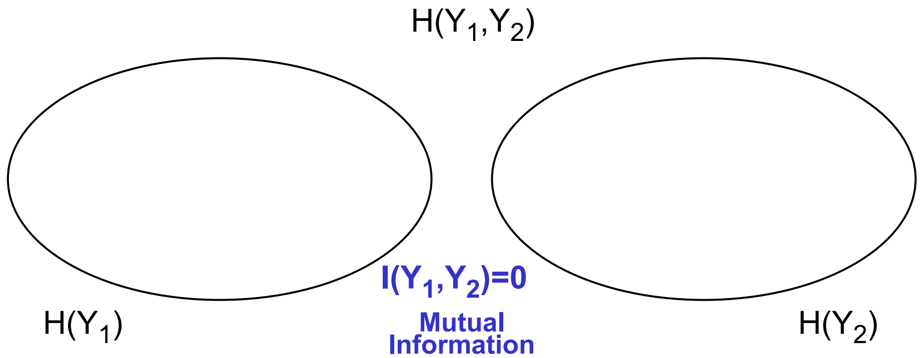


$$I(Y_1,Y_2)=\sum_{i=1}^{2}H(Y_i)-H(Y_1,Y_2)$$

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(・): Entropy

Minimization of Mutual Information

Mutual Information | Marginal Entropy | Joint Entropy

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

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

$$-\int p(Y_1,Y_2)\log\frac{1}{p(Y_1,Y_2)}dY_1dY_2$$

$$=\int p(Y_1,Y_2)\log\frac{p(Y_1,Y_2)}{p(Y_1)p(Y_2)}dY_1dY_2$$

Kullbuck-Leibler Divergence

NTT

Minimization of Mutual Information

Mutual Information

Kullbuck-Leibler Divergence

$$I(Y_1,Y_2)=\int p(Y_1,Y_2)\log\frac{p(Y_1,Y_2)}{p(Y_1)p(Y_2)}dY_1dY_2$$

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

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

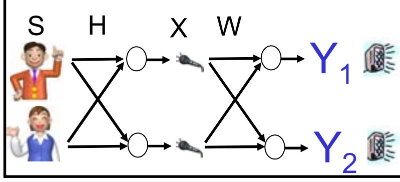
where $\underline{\phi(\mathbf{Y})=-\frac{\mathrm{d}\log p(\mathbf{Y})}{\mathrm{d}\mathbf{Y}}}$

NTT

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

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

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

Average amplitude of Y

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

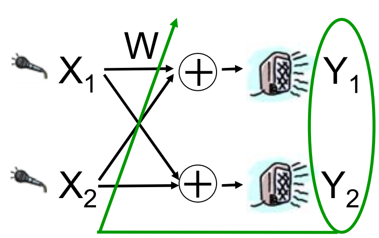
4 equations for 4 unknowns W_{ij}

NTT

ICA Learning Rule

$$W_{i+1} = W_i + \Delta W_i \qquad R_Y = \begin{bmatrix} \langle \phi(Y_1)Y_1 \rangle & \langle \phi(Y_1)Y_2 \rangle \\ \langle \phi(Y_2)Y_1 \rangle & \langle \phi(Y_2)Y_2 \rangle \end{bmatrix}$$

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



Update W so that Y_1 and Y_2 become mutually independent

Second Order Statistics vs. Higher Order Statistics

Second Order Statistics (SOS)

nonstationary sources

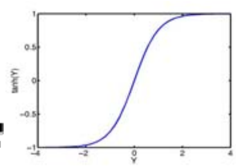
$$\phi(Y_1) = Y_1 \qquad \langle Y_1 Y_2 \rangle = 0$$

for multiple time blocks

Higher Order Statistics (HOS)

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

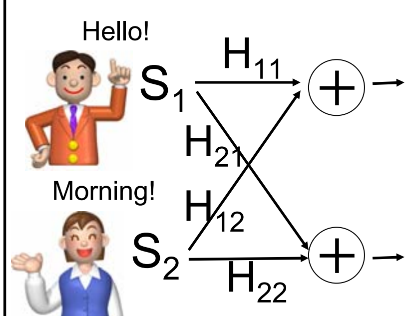
Tailor expansion



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

NTT

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

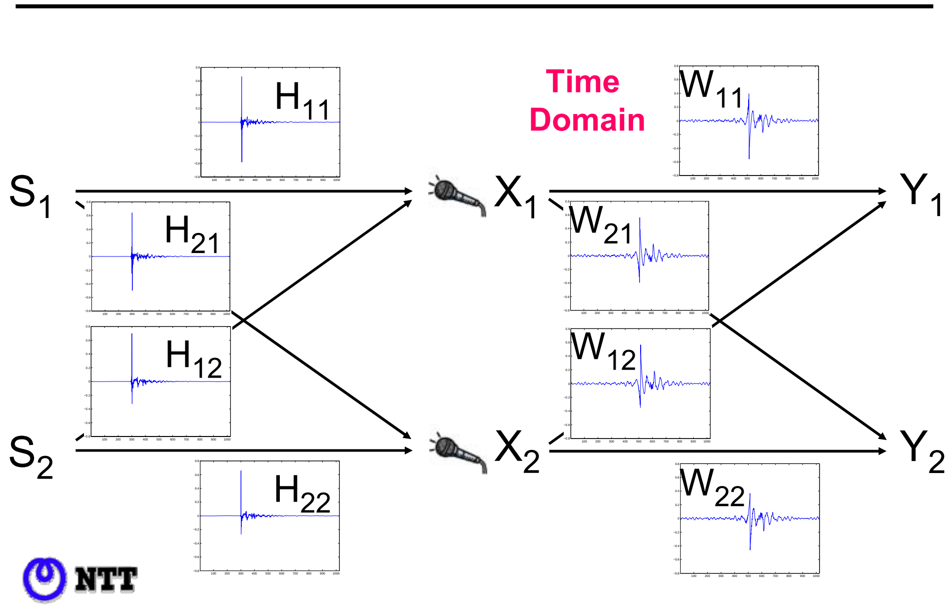
H_{ij} are **FIR filters > 1000 taps**

- sounds in a room

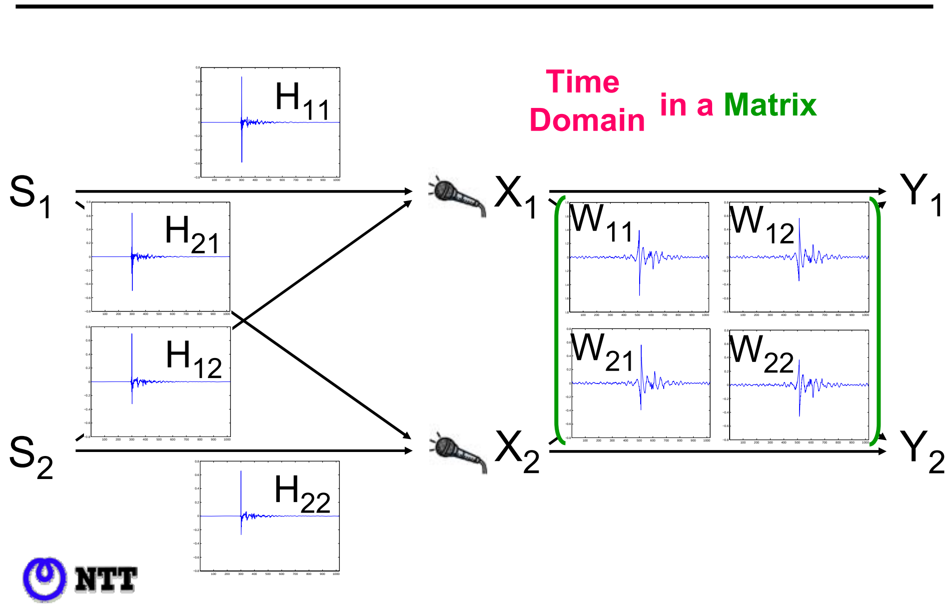
Difficult problem, relatively new

NTT

Mixing Filters and Separation Filters

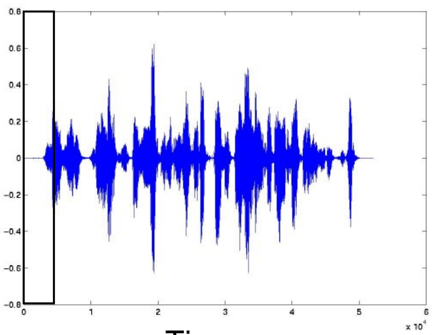


Mixing Filters and Separation Filters

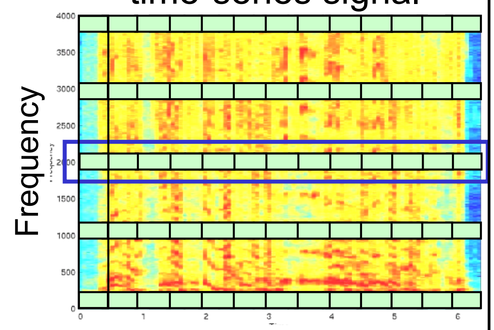


Short Time DFT (Spectrogram)

Time-domain signal



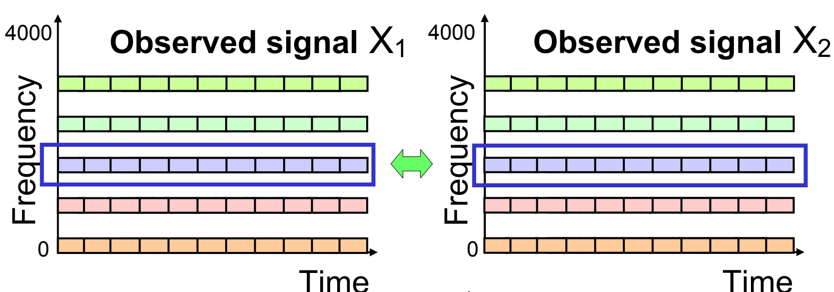
Frequency-domain time-series signal

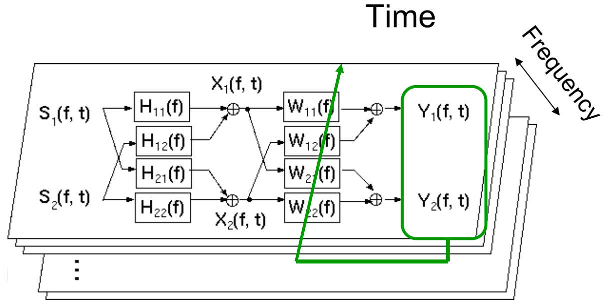


Convolutive mixture in time domain ⟷ Multiple instantaneous mixtures in frequency domain

