UNIVERSITY OF SURREY

UDRC Summer School, Edinburgh, 23-27 June, 2014

Convolutive Source Separation & Demonstrations

Wenwu Wang
Senior Lecturer in Signal Processing
Centre for Vision, Speech and Signal Processing
Department of Electronic Engineering
University of Surrey, Guildford

w.wang@surrey.ac.uk
http://personal.ee.surrey.ac.uk/Personal/W.Wang/

25/06/2014

Outline

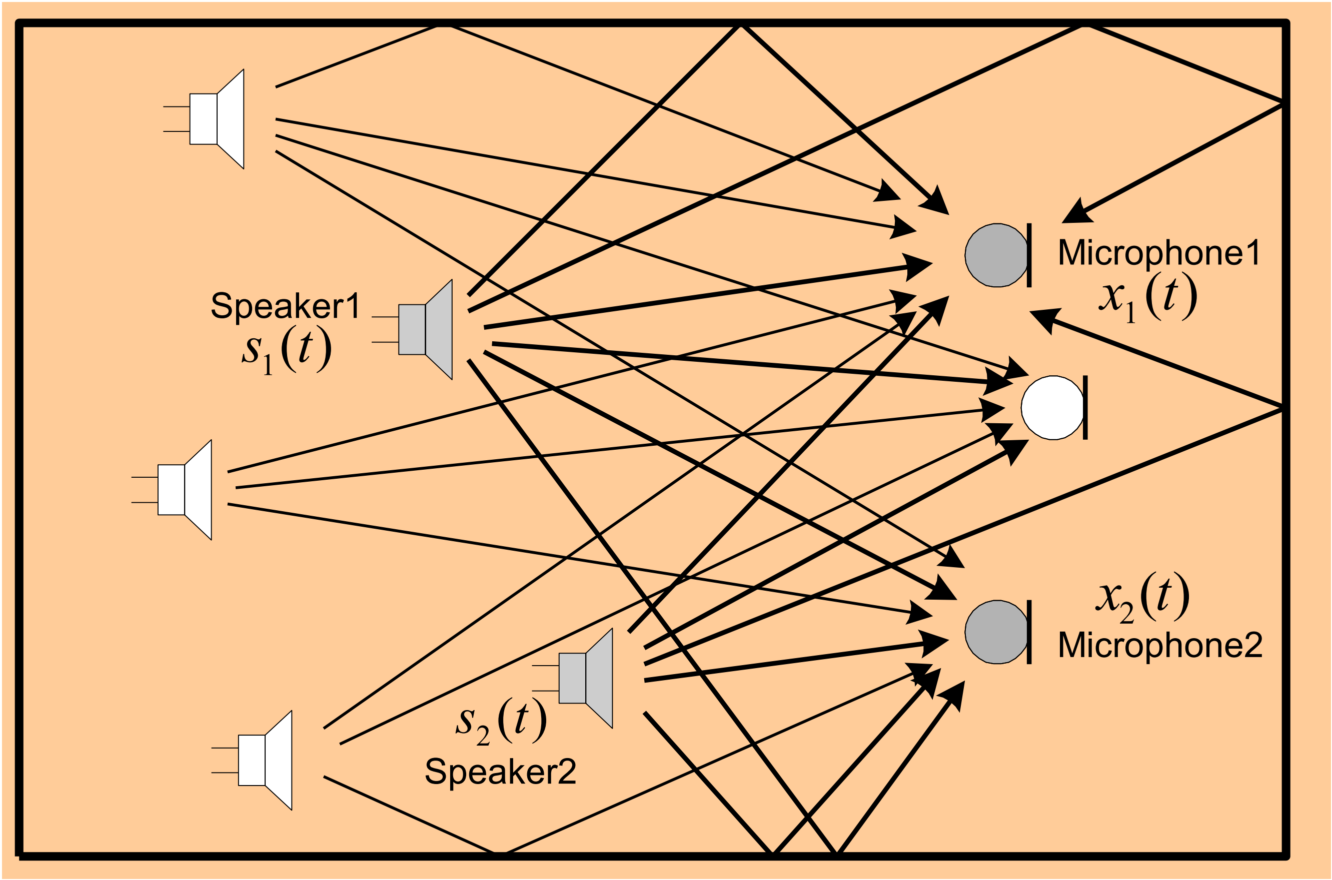
- Speech separation and cocktail party problem
- Convolutive ICA and constrained convolutive ICA
- Computational auditory scene analysis and ideal binary mask
- A hybrid approach of convolutive ICA and binary masking
- Ideal ratio mask and kurtosis ratio
- Soft time-frequency mask: A model based approach for stereo source separation (determined and underdetermined)
- Convolutive source separation of underwater acoustic signals

