Topic 15

Multi-channel Source Separation

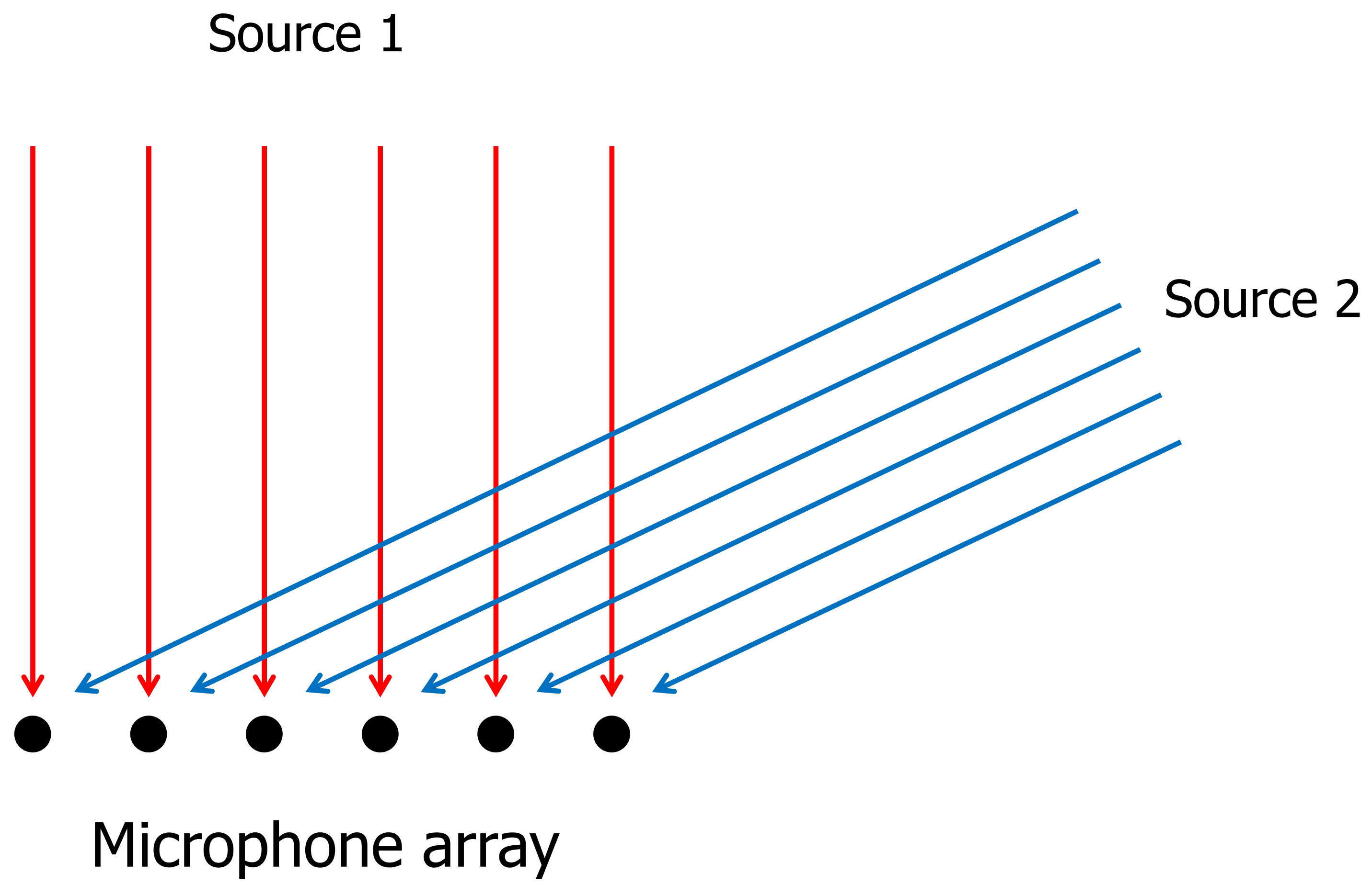
Problem Classification

- Number of sources: N
- Number of channels: M

- Over-determined: N<M
  - Beamforming, ICA
- Determined: N=M
  - ICA
- Under-determined: N>M
  - DUET