NTT

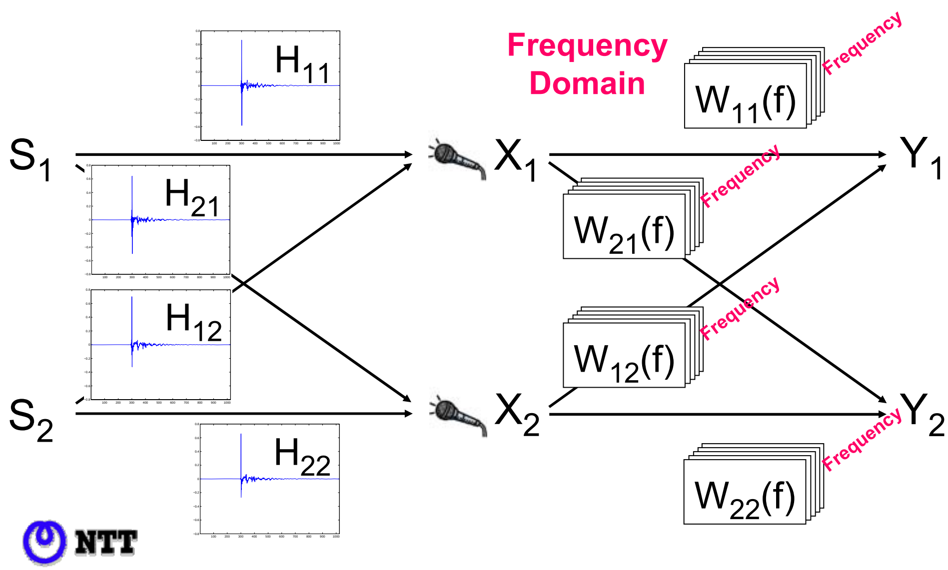
Frequency Domain BSS



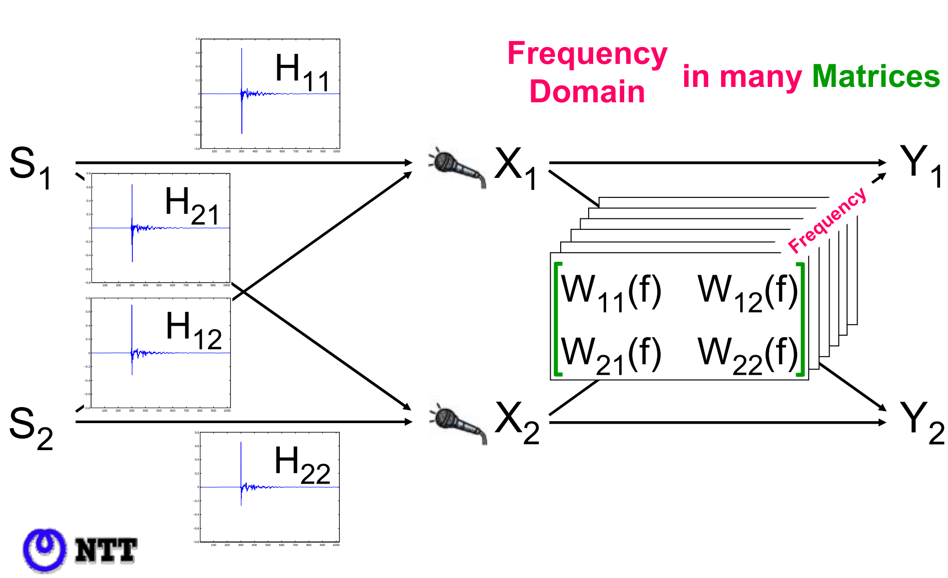


Apply instantaneous ICA approach to each frequency bin

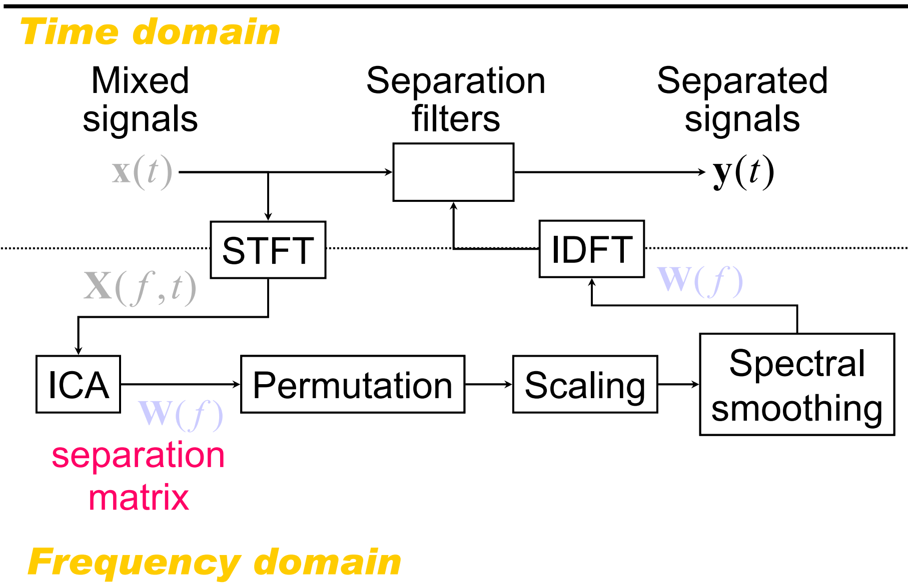
Mixing Filters and Separation Filters



Mixing Filters and Separation Filters



Flow of Frequency Domain BSS

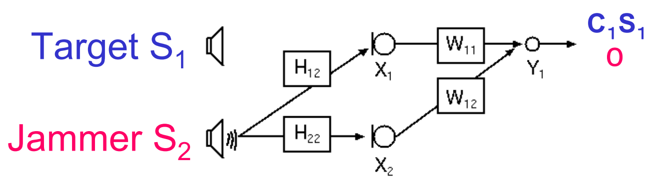


Physical Interpretation of BSS

BSS = Two sets of ABF

NTT

Adaptive Beamformer (ABF)



- Assumptions

  Direction and absence period of a target is known

- Strategy

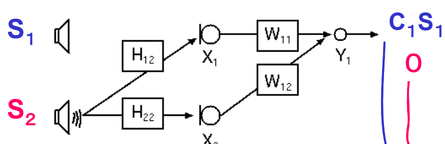
  Minimize the output when only a jammer is active but a target is not active

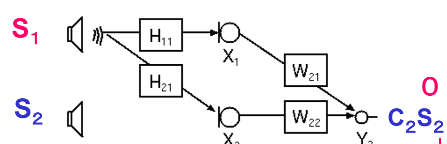
NTT

Two Sets of ABFs

(a) ABF for target S_1 and jammer S_2

(b) ABF for target S_2 and jammer S_1

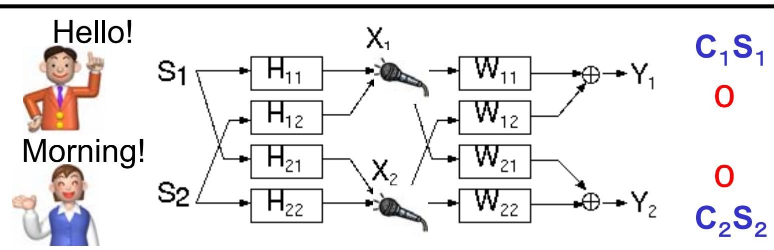




$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}\begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} = \begin{bmatrix} c_1 & 0 \\ 0 & c_2 \end{bmatrix}$$

NTT

Blind Source Separation (BSS)



- Assumptions

  Two sources are mutually independent

- Strategy

  Minimize the SOS or HOS of the outputs

NTT

Diagonalization of $\mathbf{R}_Y(\omega,k)$ in BSS

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

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

- After convergence, the off-diagonal component is

$$
\begin{aligned}
&(E[Y_1Y_2^*])^2 \\
&= \{ad^*\underbrace{E[S_1S_2^*]}_{=0} + bc^*\underbrace{E[S_2S_1^*]}_{=0} + \underbrace{\boxed{(ac^*E[S_1^2] + bd^*E[S_2^2])}}_{0}\}^2 \quad (☆) \\
&= 0
\end{aligned}
$$

(If S_1 and S_2 are ideally independent)

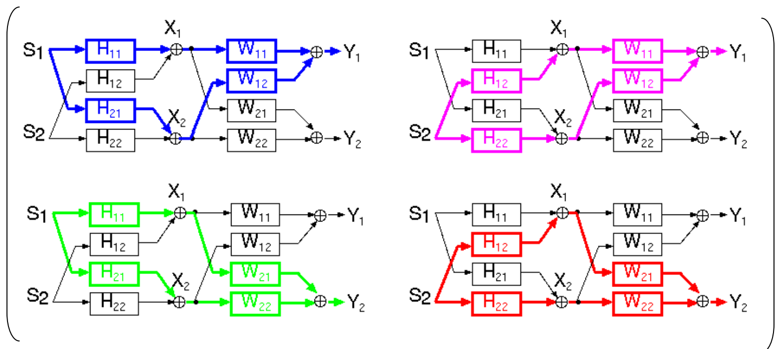
$ac = bd = 0$

NTT

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

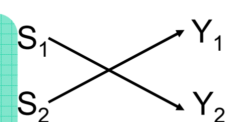
CASE 1: $a=c_1$, $c=0$, $b=0$, $d=c_2$

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

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

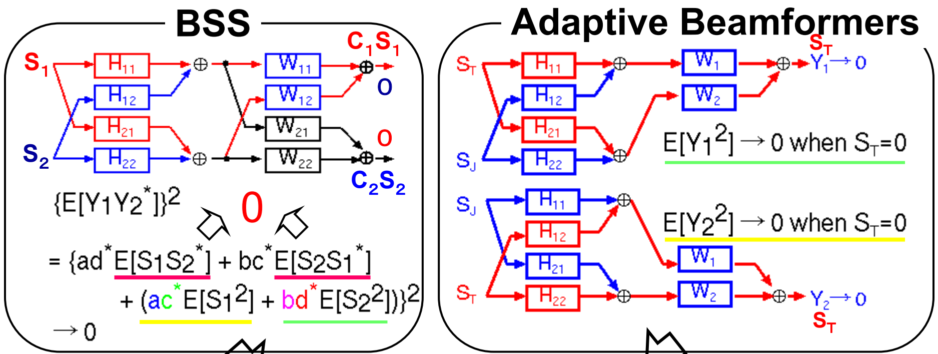
do not appear

∵ we assume

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

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

Equivalence between BSS and ABF



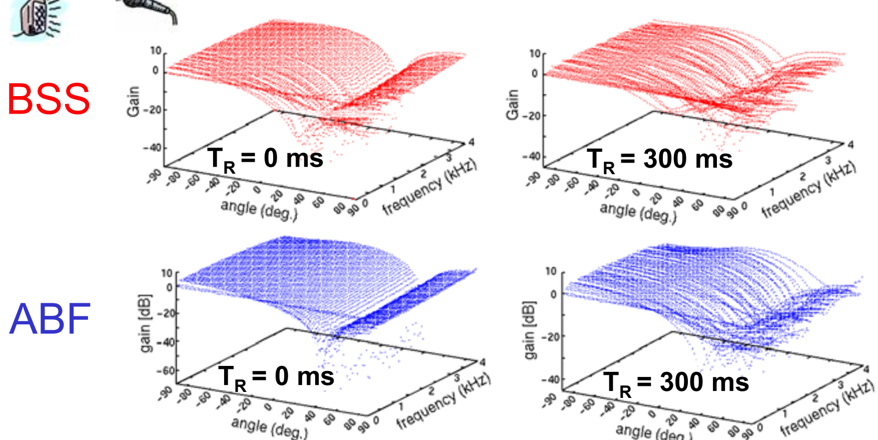
$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix} \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix} = \begin{pmatrix} C_1 & 0 \\ 0 & C_2 \end{pmatrix}$$