Speech separation problem

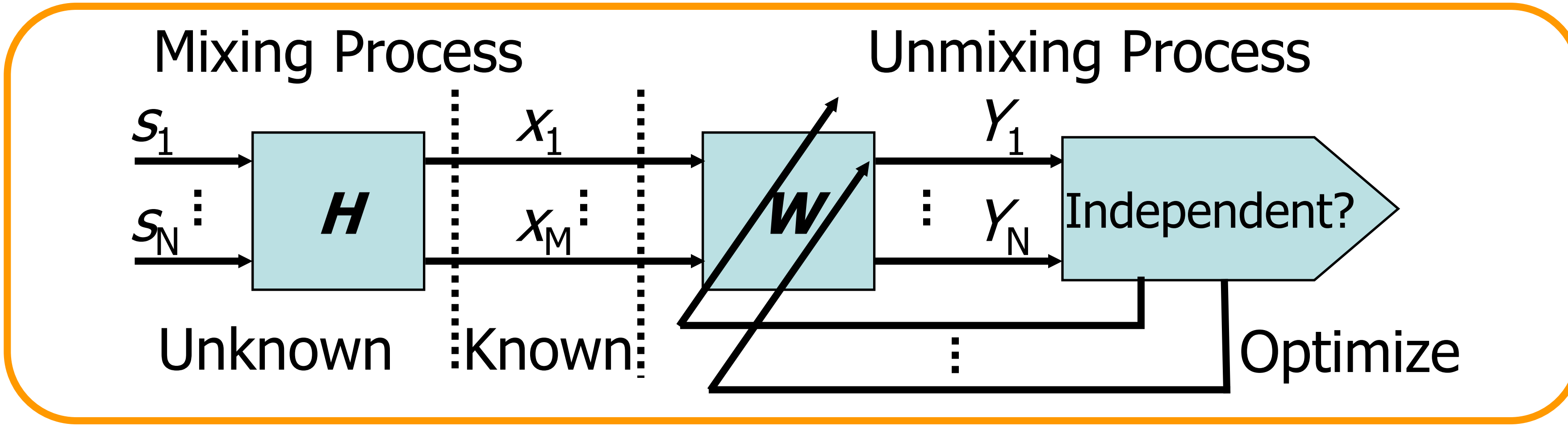
UNIVERSITY OF SURREY

- In a natural environment, target speech is usually corrupted by acoustic interference, creating a speech segregation problem
  - Also known as *cocktail-party problem* (Cherry'53) or *ball-room problem* (Helmholtz, 1863)
  - Speech segregation is critical for many applications, such as automatic speech recognition and hearing prosthesis
- Potential techniques for the speech separation problem
  - Beamforming
  - Blind source separation
  - Speech enhancement
  - Compuational auditory scene analysis
- "No machine has yet been constructed to do just that [solving the cocktail party problem]." (Cherry'57)