- Single-channel: M=1

Beamforming



Discussions

- Advantages
  - Simple, robust
- Disadvantages
  - Need many channels
  - Need to know the direction of the target source

Independent Component Analysis

- Instantaneous mixing model (ignoring delay)
- #sources = #channels

$$x_1(t) = a_{11}s_1(t) + a_{12}s_2(t)$$
$$x_2(t) = a_{21}s_1(t) + a_{22}s_2(t)$$

- Matrix notation of random variables

$$\boldsymbol{x} = A\boldsymbol{s}$$

Channel signal

Mixing matrix

Source signal

- Problem: estimate $\boldsymbol{s}$ from $\boldsymbol{x}$, where A is unknown

ICA Assumptions

- Source signals $\boldsymbol{s}$ are non-Gaussian
- Sources are independent to each other

Mixing process: $\boldsymbol{x} = A\boldsymbol{s}$

- Let demixing matrix $W = A^{-1}$, and $\boldsymbol{w}^T$ be one row, then a separated source is

$$y = \boldsymbol{w}^T\boldsymbol{x} = \boldsymbol{w}^T A\boldsymbol{s} = \boldsymbol{z}^T\boldsymbol{s}$$

- We hope that $\boldsymbol{z}^T$ has only one nonzero element whose value is 1.

Key Idea of ICA

$$y = \boldsymbol{w}^T \boldsymbol{x} = \boldsymbol{w}^T A \boldsymbol{s} = \boldsymbol{z}^T \boldsymbol{s}$$

- **Central Limit Theorem**: the sum of many independent random variables tends toward a Gaussian distribution.
- y is more Gaussian than $\boldsymbol{s}$, unless $\boldsymbol{z}^T$ only has one nonzero element, which is what we want.

- Find $\boldsymbol{w}^T$ such that y is most non-Gaussian!
  - Various ways to define non-Gaussianity

Discussions

- Advantages
  - Elegant
- Disadvantages
  - Need more (or equal) channels than sources
  - The independence assumption is too strong sometimes

DUET

- Degenerate Unmixing Estimation Technique

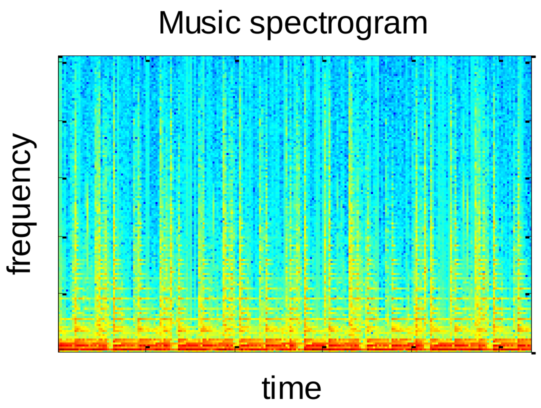
[Yilmaz, Rickard' 04]

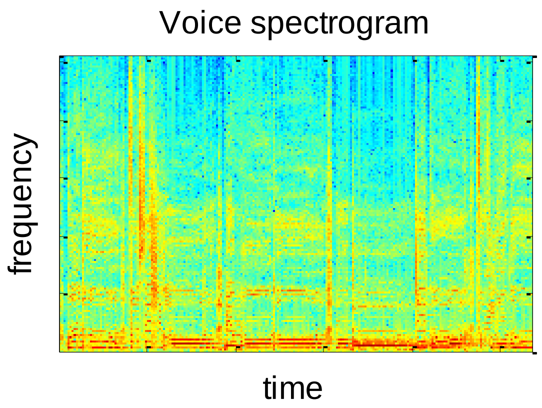
- Separates N>2 sources from 2 mixtures

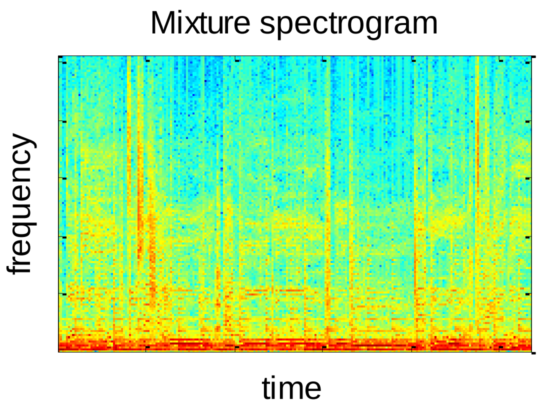
- Assumes that spectrograms of sources do not overlap much