BSS and ABF are equivalent
if the independent assumption ideally holds

NTT

Physical Understanding of BSS

BSS, ABF ⇒ Spatial notch to jammer direction

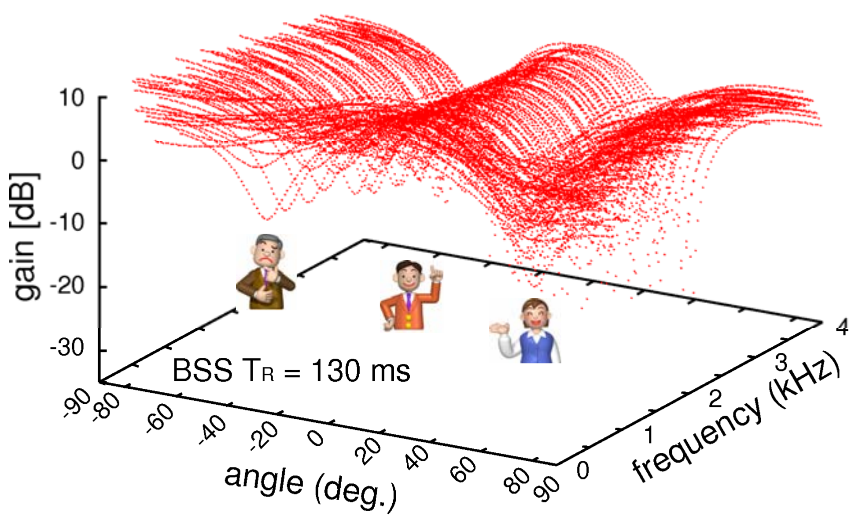


Directivity Patterns

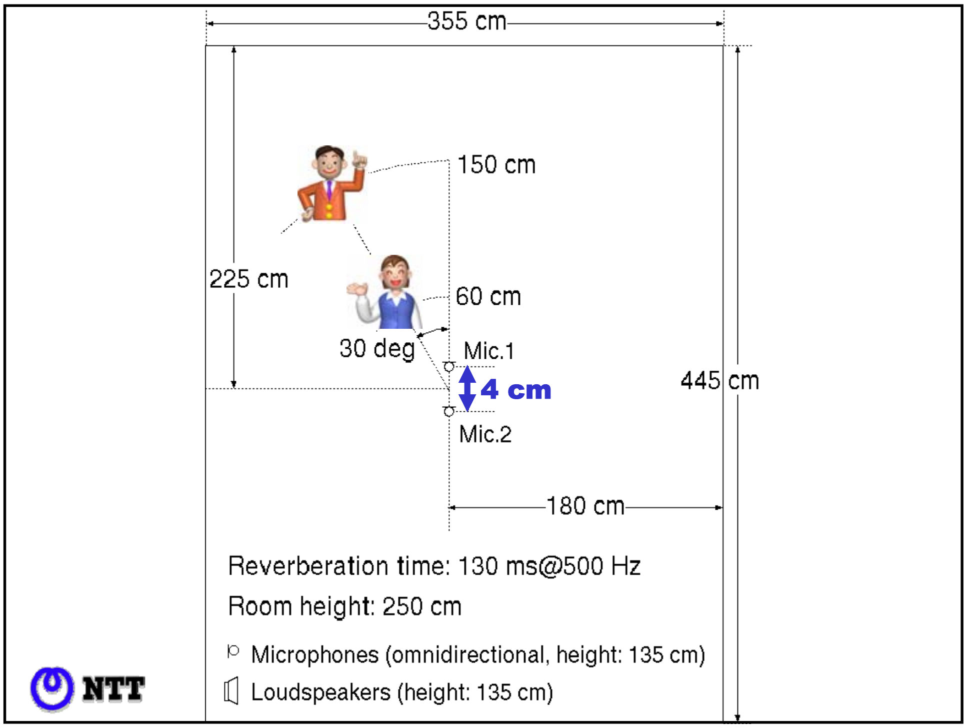
NTT

BSS of Three Speeches

3 sources × 3 sensors BSS

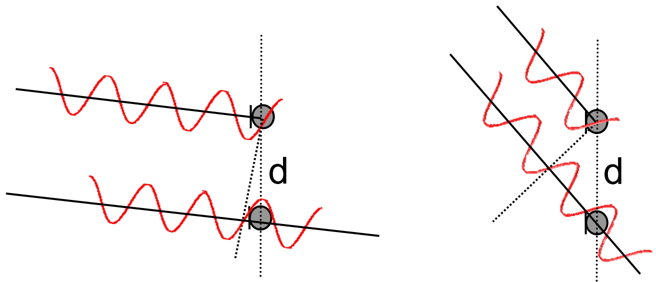


Directivity Patterns



Spatial Aliasing

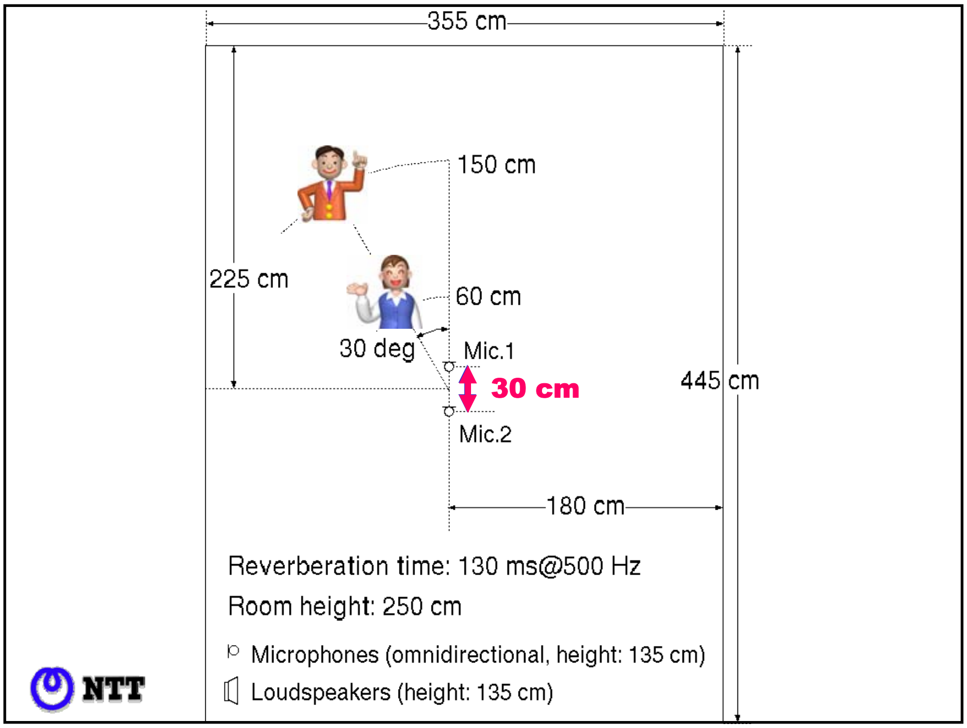
When microphone spacing *d* is too wide…

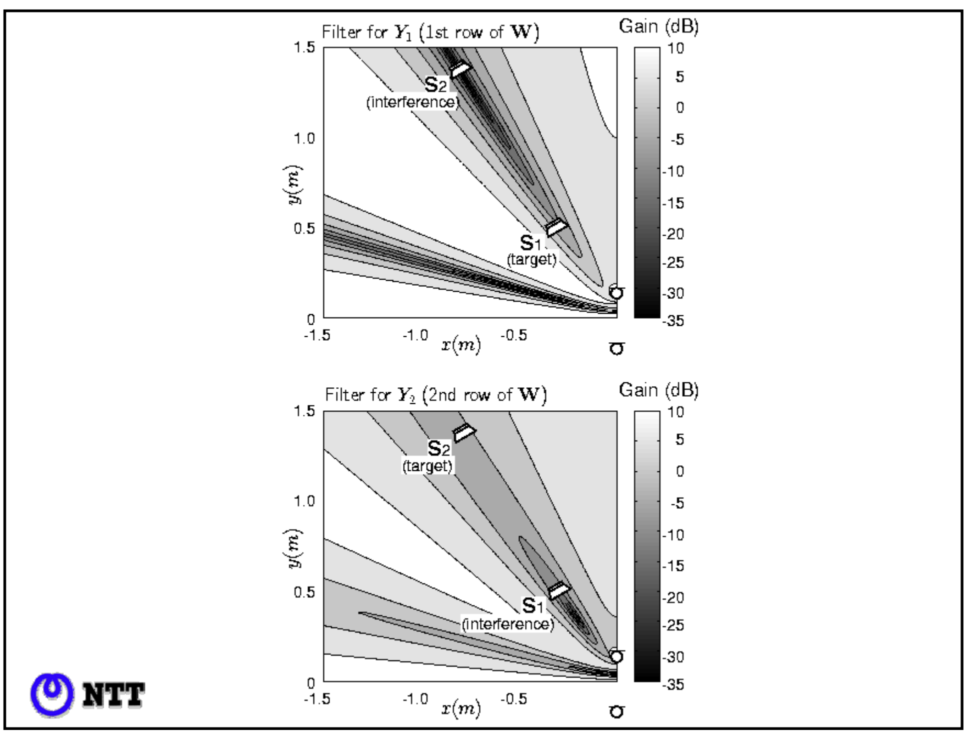


Spatial aliasing does not occer when $d < \frac{\lambda}{2}$

λ : wave length of the highest frequency

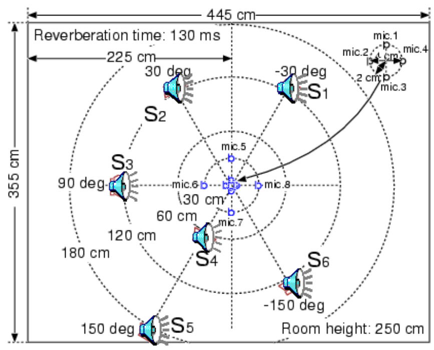
NTT


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)
NTT


Filter for Y_1 (1st row of $\mathbf{W}$)
Gain (dB)
S2
(interference)
S1
(target)
Filter for Y_2 (2nd row of $\mathbf{W}$)
S2
(target)
S1
(interference)
$y(m)$
$x(m)$
NTT

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

	S1	S2	S3	S4	S5	S6	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?

(1) Direct sound
(2) Reverberation
of target

(3) Direct sound
(4) Reverberation
of jammer

Should be extracted

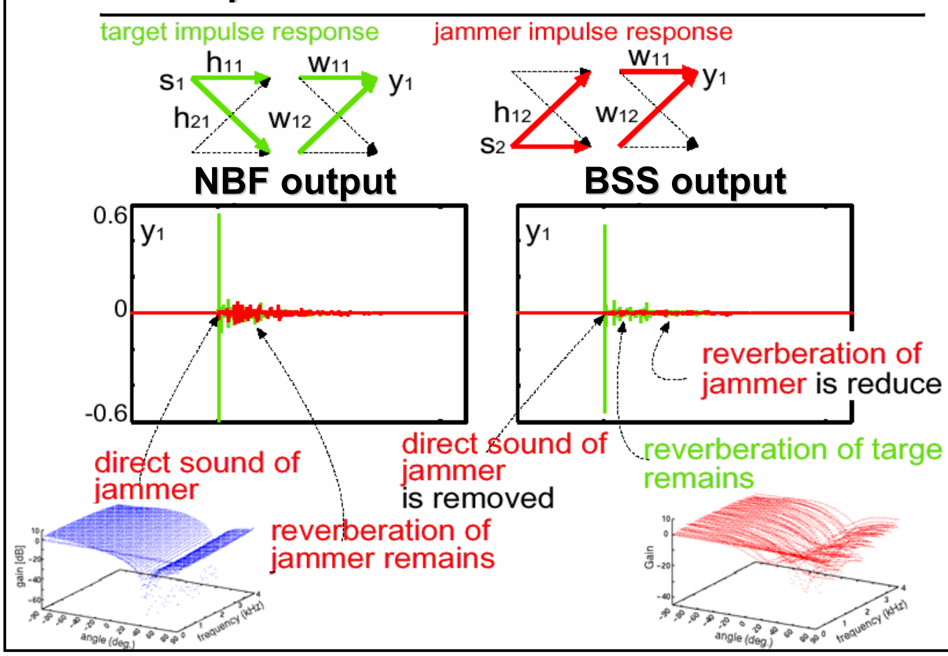
Should be removed

NTT

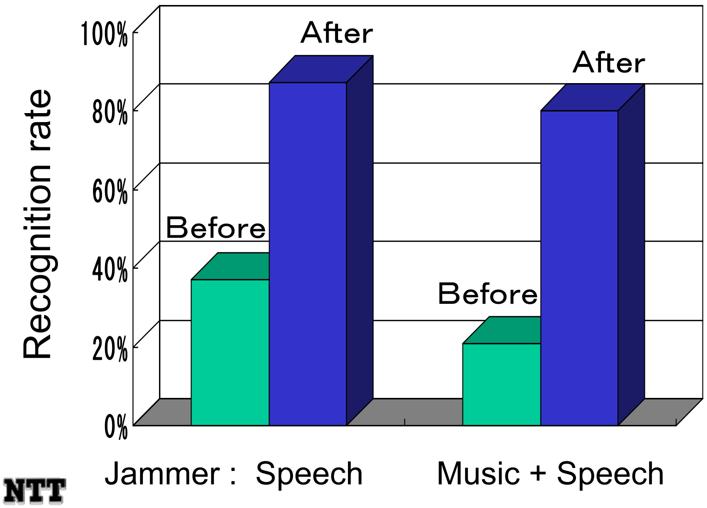
What is separated, and what remains?