Cocktail party problem



Blind source separation & independent component analysis



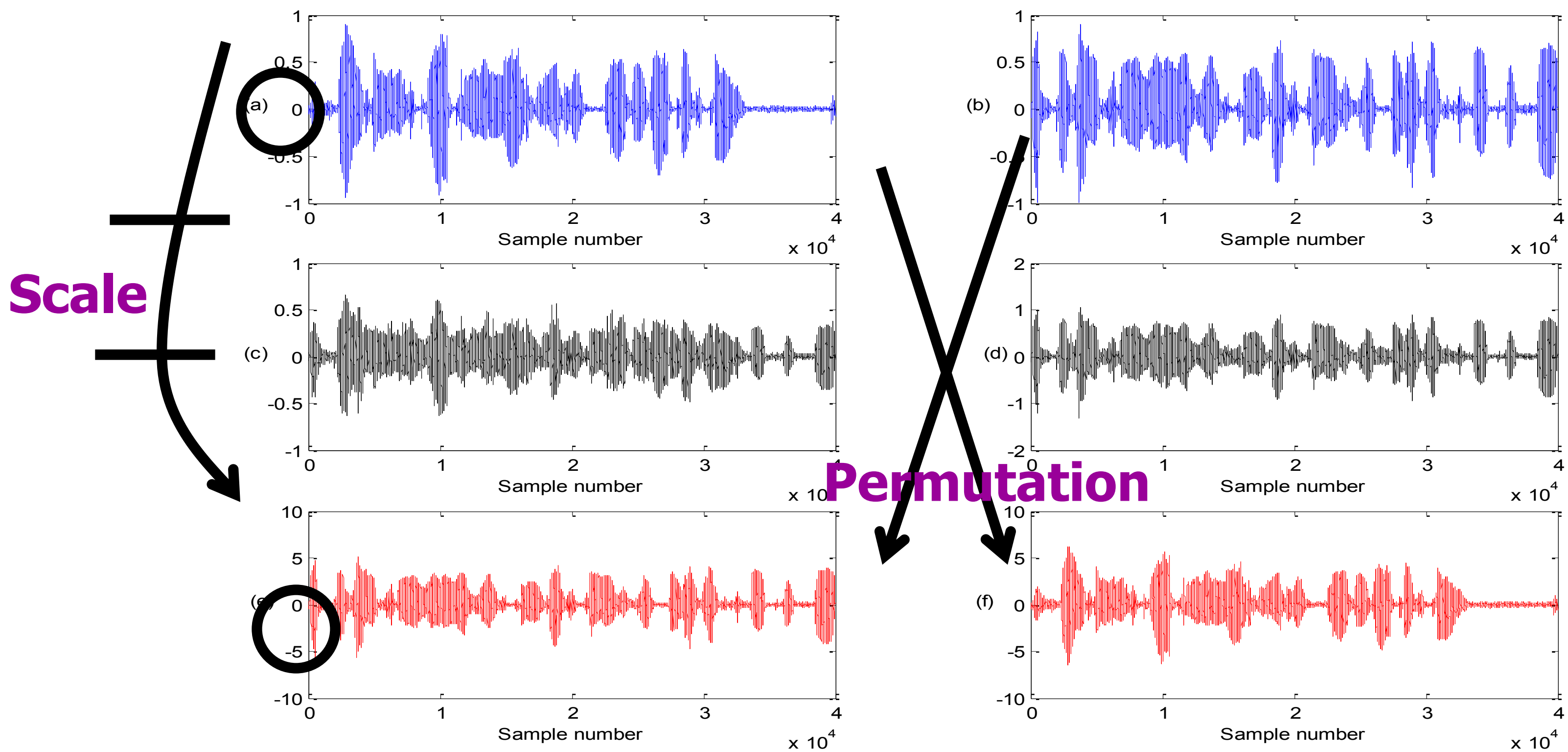
Mixing Model: $\mathbf{x} = \mathbf{H}\mathbf{s}$

De-mixing Model: $\mathbf{y} = \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{H}\mathbf{s} = \mathbf{P}\mathbf{D}\mathbf{s}$

Diagonal Scaling Matrix

Permutation Matrix

Scale and permutation ambiguities: an example



Blind source separation for instantaneous mixtures with the JADE algorithm (SNR=30dB): (a)(b) original sources; (c)(d) mixtures; (e)(f) separated sources