- Binary time-frequency masking

Time-frequency Masking







- Multiply the mixture spectrogram with a mask (a matrix of the same size as the spectrogram) to get the source spectrogram
  - Soft mask: mask takes real values
  - Binary mask: mask takes binary values

W-disjoint orthogonal (W-DO)

- The support (non-zero time-freq points) of different sources do not overlap w.r.t. window W
  - Sources are of different frequencies
  - Sources are active at different times
- Not realistic in practice

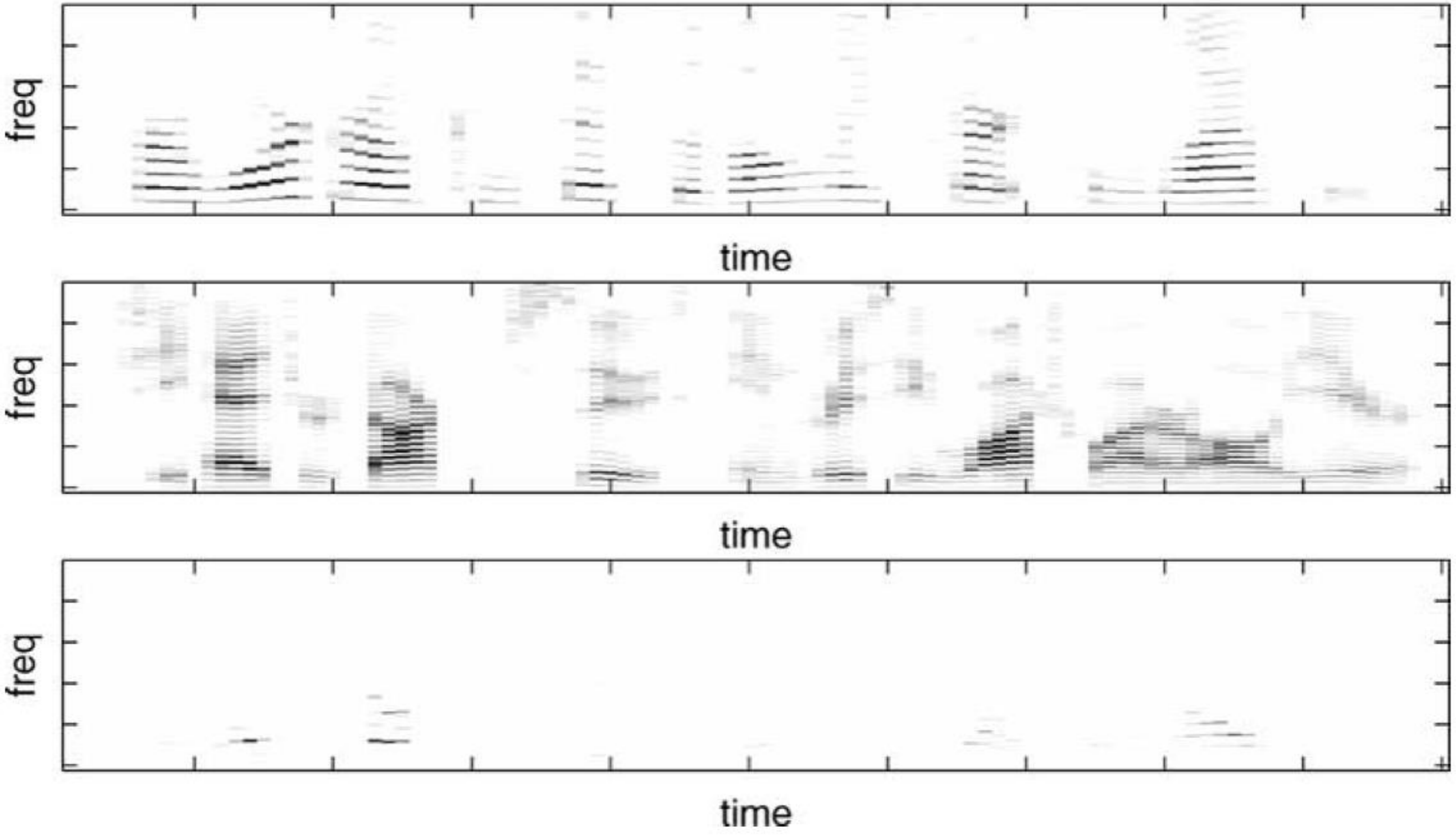
- Approximately W-DO: Only one source has strong energy at each time-freq point.
  - Binary masking would give pretty good separation