BSS = Two sets of ABF

Comparison of NBF and BSS



Effect on Speech Recognition



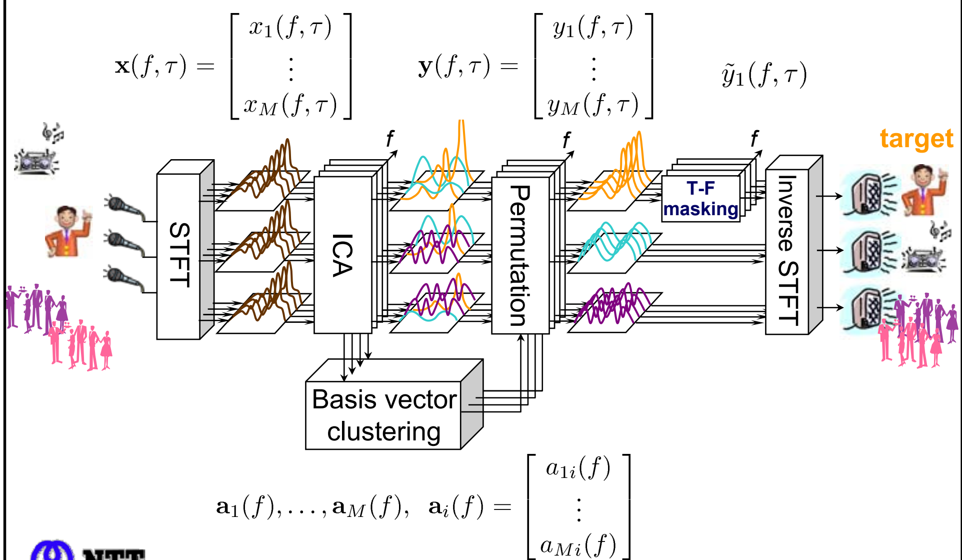
NTT

Blind Source Separation of Many Sounds

VIDEO

NTT

Flow of the F-Domain Method



NTT

Basis Vectors

- Mixtures

$$\mathbf{x}(f,t) = \sum_{k=1}^{N} \mathbf{h}_k(f) s_k(f,t) \qquad \mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_M \end{bmatrix}$$

- ICA

$$\mathbf{x}(f,t) = \mathbf{W}(f)^{-1}\mathbf{y}(f,t)$$

- The inverse of separation matrix

$$[\mathbf{a}_1, \cdots, \mathbf{a}_N] = \mathbf{W}(f)^{-1} \qquad \mathbf{a}_i = \begin{bmatrix} a_{1i} \\ \vdots \\ a_{Mi} \end{bmatrix}$$

basis vector

- Decomposition of mixture

$$\mathbf{x}(f,t) = \sum_{i=1}^{N} \mathbf{a}_i(f) y_i(f,t)$$

NTT

Basis Vectors

- Basis vector $\mathbf{a}_i(f)$
  - represents the frequency responses from source to all sensors
  - Implies information on the source location

Outline

1. Introduction
2. Convolutive blind source separation (BSS) - Formulation
3. Independent component analysis - Concepts
4. Frequency-domain approach for convolutive mixtures
5. Relationship between BSS and adaptive beamformer - Physical interpretation
6. ➡ (Coffee break)
7. Permutation and scaling problems
8. Dependence on separated signals across frequencies
9. Time-difference-of-arrival (TDOA) and direction-of-arrival (DOA) estimation
10. Sparse source separation

Audio Source Separation based on Independent Component Analysis

Part II

Main topics of first and second parts

- Main topic of the first part
  - Basic concepts of BSS and ICA
  - Convolutive BSS
  - Frequency-domain approach
  - BSS and adaptive beamformer
- Main topic of the second part
  - Detailed procedure of frequency-domain BSS
  - Especially, how to solve permutation problem
  - Sparse source separation

Approaches to convolutive BSS

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

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

t Time
l Filter tap

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

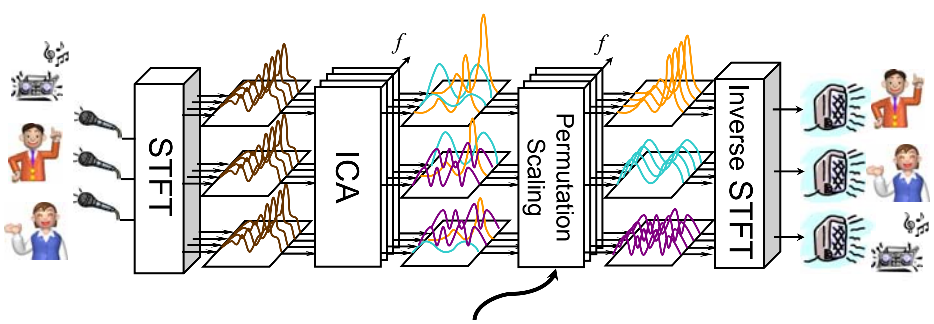
$$y_i(n,f) = \sum_{j=1}^{M} w_{ij}(f) x_j(n,f)$$

n Time frame index
f Frequency

NTT 3

Flow of frequency-domain BSS

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



The second part mainly explains these operations

NTT 4

Outline

- Part I by Shoji Makino

-------- Coffee break --------

- Part II by Hiroshi Sawada
  1. Permutation and scaling problems
  2. Mutual dependence of separated signals across frequencies
  3. Time-difference-of-arrival (TDOA) and direction-of-arrival (DOA) estimation
  4. Sparse source separation

NTT 5

Permutation and Scaling problems

- How important are they?
  - Cannot obtain proper separated signals without considering them
  - Almost all papers on frequency-domain BSS discuss or at least mention these problems
- Number of ICASSP papers that discuss or mention the permutation problem

Year	2000	2001	2002	2003	2004	2005	2006
# papers	2	6	8	8	10	10	11

  - Still increasing!

NTT 6

Permutation and Scaling problem

- Ambiguities of ICA solutions

If $\mathbf{y}(n,f) = \mathbf{W}(f)\,\mathbf{x}(n,f)$ is a solution, then

$\mathbf{y}(n,f) \leftarrow \mathbf{\Lambda}(f)\,\mathbf{P}(f)\,\mathbf{y}(n,f)$ is also a solution

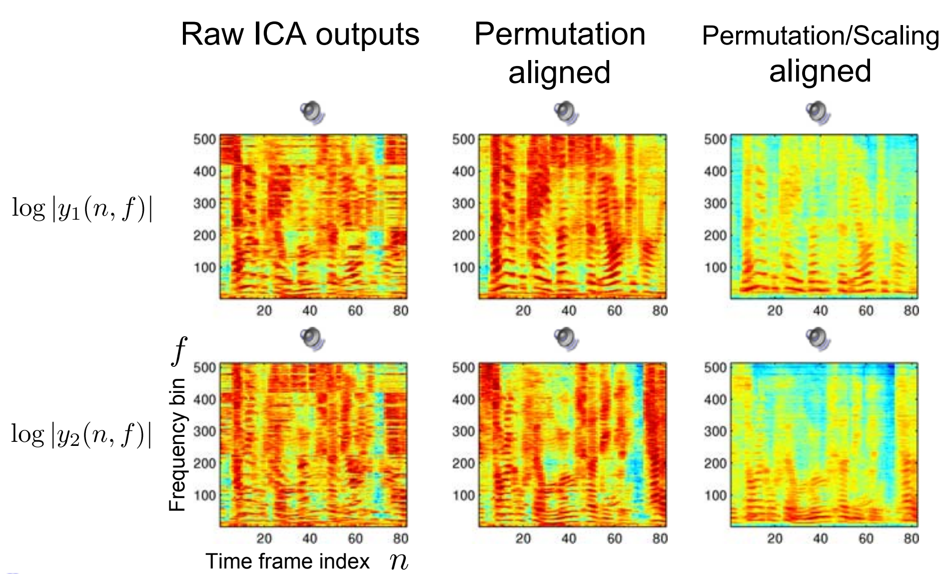
for any diagonal $\mathbf{\Lambda}$ and permutation $\mathbf{P}$ matrix

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \leftarrow \begin{bmatrix} 3 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

Independence of y_1, y_2, y_3 does not change

NTT 7

Permutation and Scaling problem



NTT 8

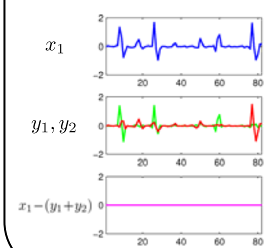
Relationship among sections

Section 1

Estimating mixing situation

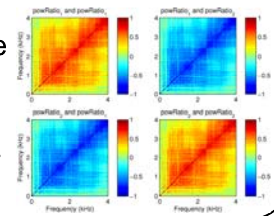
Permutation alignment

Scaling alignment



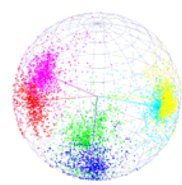
Section 2

Mutual dependence of separated signals across frequencies



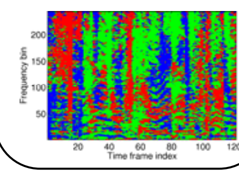
Section 3

TDOA, DOA



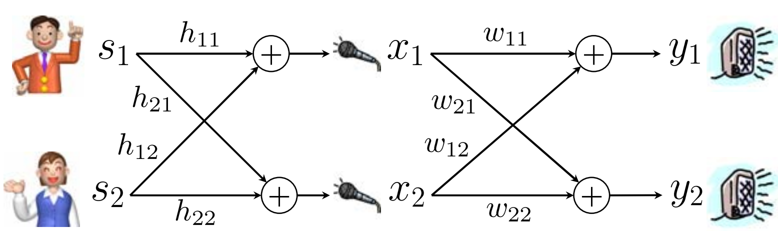
Section 4

Sparse source separation



Mixing model and ICA solution

Frequency-bin view: instantaneous mixture model



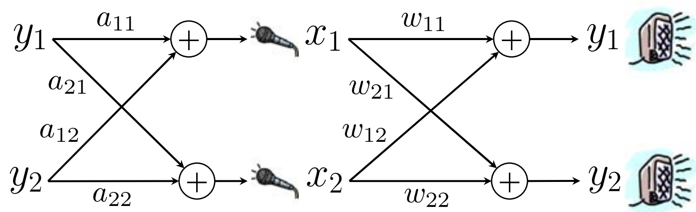
$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} \qquad \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\mathbf{x} = \sum_{k=1}^{N} \mathbf{h}_k s_k \qquad \mathbf{y}(n,f) = \mathbf{W}(f)\,\mathbf{x}(n,f)$$

Mixing model — ICA solution

Estimating mixing situation with ICA

Frequency-bin view: instantaneous mixture model



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

$$\mathbf{x} = \sum_{i=1}^{N} \mathbf{a}_i y_i \Longleftarrow \mathbf{y}(n,f) = \mathbf{W}(f)\,\mathbf{x}(n,f)$$

Estimated mixing situation ICA solution

NTT

Basis vector calculation

Estimated mixing situation ICA solution

$$\mathbf{x} = \sum_{i=1}^{N} \mathbf{a}_i y_i = \mathbf{A}\mathbf{y} \Longleftarrow \mathbf{y} = \mathbf{W}\,\mathbf{x}$$

$$\mathbf{a}_i = \begin{bmatrix} a_{1i} \\ \vdots \\ a_{Mi} \end{bmatrix} \qquad \mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_N] \qquad \mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_M \end{bmatrix}$$

Basis vector

- How to calculate matrix $\mathbf{A}$
  - If $\mathbf{W}$ has an inverse
    $\mathbf{A} = \mathbf{W}^{-1}$
  - Otherwise ($N < M$)
    $\mathbf{A} = \mathrm{E}\{\mathbf{x}\mathbf{y}^H\}(\mathrm{E}\{\mathbf{y}\mathbf{y}^H\})^{-1}$
    $\mathbf{A} = \mathbf{W}^{+}$

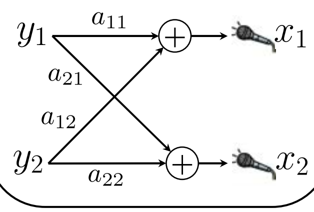
- Least-mean-square estimator that minimizes $\mathrm{E}\{||\mathbf{x} - \mathbf{A}\mathbf{y}||^2\}$
- Moore-Penrose pseudo inverse

NTT

Comparison

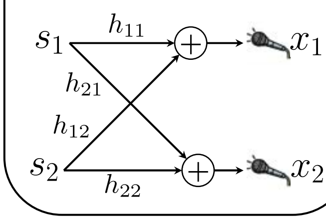
Estimated mixing situation

$$\mathbf{x} = \sum_{i=1}^{N} \mathbf{a}_i y_i$$



Mixing model

$$\mathbf{x} = \sum_{k=1}^{N} \mathbf{h}_k s_k$$



If ICA works well, we expect that

$$\mathbf{a}_i(f) y_i(n, f) = \mathbf{h}_k(f) s_k(n, f)$$

with some correspondence between i and k

frequency dependent

13

Important formula

- Mixing system estimation with basis vectors
  - calculated from ICA solution

$$\boxed{\mathbf{a}_i y_i = \mathbf{h}_k s_k}$$

- Permutation ambiguity
  - Correspondence between i and k is unknown
- Scaling ambiguity

$$\mathbf{a}_i y_i = (\mathbf{a}_i \alpha)\left(\frac{y_i}{\alpha}\right) = \mathbf{h}_k s_k \qquad \text{for any scalar } \alpha$$