Convolutive BSS: mathematical model

Compact form: $\mathbf{x} = \mathbf{H} * \mathbf{s}$

Convolution

$$\begin{bmatrix} H_{11}(t) & \cdots & H_{1N}(t) \\ \vdots & \ddots & \vdots \\ H_{M1}(t) & \cdots & H_{MN}(t) \end{bmatrix} * \begin{bmatrix} s_1(t) \\ \vdots \\ s_N(t) \end{bmatrix} = \begin{bmatrix} x_1(t) \\ \vdots \\ x_M(t) \end{bmatrix}$$

Expansion form: $x(n) = \sum_{j=1}^{N} \sum_{p=0}^{P-1} h_{ij}(p) s_j(n-p)$

Transform convolutive BSS into the frequency domain

$$\mathbf{x} = \mathbf{H} * \mathbf{s}$$

DFT

$$\begin{bmatrix} x_1(\omega) \\ \vdots \\ x_M(\omega) \end{bmatrix} = \begin{bmatrix} H_{11}(\omega) & \cdots & H_{1N}(\omega) \\ \vdots & \ddots & \vdots \\ H_{M1}(\omega) & \cdots & H_{MN}(\omega) \end{bmatrix} \cdot \begin{bmatrix} s_1(\omega) \\ \vdots \\ s_N(\omega) \end{bmatrix}$$

Convolutive BSS problem → Multiple complex-valued instantaneous BSS problems

De-mixing operation

UNIVERSITY OF SURREY

$$\mathbf{Y}(\omega,t) = \mathbf{W}(\omega)\mathbf{X}(\omega,t) \qquad \forall\,\omega$$

where $\mathbf{W}(\omega) \in C^{N\times M} \qquad \forall\,\omega$

$$\mathbf{Y}(\omega,t) = [Y_1(\omega,t),\ldots,Y_N(\omega,t)]^T \in C^N$$

Parameters in $\mathbf{W}(\omega)$ determined such that $\mathrm{Y}_1(\omega,\mathrm{t}),\ldots,\mathrm{Y}_\mathrm{N}(\omega,\mathrm{t})$ become mutually independent.

L. Parra and C. Spence, "Convolutive blind source separation of nonstationary sources," *IEEE Trans. Speech Audio Process., vol. 8, no. 3, pp.* 320–327, May 2000.

Joint diagonalisation criterion

Exploiting the non-stationarity of signals measured by the cross-spectrum of the output signals,

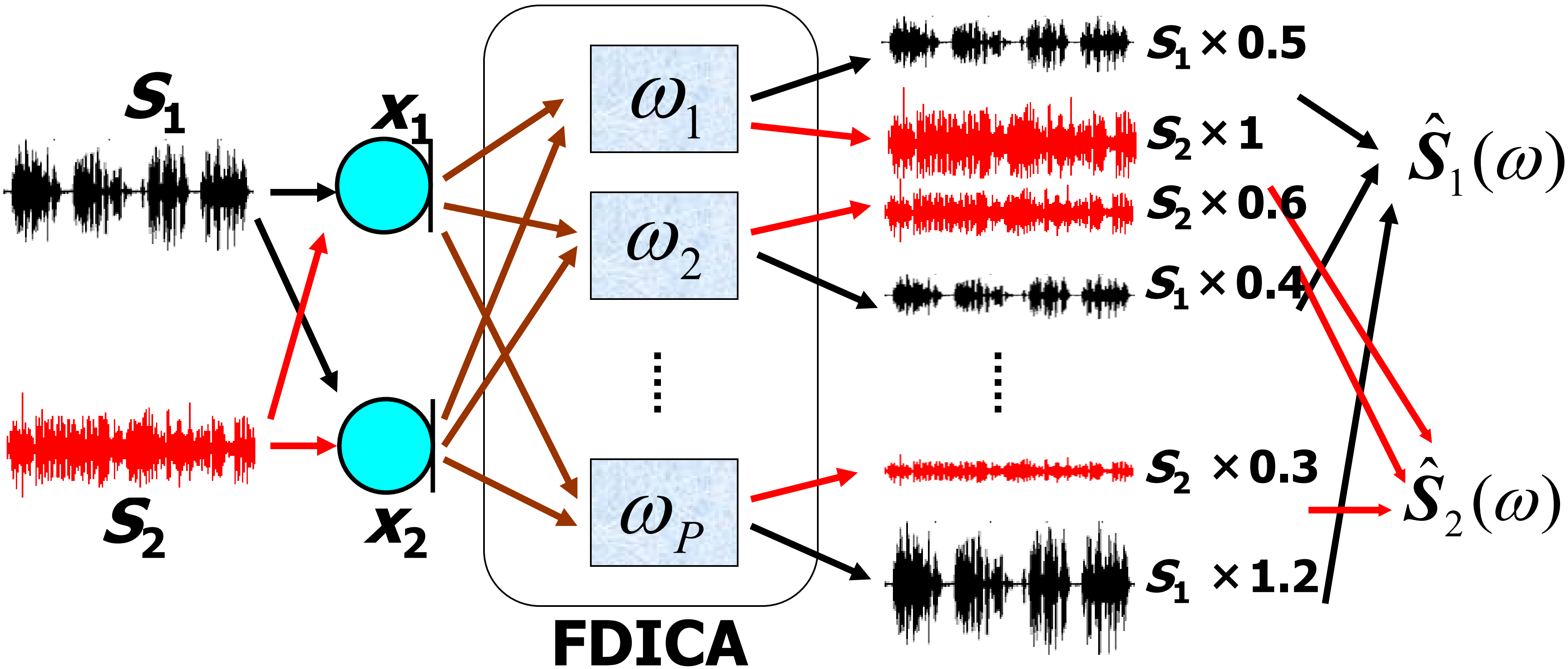
$$\mathbf{R}_Y(\omega,k) = \mathbf{W}(\omega)[\mathbf{R}_X(\omega,k)]\mathbf{W}^H(\omega)$$