Speech signals are approximately W-DO

Speech 1

Speech 2

Multiplication



Measuring W-Disjoint Orthogonality

A mask for target source

Target source spectrogram

All interferences spectrogram

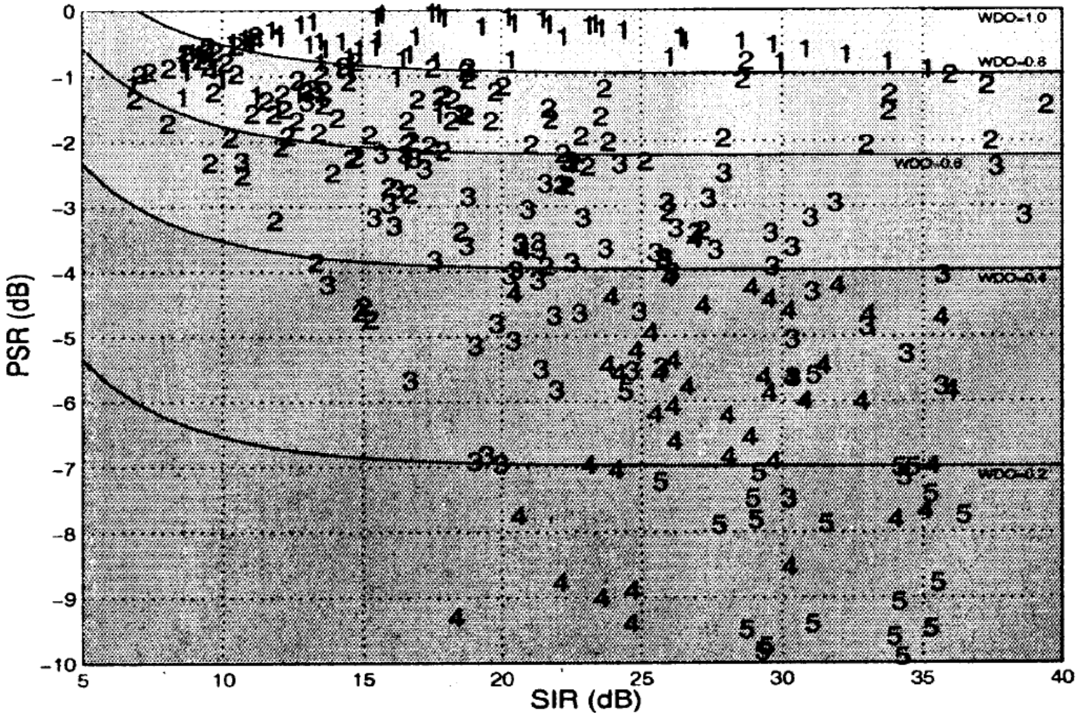
$$\mathrm{WDO}_M := \frac{\|M(\tau,\omega)\hat{s}_j(\tau,\omega)\|^2 - \|M(\tau,\omega)\hat{y}_j(\tau,\omega)\|^2}{\|\hat{s}_j(\tau,\omega)\|^2}$$

time

frequency

- Ideal Binary Mask (IBM): takes 1 in the time-freq points where the target source is at least x dB louder than interferences, and 0 otherwise.
- WDO_IBM measures W-Disjoint orthogonality.