  - However, no ambiguity as to the term itself $\mathbf{a}_i y_i$

14

Scaling alignment

- Dereverberation (deconvolution) $y_i = s_k$
  - Eliminating all the effect of impulse responses
  - Difficult task in the blind scenario [references]
    - Even for a single source
- Adjusting to a microphone observation $y_i = h_{Jk} s_k$
  - Popular approach [references]
  - Easily performed if basis vectors are obtained

$$x_J = \underbrace{a_{J1} y_1}_{y_1} + \underbrace{a_{J2} y_2}_{y_2}$$

J :reference microphone

15

Scaling alignment

reference microphone $J = 1$

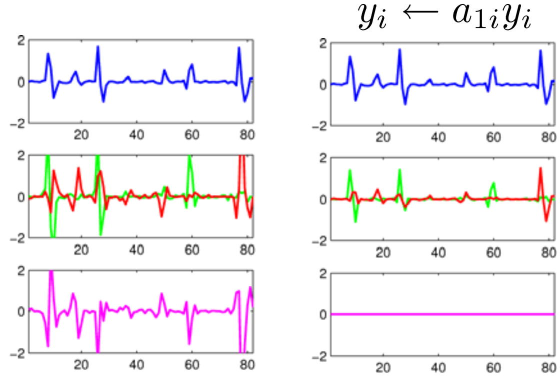
Untouched

Aligned $y_i \leftarrow a_{1i} y_i$

x_1

y_1, y_2

$x_1 - (y_1 + y_2)$



At frequency 773 Hz; only the real part is plotted

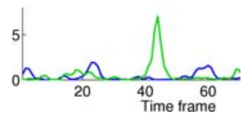
16

Section 1 Summary

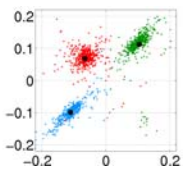
- Permutation and scaling ambiguity
  - Inherent to ICA
  - Serious problem for frequency-domain BSS
- Estimating mixing situation
  - From ICA solution
- Scaling alignment
  - Adjusting to a microphone observation

Permutation alignment

- Various approaches and methods [references]
- In this tutorial, methods based on clustering
  - Bin-wise separated signals $y_i(n, f)$
    - according to their activities



  - Time difference of arrival (TDOA)
    - estimated from basis vectors $\mathbf{a}_i(f)$



- Permutation $\approx$ Clustering
  - Membership assignment is restricted
    - to a permutation $\mathbf{P}(f)$ in each frequency bin

Outline

- Part I

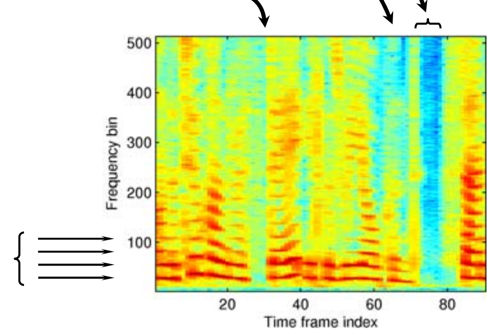
-------- Coffee break --------

- Part II
  1. Permutation and scaling problems
  2. Mutual dependence of separated signals across frequencies
  3. Time-difference-of-arrival (TDOA) and direction-of-arrival (DOA) estimation
  4. Sparse source separation

NTT 19

Dependence across frequencies

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



➡ Mutual dependence of separated signals across frequencies

NTT 20

Approaches to exploit dependencies

- Correlation coefficients [references]

  Will be explained in this section

  - Envelopes $|y_i|$
  - Dominance measure $powRatio_i(n, f)$
- Multivariate density function [references]
  - Models the separated signals of all frequencies
  - ICA algorithm should be modified to accommodate the multivariate density function
    - Natural gradient and FastICA type updates were proposed

NTT 21

Correlation coefficients

- Correlation coefficients 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

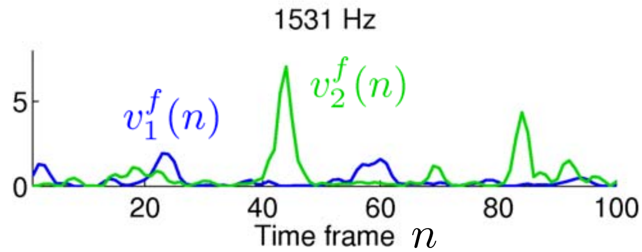
NTT 22

Envelopes of separated signals

- Envelope of bin-wise separated signal

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

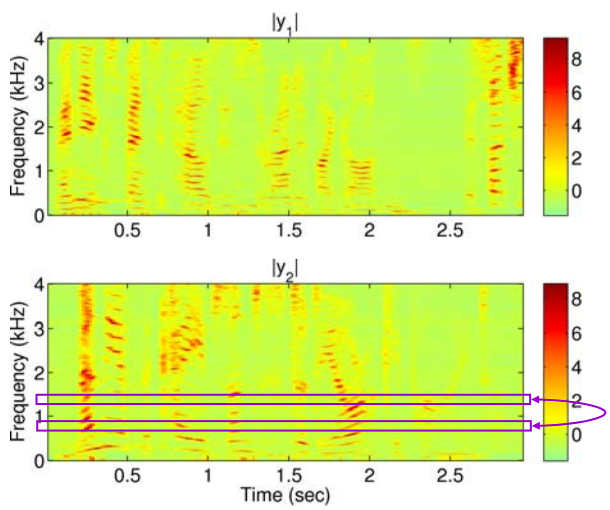
  - At frequency f and at channel i
- Shows the signal activity at the frequency



Envelope examples

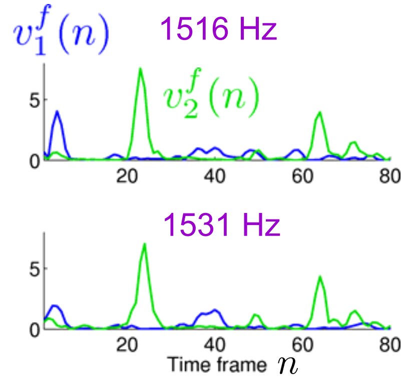
Two separated signals

Normalized to zero-mean and unit-norm



Neighboring frequencies

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

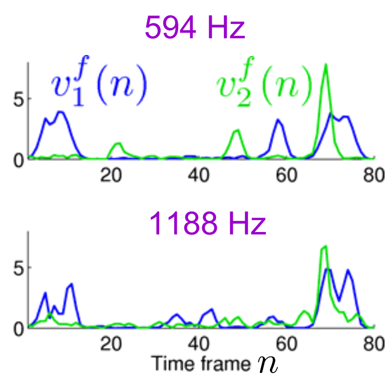


$$\begin{bmatrix} \mathrm{cor}(v_1^f, v_1^{f+1}) & \mathrm{cor}(v_1^f, v_2^{f+1}) \\ \mathrm{cor}(v_2^f, v_1^{f+1}) & \mathrm{cor}(v_2^f, v_2^{f+1}) \end{bmatrix} = \begin{bmatrix} 0.78 & -0.12 \\ -0.15 & 0.92 \end{bmatrix}$$

NTT 25

Harmonic frequencies

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



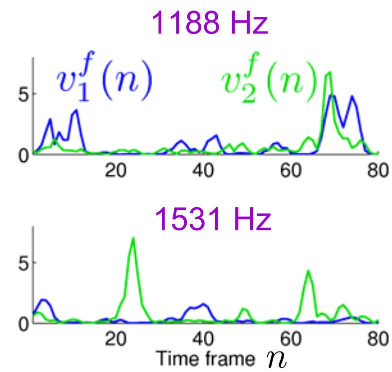
$$\begin{bmatrix} \mathrm{cor}(v_1^f, v_1^{2f}) & \mathrm{cor}(v_1^f, v_2^{2f}) \\ \mathrm{cor}(v_2^f, v_1^{2f}) & \mathrm{cor}(v_2^f, v_2^{2f}) \end{bmatrix} = \begin{bmatrix} 0.76 & 0.36 \\ 0.48 & 0.89 \end{bmatrix}$$

NTT 26

Arbitrary pairs of frequencies

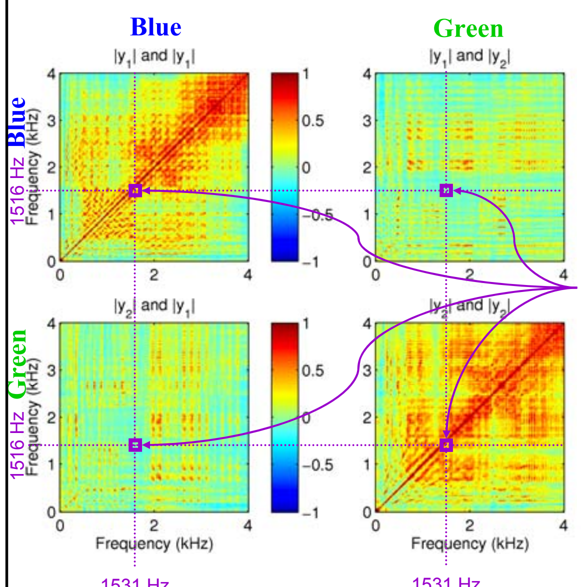
- Among frequencies that have no specific relation
  - May end up with almost zero correlation

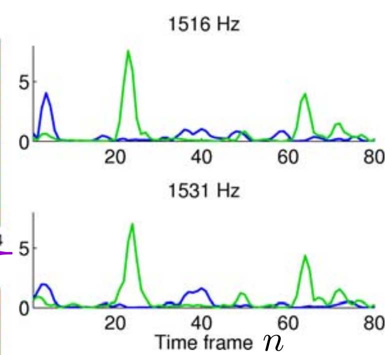
$$\begin{bmatrix} \mathrm{cor}(v_1^f, v_1^g) & \mathrm{cor}(v_1^f, v_2^g) \\ \mathrm{cor}(v_2^f, v_1^g) & \mathrm{cor}(v_2^f, v_2^g) \end{bmatrix} = \begin{bmatrix} 0.10 & -0.14 \\ -0.11 & 0.06 \end{bmatrix}$$



NTT 27

Correlation of envelopes: global view

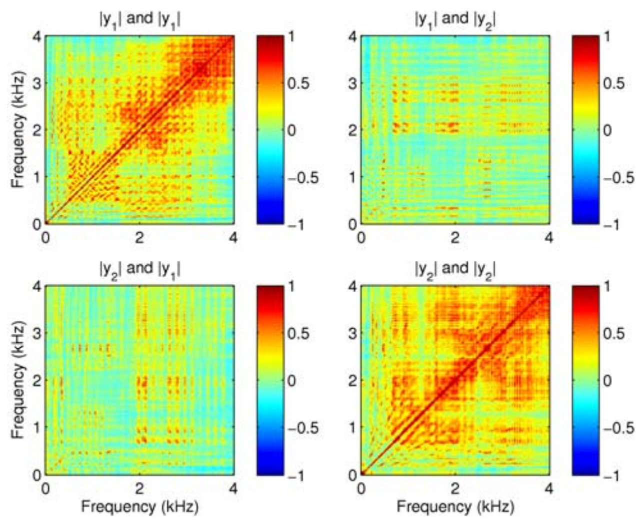




$$= \begin{bmatrix} 0.78 & -0.12 \\ -0.15 & 0.92 \end{bmatrix}$$

NTT 28

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

NTT

Local optimization

- High correlation of envelopes can mostly be seen only when two frequencies are
  - close together or in harmonic relationship
- Local optimization
  - One local mistake leads to a big mistake for the whole

Output1 S_2 S_2 S_2 S_2 S_2 S_2 S_2 S_1 S_1 S_1 S_1 S_1

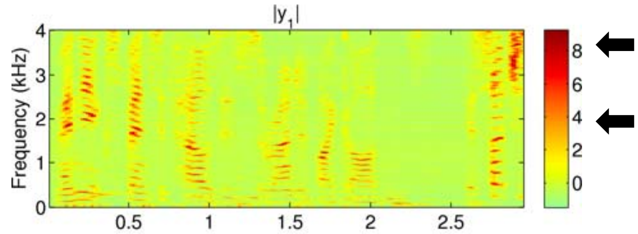
Output2 S_1 S_1 S_1 S_1 S_1 S_1 S_1 S_2 S_2 S_2 S_2 S_2

0 1000 2000 3000 Frequency

NTT

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



- Another type of sequence where active signals are represented uniformly?