Cost Function:

$$J(\mathbf{W}(\omega)) = \arg\min_{\mathbf{W}} \sum_{\omega=1}^{T}\sum_{t=1}^{K} F(\mathbf{W})(\omega,t)$$

where $F(\mathbf{W})(\omega,t) = \left\|\mathbf{R}_Y(\omega,t) - diag[\mathbf{R}_Y(\omega,t)]\right\|_F^2$

Frequency domain BSS & permutation problem



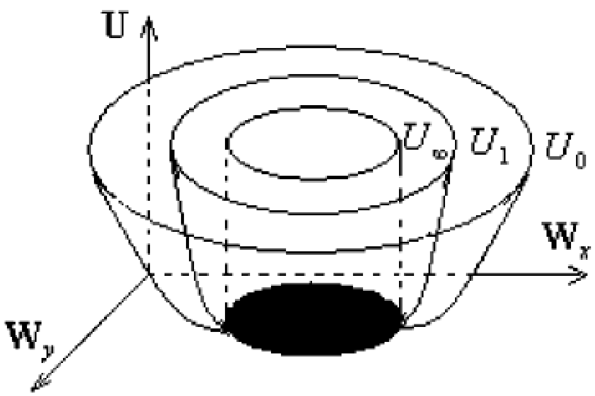
Solutions:

- Beamforming
- Spectral envelope correlation

Constrained convolutive ICA: penalty function approach

UNIVERSITY OF SURREY

- Introducing additional constraints could further improve the separation performance, such as unitary and non-unitary constraints to prevent the degenerate trivial solutions to the unmixing matrix, as shown in Wang et al. (2005).
- Penalty function can be used to convert the constrained optimisation problem into an unconstrained one.



$$\mathfrak{J}(\mathbf{W}(\omega)) = \mathcal{J}(\mathbf{W}(\omega)) + \sum_{i=1}^{r} \kappa_i \mathcal{U}_i(\mathbf{W}(\omega))$$

$$\mathfrak{J}(\mathbf{W}) = \underset{\mathbf{W}}{\operatorname{argmin}} \sum_{\omega=1}^{T} \sum_{k=1}^{K} \{\mathcal{F}(\mathbf{W})(\omega, k) + \kappa \mathcal{G}(\mathbf{W})(\omega, k)\}$$

Sound demonstration

	Sources	Mixtures	Parra&Spence	Our approach
Two speaking sentences artificially mixed together				
A man speaking with TV on				

W. Wang, S. Sanei, and J. A. Chambers, Penalty function based joint diagonalization approach for convolutive blind separation of nonstationary sources, in IEEE Trans. Signal Processing, vol. 53, no. 5, pp. 1654-1669, May 2005.