How large WDO do we need?



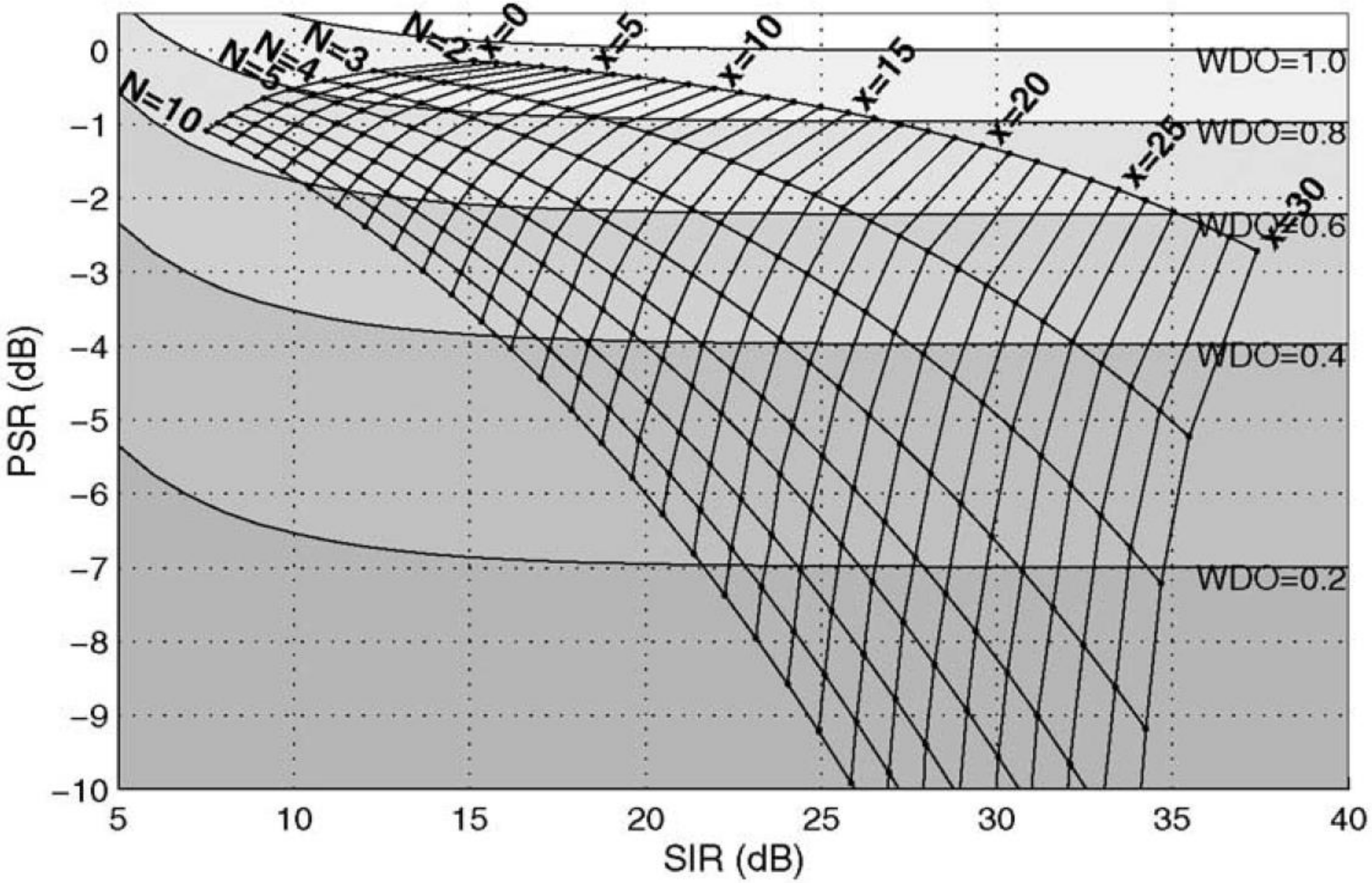
rating	meaning
1	perfect
2	minor artifacts or interference
3	distorted but intelligible
4	very distorted and barely intelligible
5	not intelligible