➡ High correlation among many frequencies

Dominance measure

- Estimated mixing situation (explained in **Section 1**)

$$\mathbf{x}(n,f) = \sum_{i=1}^{N} \mathbf{a}_i(f) y_i(n,f)$$

- Dominance of i-th signal in mixture [references]

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

$$0 \le powRatio_i \le 1$$

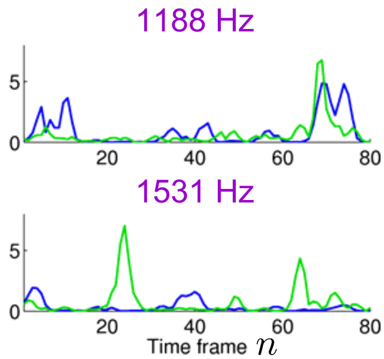
Other signals are dominant — The i-th signal is dominant

  - If sources follow the sparseness assumption (explained in **Section 4**), active signals are represented uniformly with a value close to 1

Envelope and dominance measure

Envelope

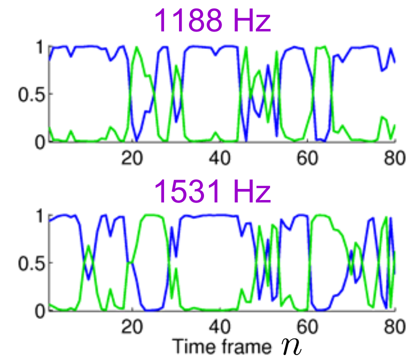
$|y_i(n, f)|$



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

Dominance measure

$powRatio_i(n, f)$



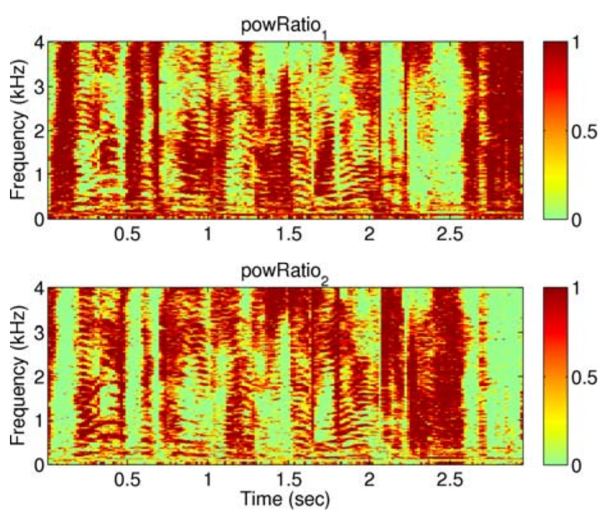
$$= \begin{bmatrix} 0.54 & -0.54 \\ -0.54 & 0.54 \end{bmatrix}$$

NTT

33

powRatio values (dominance measure)

Two separated signals



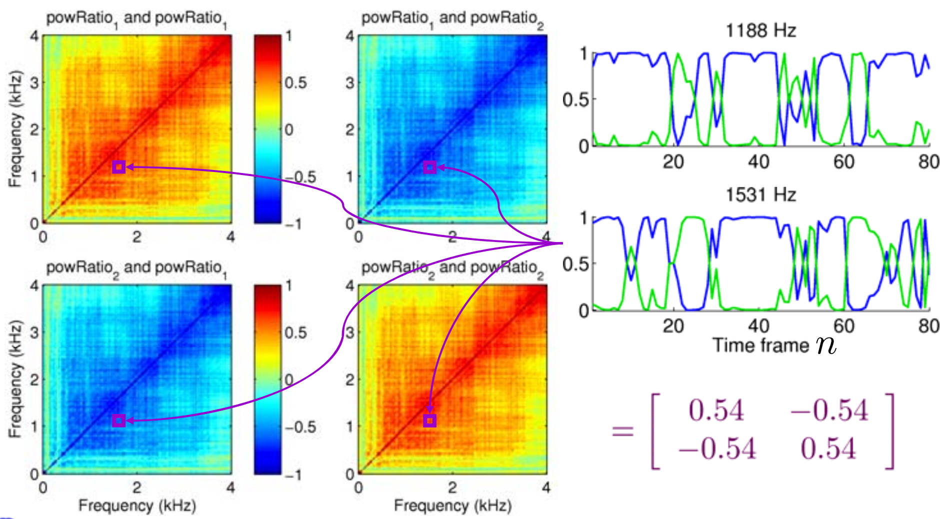
1. Active signal uniformly close to 1
2. Exclusive: if one is close to 1, then the other is close to 0

NTT

34

Correlation of powRatio: global view

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

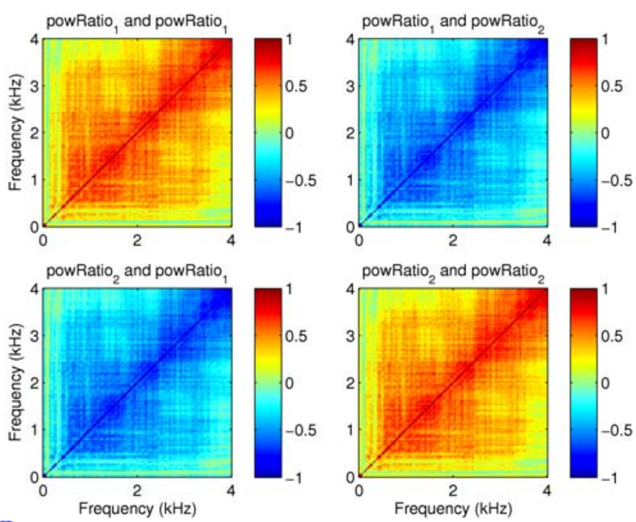


$$= \begin{bmatrix} 0.54 & -0.54 \\ -0.54 & 0.54 \end{bmatrix}$$

NTT

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

NTT

Strategies for permutation alignment

- Local optimization
  - Among neighboring or harmonic frequencies
  - Effective for fine tuning
    - Improves a fairly good solution
- Global optimization
  - Applicable if high correlations within many frequency pairs
  - Efficient and robust algorithms
    - k-means clustering, EM algorithm
    - Centroid or model for each source (cluster)

NTT

Global optimization

Optimization algorithm similar to k-means clustering

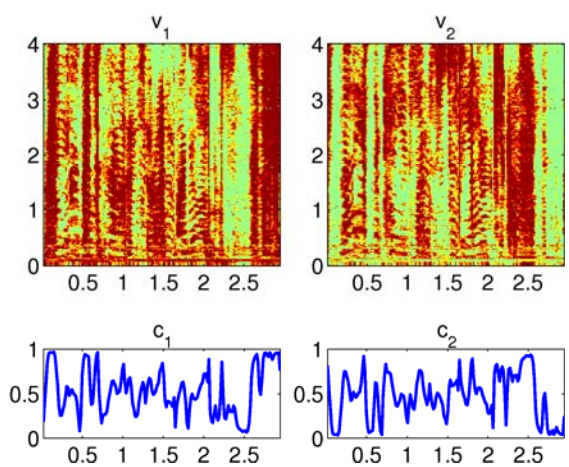
Permutation Π_f aligned to maximize

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

⬇ ⬆

Centroid $c_k(n)$ Member average

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

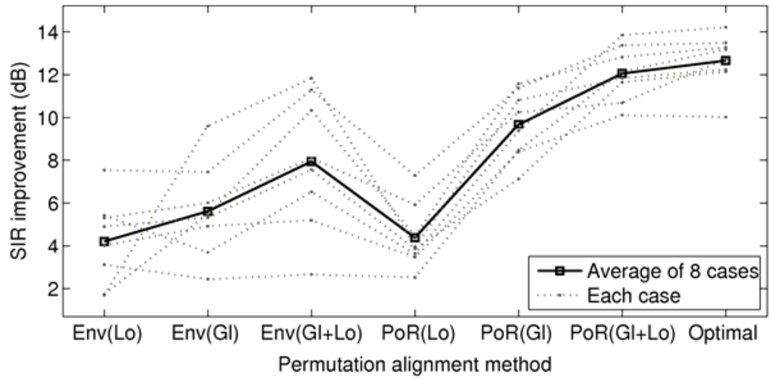


Represents the signal activity for each source

NTT

Experimental results

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.

NTT 39

Section 2 Summary

- Mutual dependence of separated signals
  - Active time frames are expected to coincide across frequencies for the same source
- Signal activity
  - Envelope $|y_i|$
  - Dominance measure $powRatio_i(n, f)$
- Permutation alignment strategies
  - Local optimization
  - Global optimization - clustering

NTT 40

Outline

- Part I

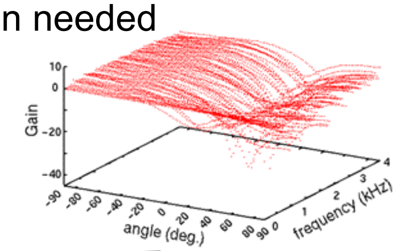
-------- Coffee break --------

- Part II
  1. Permutation and scaling problems
  2. Mutual dependence of separated signals across frequencies
  3. Time-difference-of-arrival (TDOA) and direction-of-arrival (DOA) estimation
  4. Sparse source separation

NTT

Permutation alignment (Spatial information)

- Beamforming approach [references]
  - Directivity patterns calculated with $\mathbf{W}(f)$
  - Direction of arrival (DOA) estimated & clustered
  - Array geometry information needed
- Time difference of arrival (TDOA)
  - Estimated from basis vectors $\mathbf{a}_i(f)$
  - No need for array information

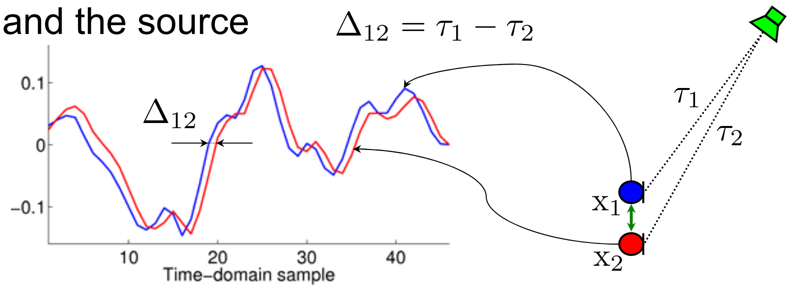


DOA

NTT

Time-difference-of-arrival (TDOA)

- Estimated for each source
- Caused by the positions of microphones and the source $\Delta_{12} = \tau_1 - \tau_2$



- (Generalized) cross correlation

$$\Delta_{12} = \mathrm{argmax}_{\Delta}\, \mathrm{x}_1(t)\, \mathrm{x}_2(t-\Delta)$$

$$\Delta_{12} = \mathrm{argmax}_{\Delta} \sum_f \Phi(f)\, x_1(f) x_2^*(f) e^{\imath 2\pi f \Delta} \qquad \Phi(f) = \frac{1}{|x_1(f)x_2^*(f)|}$$

GCC-PHAT

Frequency-dependent TDOA

- Estimated with observations
  - For each time-frequency slot (n, f)

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_M \end{bmatrix}$$

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

- Estimated with basis vectors
  - For frequency f and output channel i

$$\mathbf{a}_i = \begin{bmatrix} a_{1i} \\ \vdots \\ a_{Mi} \end{bmatrix}$$

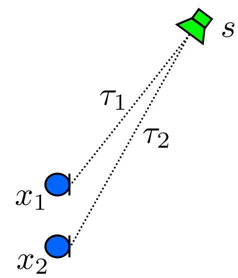
$$\Delta_{12}^i(f) = \frac{\arg[a_{1i}(f)/a_{2i}(f)]}{-2\pi f}$$

Remember the relationship between observation and basis vectors: $\mathbf{x}(n,f) = \sum_{i=1}^N \mathbf{a}_i(f) y_i(n,f)$

Derivation of the estimation formula

- Single source, frequency domain

$$\begin{bmatrix} x_1(n,f) \\ x_2(n,f) \end{bmatrix} = \begin{bmatrix} h_1(f) \\ h_2(f) \end{bmatrix} s(n,f)$$