Auditory scene analysis

- Listeners parse the complex mixture of sounds arriving at the ears in order to form a mental representation of each sound source
- This perceptual process is called auditory scene analysis (Bregman’90)
- Two conceptual processes of auditory scene analysis (ASA):
  - **Segmentation**. Decompose the acoustic mixture into sensory elements (segments)
  - **Grouping**. Combine segments into groups, so that segments in the same group likely originate from the same sound source

Computational auditory scene analysis (CASA)

- Computational auditory scene analysis (CASA) approaches sound separation based on ASA principles
  - Feature based approaches
  - Model based approaches
- CASA has made significant advances in speech separation using monaural and binaural analysis
- CASA challenges
  - Reliable pitch tracking of noisy speech
  - Unvoiced speech
  - Room reverberation

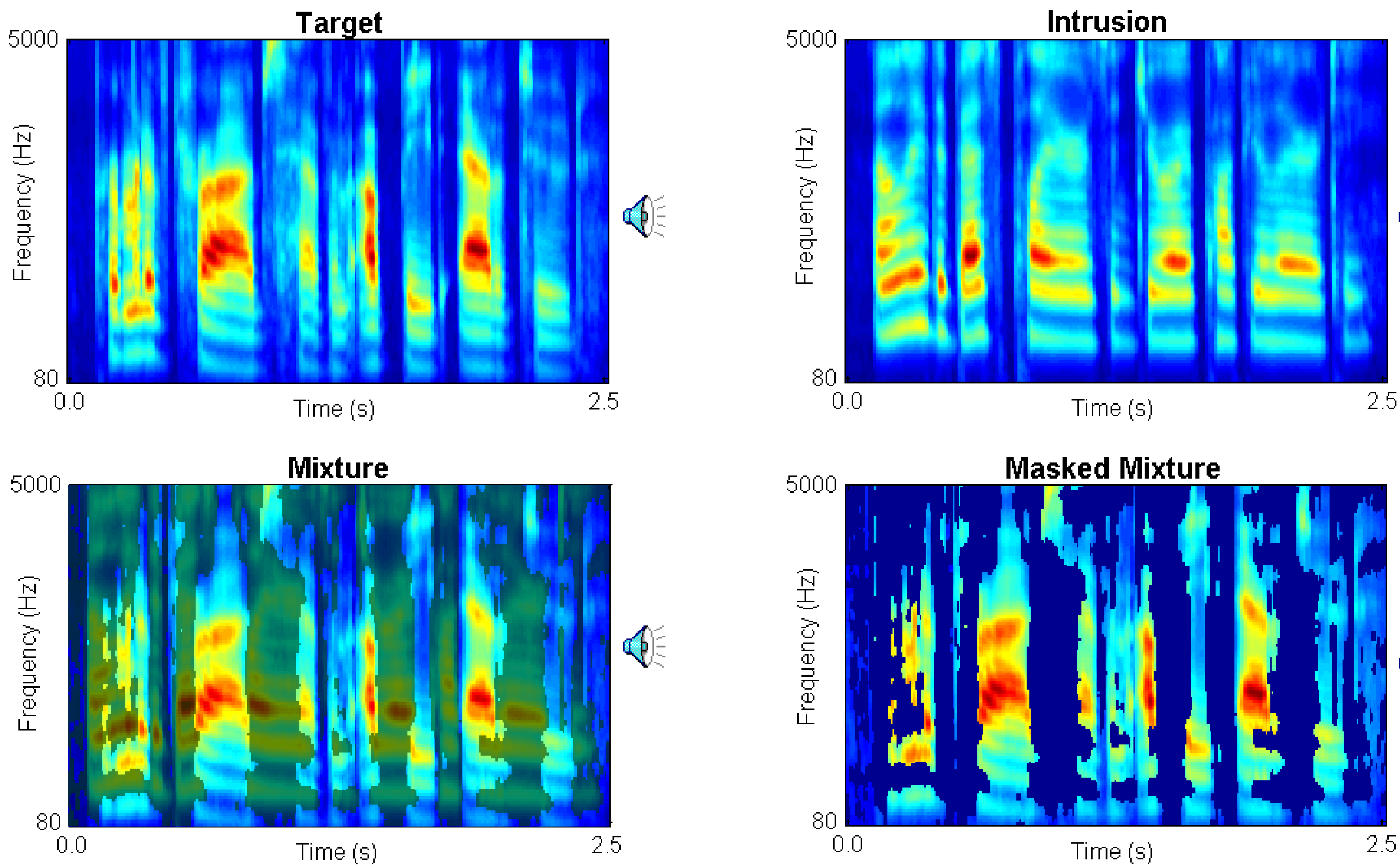
Ideal binary mask (IBM)