WDO = PSR – PSR/SIR

- PSR: preserved signal ratio
- SIR: signal to interference ratio

Speech signals are approximately W-DO

N: #sources; x: energy threshold to derive IBM



Anechoic Mixing Model

$$x_k(t) = \sum_{j=1}^{N} \underbrace{a_{kj}}_{\text{Attenuation coefficients}} s_j(t - \underbrace{\delta_{kj}}_{\text{Time delays}}), \qquad k = 1,2$$

- Without loss of generality, we can set $a_{1j} = 1$ and $\delta_{1j} = 0$ for all $j = 1, \ldots, N$. And rename a_{2j} as a_j and δ_{2j} as δ_j, which are relative attenuation and time delay.
- Take STFT:

$$\begin{bmatrix} \hat{x}_1(\tau,\omega) \\ \hat{x}_2(\tau,\omega) \end{bmatrix} = \begin{bmatrix} 1 & \ldots & 1 \\ a_1 e^{-i\omega\delta_1} & \ldots & a_N e^{-i\omega\delta_N} \end{bmatrix} \begin{bmatrix} \hat{s}_1(\tau,\omega) \\ \vdots \\ \hat{s}_N(\tau,\omega) \end{bmatrix}$$

How to derive the mask?

$$\begin{bmatrix} \hat{x}_1(\tau,\omega) \\ \hat{x}_2(\tau,\omega) \end{bmatrix} = \begin{bmatrix} 1 & \dots & 1 \\ a_1 e^{-i\omega\delta_1} & \dots & a_N e^{-i\omega\delta_N} \end{bmatrix} \begin{bmatrix} \hat{s}_1(\tau,\omega) \\ \vdots \\ \hat{s}_N(\tau,\omega) \end{bmatrix}$$

- When sources are W-DO, each t-f point contains only one source, and its relative attenuation and delay correspond to those of that source!

$$R_{21}(\tau,\omega) := \frac{\hat{x}_2(\tau,\omega)}{\hat{x}_1(\tau,\omega)} = a_j e^{-i\delta_j\omega}$$

If only source j is active at (τ, ω)

Group T-F Points

- T-F points dominated by the same source have very similar relative attenuation and delay

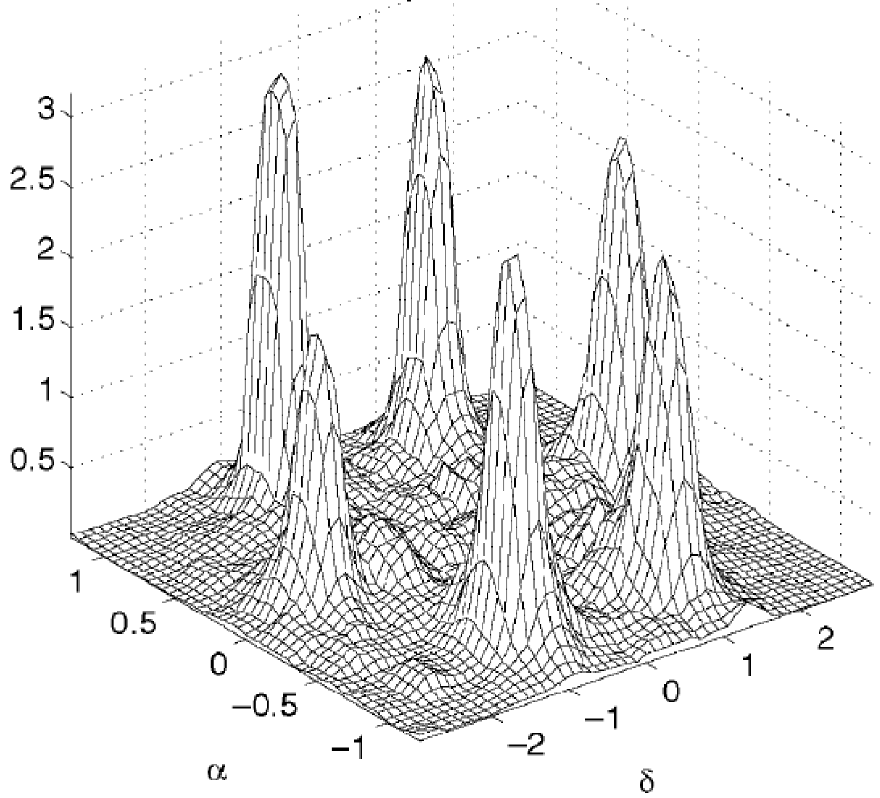
$$\tilde{a}(\tau,\omega) := |R_{21}(\tau,\omega)|$$
$$\tilde{\delta}(\tau,\omega) := -\frac{1}{\omega}\angle R_{21}(\tau,\omega)$$