- Simplified model with time delay

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

- 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 estimation formula

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

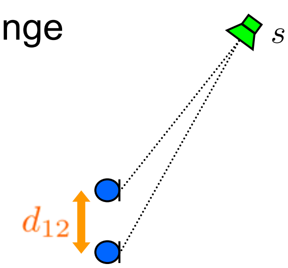
NTT 45

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 $d_{12} = 4\ \mathrm{cm}$

then $0 < f < 4250\ \mathrm{Hz}$

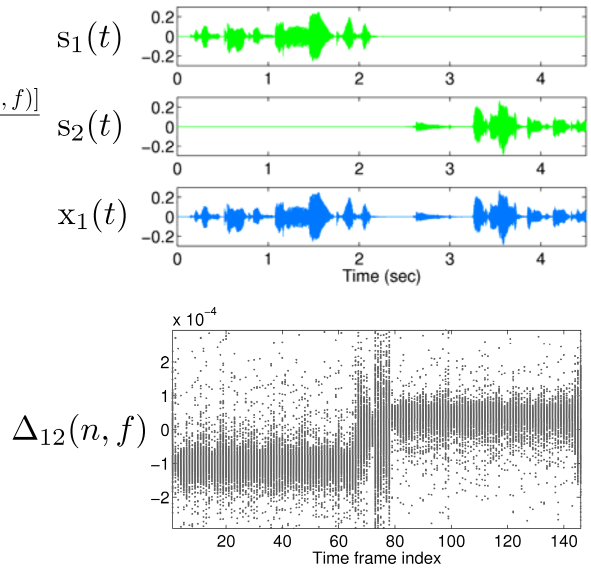
NTT 46

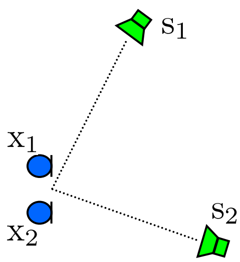
Frequency-dependent TDOA

With observations

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

Non-overlapped





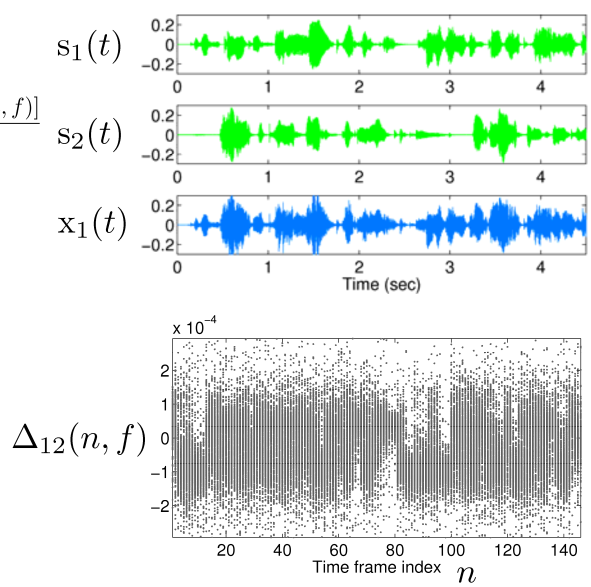
NTT 47

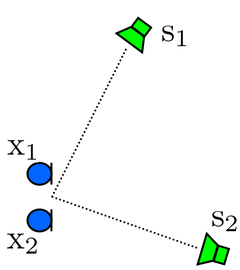
Frequency-dependent TDOA

With observations

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

Overlapped





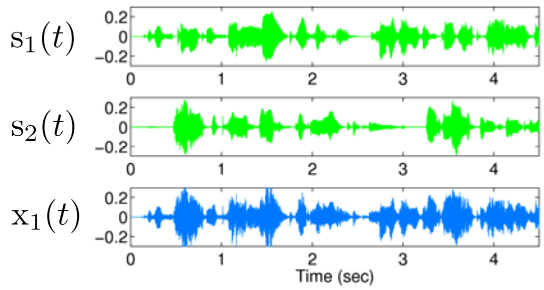
NTT 48

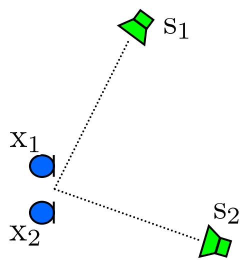
Frequency-dependent TDOA

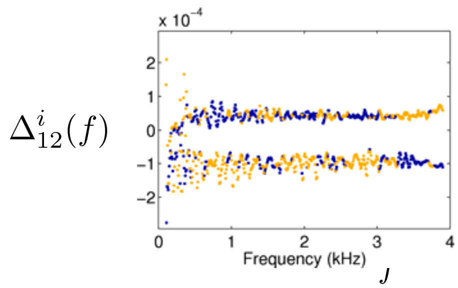
With basis vectors

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

Overlapped



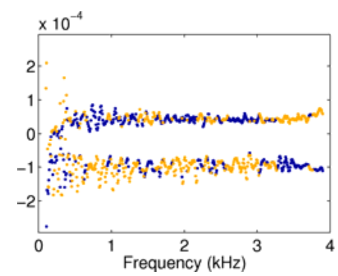




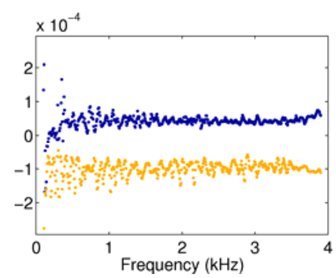
NTT

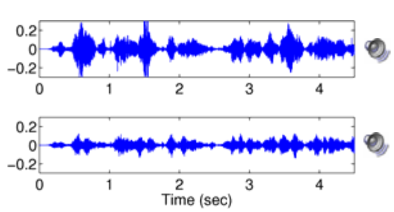
Permutation alignment

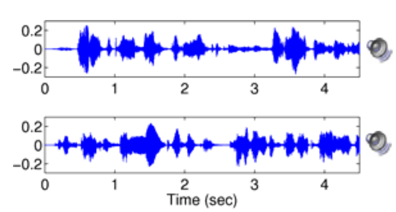
TDOA estimations
with basis vectors



Sorting
Clustering



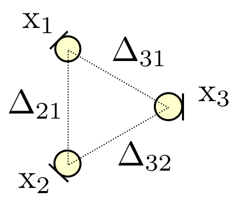




NTT

Multiple microphone pairs

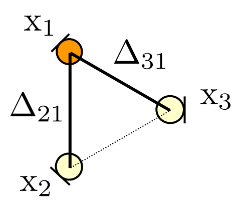
$\frac{M(M-1)}{2}$ pairs for M microphones



Redundant, for example

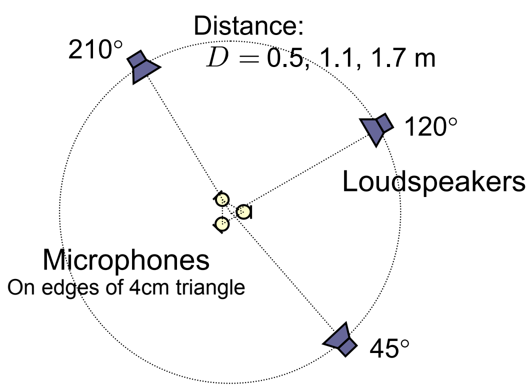
$$
\begin{aligned}
\Delta_{32} &= \tau_3 - \tau_2 \\
&= (\tau_3 - \tau_1) - (\tau_2 - \tau_1) \\
&= \Delta_{31} - \Delta_{21}
\end{aligned}
$$

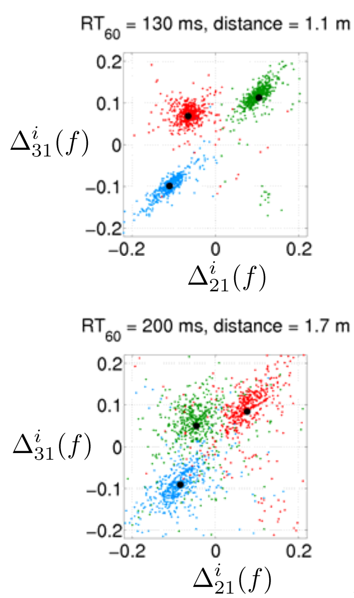
Considers only $M-1$ pairs with a reference microphone



Clustering TDOA estimations in an $M-1$ dimensional space for permutation alignment

3-source 3-microphone case

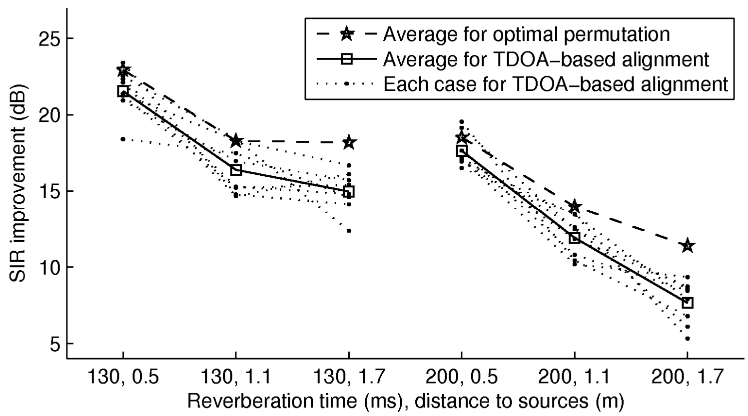




- Reverberation time: RT_{60} = 130, 200 ms

Experimental results

Setup shown in the previous slide

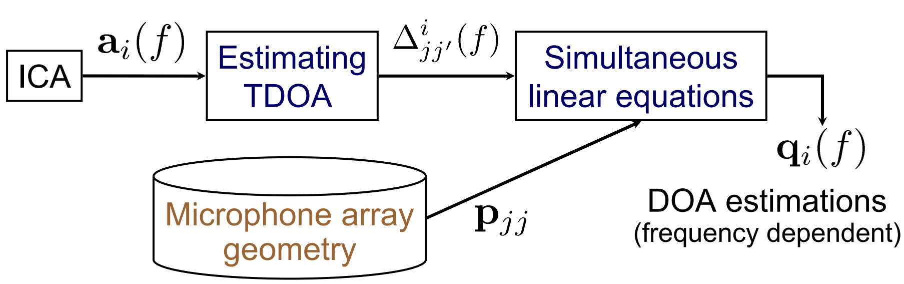


The reverberation time and the distance from sources to microphones affect the separation performance.

NTT

Estimating DOAs of sources

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



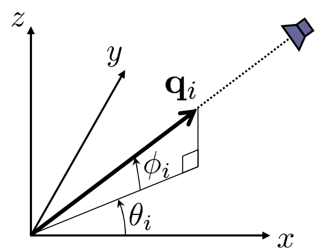
Needs to know: semi-blind

NTT

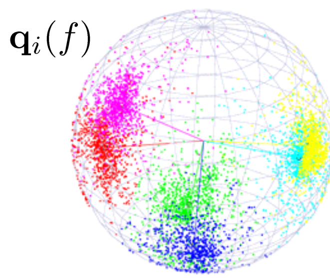
DOA: definition and example

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



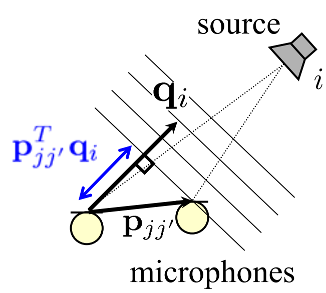
Linear equation for a pair

- Path difference

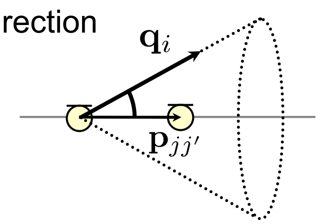
$$\mathbf{p}_{jj'}^T \mathbf{q}_i = \Delta_{jj'}^i \cdot v$$

velocity

TDOA



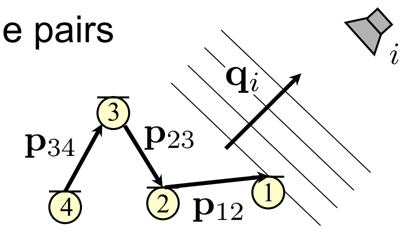
- Cone ambiguity
  - Need more pairs to specify a direction



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

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

Section 3 Summary

- Frequency-dependent TDOA
  - Estimated with observations $\mathbf{x}(n, f)$
  - Estimated with basis vectors $\mathbf{a}_i(f)$
- Permutation alignment
  - Sorting or clustering TDOAs estimated with $\mathbf{a}_i(f)$
  - Effective in low reverberant conditions or when the distance from source to microphone is small
- Direction of arrival (DOA) estimation
  - Together with information on microphone array geometry

Outline

- Part I

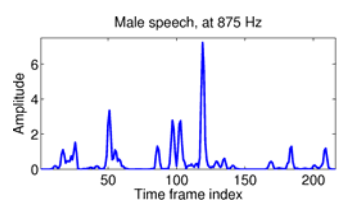
-------- Coffee break --------

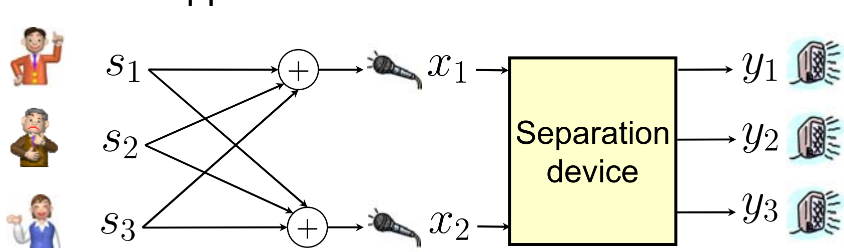
- Part II
  1. Permutation and scaling problems
  2. Mutual dependence of separated signals across frequencies
  3. Time-difference-of-arrival (TDOA) and direction-of-arrival (DOA) estimation
  4. Sparse source separation