- Plot a 2-D histogram

- Here we use symmetric attenuation for better numerical results

$$\tilde{\alpha}(\tau,\omega) \coloneqq \tilde{a}(\tau,\omega) - \frac{1}{\tilde{a}(\tau,\omega)}$$

6 sources

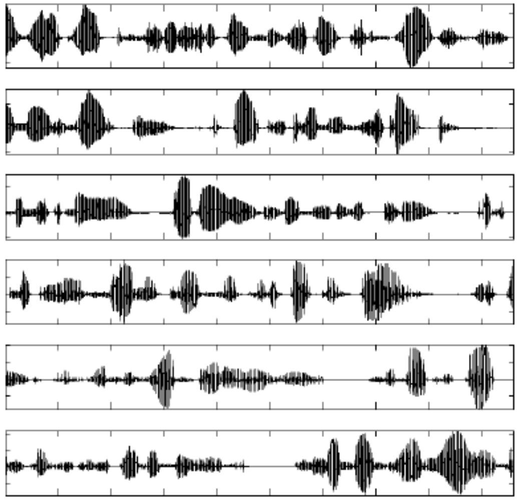


DUET Algorithm

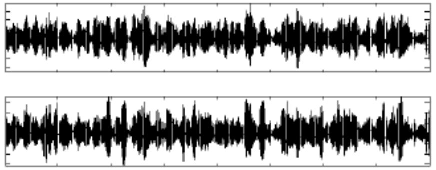
- 1) STFT on both channels
- 2) calculate DUET parameters (i.e., relative symmetric attenuation and delay) for each T-F point
- 3) construct a 2-D histogram and locate peaks, where each peak will correspond to a source
- 4) for each peak, construct a binary mask by collecting T-F points whose DUET parameters are close to the peak
- 5) apply the mask to the mixture and do inverse-STFT

Experiments

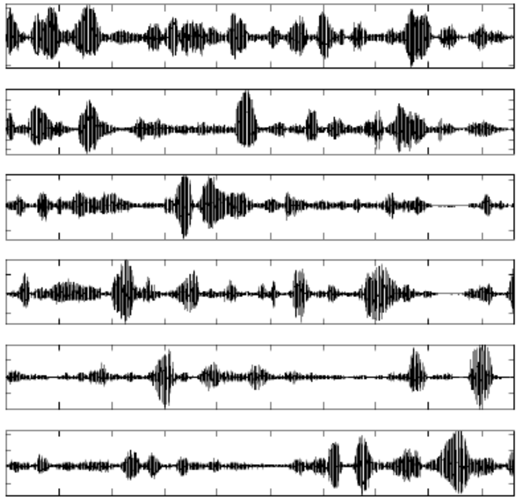
Speech sources



Artificial mixtures using anechoic mixing model



Separated by DUET



source	SIR in (dB)	SIR out (dB)	WDO DUET	WDO 0dB
s_1	-7.29	5.92	0.57	0.80
s_2	-7.29	5.24	0.55	0.78
s_3	-5.08	6.60	0.62	0.81
s_4	-9.29	5.35	0.56	0.69
s_5	-5.03	7.06	0.63	0.81
s_6	-9.28	5.47	0.55	0.66

Experiments in Real Environments

Anechoic room

test	SIR in (dB)	SIR out (dB)	WDO DUET	WDO 0dB
M1 0^o	-2.72	13.67	0.88	0.90
F1 90^o	-2.05	7.96	0.80	0.93
M2 180^o	-4.37	13.32	0.84	0.87
F1 0^o	-9.77	7.97	0.62	0.76
M1 60^o	-4.30	7.16	0.67	0.86
F2 90^o	-3.77	5.99	0.68	0.91
M2 120^o	-5.60	7.05	0.65	0.85
F3 180^o	-8.59	8.53	0.65	0.82

Echoic room (reverberation time ~500ms)

test	SIR in (dB)	SIR out (dB)	WDO DUET	WDO 0dB
M1 0^o	-5.20	5.38	0.40	0.81
M2 90^o	0.07	4.33	0.56	0.91
F1 180^o	-4.48	6.03	0.49	0.87

Histograms

Anechoic room

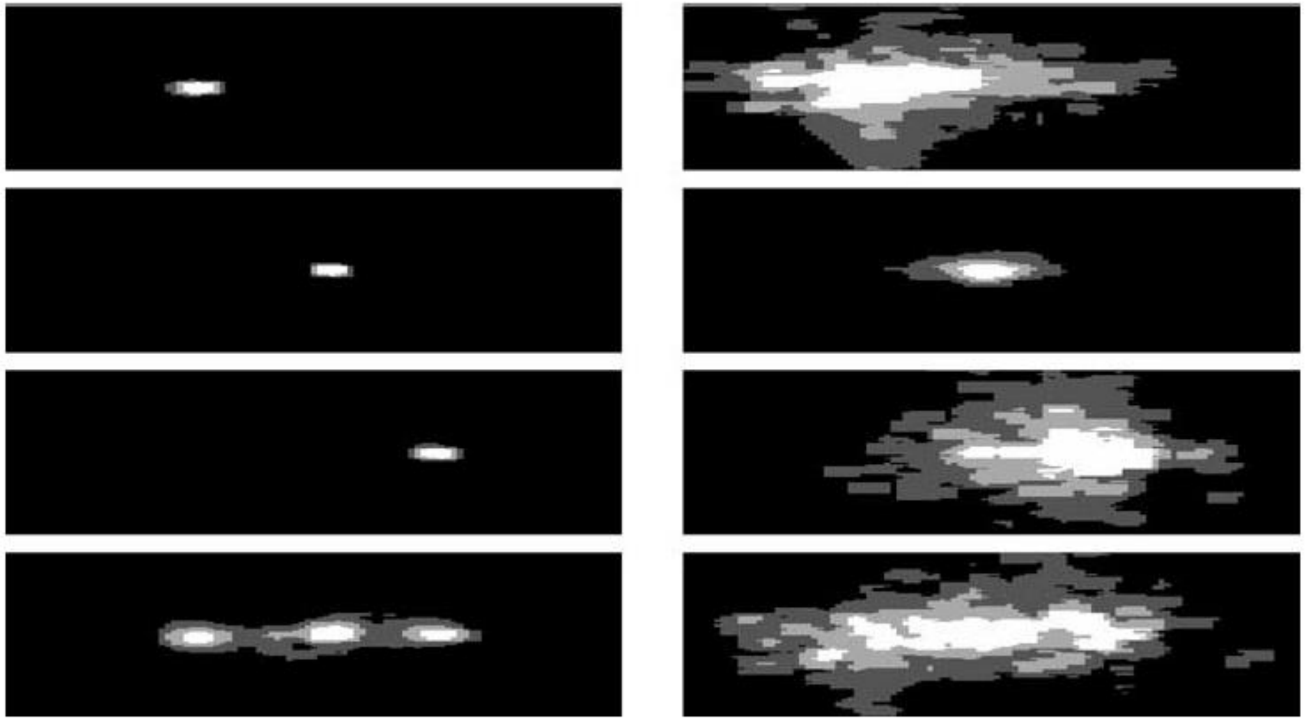
Echoic room (reverberation time ~500ms)

Source 1

Source 2

Source 3

Mixture



Discussions

- Advantages
  - Blind
  - Simple
  - Works pretty well for speech sources in anechoic rooms

- Disadvantages
  - Would fail if sources overlap much
  - Can’t deal with reverberation well