NTT 59

Sparse source separation

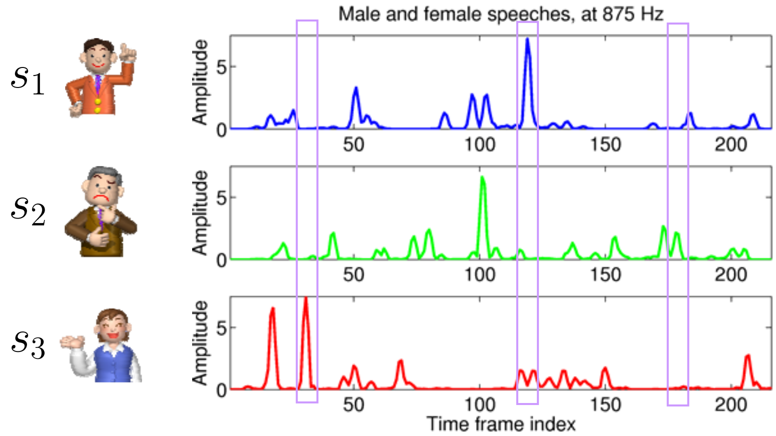
- Sparse source
  - Close to zero most of the time
  - Frequency-domain speech
- Time-variant filtering
  - ICA: time-invariant
- Can be applied even to underdetermined case





NTT 60

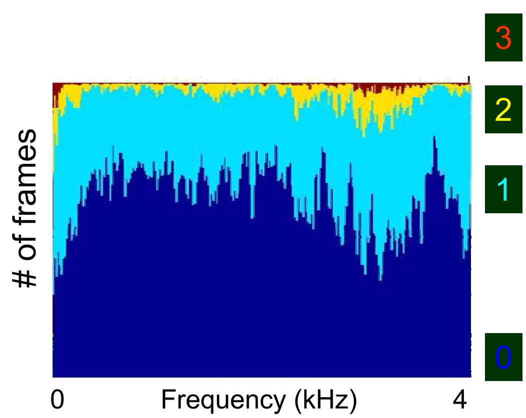
Sparseness



Most samples are close to zero

⇒ ASSUMPTION: At most one source is loud at the same time

Number of sources in each frame



- Anechoic
- Male-male-female
- Threshold: –20dB from the maximum

More than two sources are rarely active simultaneously

Time-frequency masking

- Popular separation method for sparse sources
  [references]
  - Time-variant filtering

$$\begin{bmatrix} y_1(n,f) \\ y_2(n,f) \\ y_3(n,f) \end{bmatrix} = \begin{bmatrix} \mathcal{M}_1(n,f) \\ \mathcal{M}_2(n,f) \\ \mathcal{M}_3(n,f) \end{bmatrix} x_1(n,f)$$

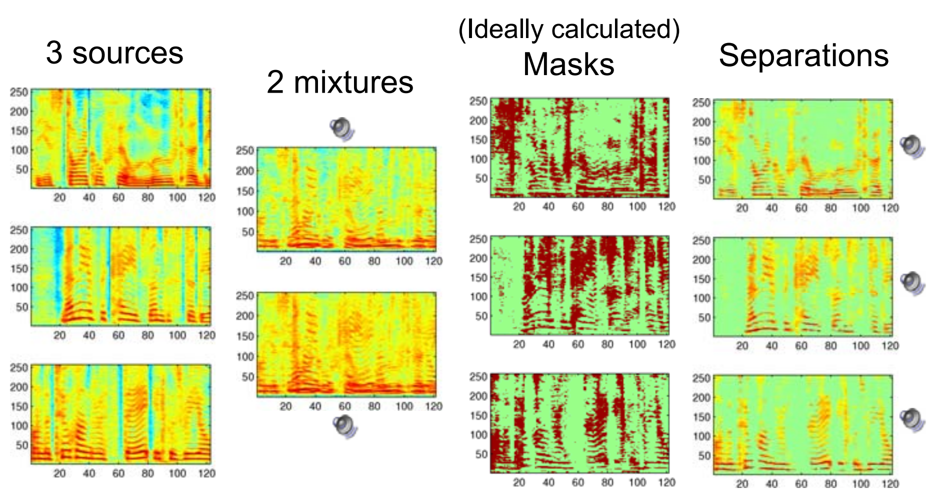
Masks defined for each time-frequency slot (n,f)

$\leftrightarrow$ Time-invariant filtering (ICA)

$$\begin{bmatrix} y_1(n,f) \\ y_2(n,f) \end{bmatrix} = \begin{bmatrix} w_{11}(f) & w_{12}(f) \\ w_{21}(f) & w_{22}(f) \end{bmatrix} \begin{bmatrix} x_1(n,f) \\ x_2(n,f) \end{bmatrix}$$

Depend only on frequency

Time-frequency masking



Mask design

- Time-frequency mask

$$0 \leq \mathcal{M}_i(n, f) \leq 1$$

  - 0 if the i-th source is inactive at (n, f)
  - 1 if the i-th source is active at (n, f)
- How to design in a blind scenario?
  - Utilize TDOA estimations

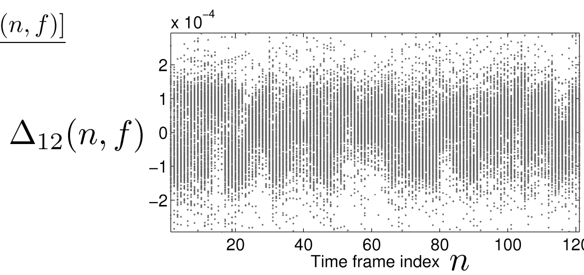
$$\Delta_{12}(n, f) = \frac{\arg[x_1(n, f)/x_2(n, f)]}{-2\pi f}$$

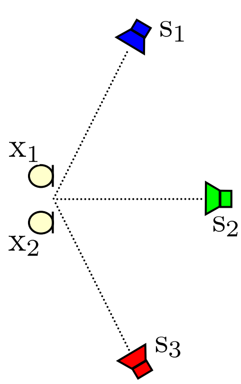
    - Discussed in the permutation alignment method

NTT

Frequency-dependent TDOA

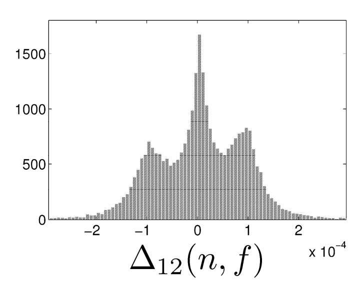
$$\Delta_{12}(n, f) = \frac{\arg[x_1(n, f)/x_2(n, f)]}{-2\pi f}$$





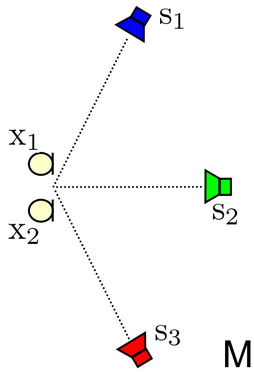
Histogram

3 peaks

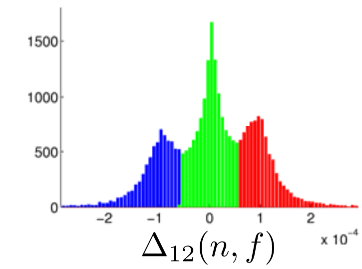


NTT

Frequency-dependent TDOA



Clustering

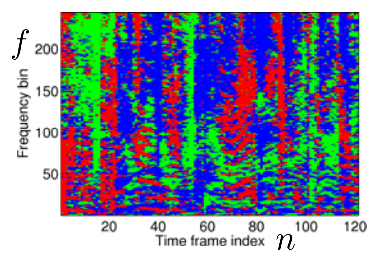


Mask designed

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

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

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

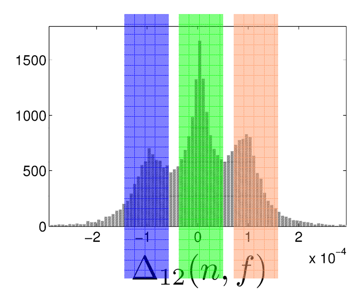


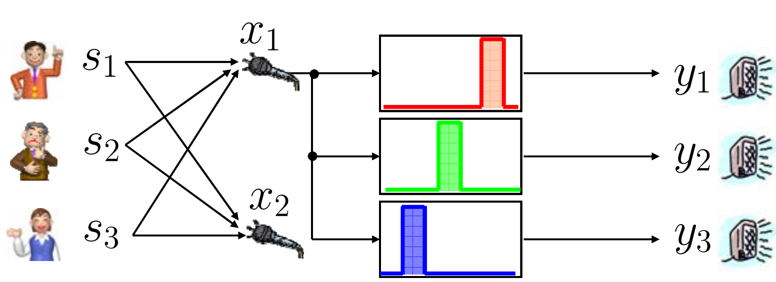
NTT

67

T-F masking separation using TDOA

1. TDOA estimations from microphone observations
2. Clustering & T-F masking
3. Sound output

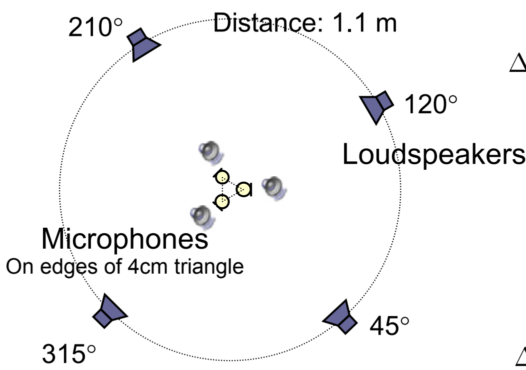




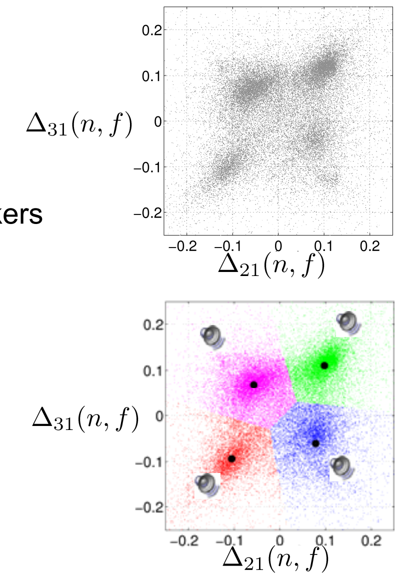
NTT

68

4-source 3-microphone case



• Reverberation time: RT_{60} = 130 ms



Section 4 Summary

- Sparseness
  - well recognized in the frequency domain
- Separation with time-frequency masking
  - Time-variant filtering
  - Applicable to underdetermined case
- T-F mask design
  - Based on frequency-dependent TDOA estimated with observations $\mathbf{x}(n, f)$

Concluding remarks

- Frequency-domain approach for the separation of speech/audio sounds mixed in a real room
- If the situation is properly setup, the source separation task can be performed effectively
  - Sound sources are mostly active for the observation time period
  - Source positions are not changed
- The real challenge lies in a situation where the above conditions are not satisfied
  - Short utterance, unknown number of sources, …

NTT

References

ICA and BSS books

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

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 Laheld, 1996]
- FastICA [Hyvärinen et al., 2001]

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., 2007]

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., 2007]

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]

Blind dereverberation for speech signals

[Nakatani et al., 2007], [Delcroix et al., 2007], [Kinoshita et al., 2006]

Scaling adjustment to microphone observations

[Cardoso, 1998], [Murata et al., 2001], [Matsuoka and Nakashima, 2001], [Takatani et al., 2004]

Linear estimation

[Kailath et al., 2000]

Time difference of arrival (TDOA)

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

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*, volume III, pages 81–84.

[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.

[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, 1998] Cardoso, J.-F. (1998). Multidimensional independent component analysis. In *Proc. ICASSP 1998*, volume 4, pages 1941–1944.

[Cardoso and Laheld, 1996] Cardoso, J.-F. and Laheld, B. H. (1996). Equivariant adaptive source separation. *IEEE Trans. Signal Processing*, 44(12):3017–3030.

[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.

[Delcroix et al., 2007] Delcroix, M., Hikichi, T., and Miyoshi, M. (2007). Precise dereverberation using multi-channel linear prediction. *IEEE Trans. Audio, Speech and Language Processing*, 15(2):430–440.

[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.

[Kailath et al., 2000] Kailath, T., Sayed, A. H., and Hassibi, B. (2000). *Linear Estimation*. Prentice Hall.

[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.

[Kinoshita et al., 2006] Kinoshita, K., Nakatani, T., and Miyoshi, M. (2006). Spectral subtraction steered by multi-step forward linear prediction for single channel speech dereverberation. In *Proc. ICASSP 2006*, volume I, pages 817–820.

[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.

[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 near-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.

[Nakatani et al., 2007] Nakatani, T., Kinoshita, K., and Miyoshi, M. (2007). Harmonicity-based blind dereverberation for single-channel speech signals. *IEEE Trans. Audio, Speech and Language Processing*, 15(1):80–95.

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

[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*, pages 2165–2173.

[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.