- Auditory masking phenomenon: In a narrowband, a stronger signal masks a weaker one
- Motivated by the auditory masking phenomenon, the ideal binary mask has been suggested as a main goal of CASA (D.L. Wang'05)
- The definition of the ideal binary mask

$$m(t,f) = \begin{cases} 1 & \text{if } s(t,f) - n(t,f) \geq \theta \\ 0 & \text{otherwise} \end{cases}$$

  - $s(t, f)$: Target energy in unit (t, f)
  - $n(t, f)$: Noise energy
  - θ: A local SNR criterion in dB, which is typically chosen to be 0 dB
  - Optimality: Under certain conditions the ideal binary mask with $\theta = 0$ dB is the optimal binary mask from the perspective of SNR gain
  - It does not actually separate the mixture!

IBM illustration (after DeLiang Wang)



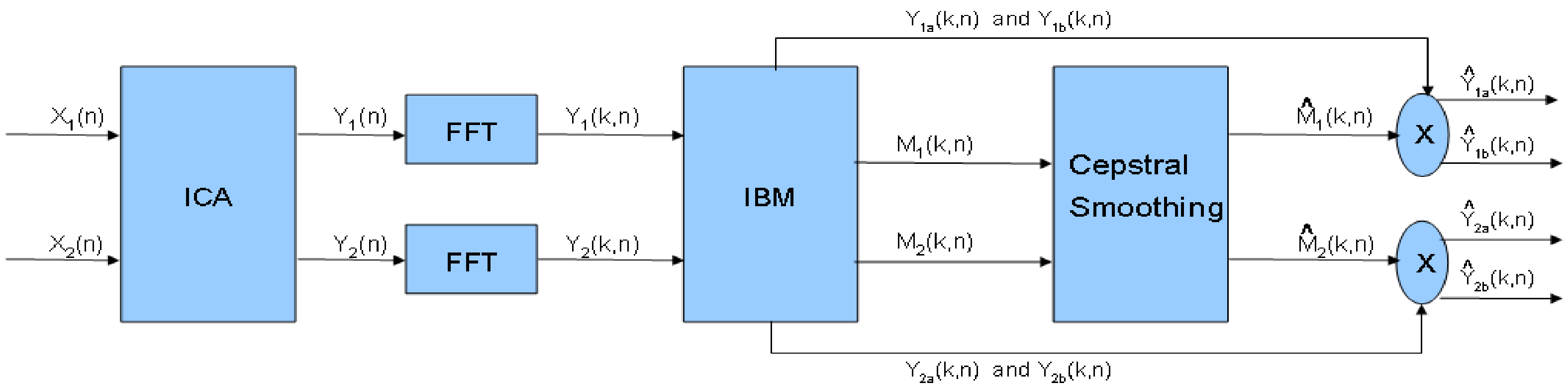
Recent psychophysical tests show that the ideal binary mask results in dramatic speech intelligibility improvements (Brungart et al.'06; Li & Loizou'08)

ICA versus IBM

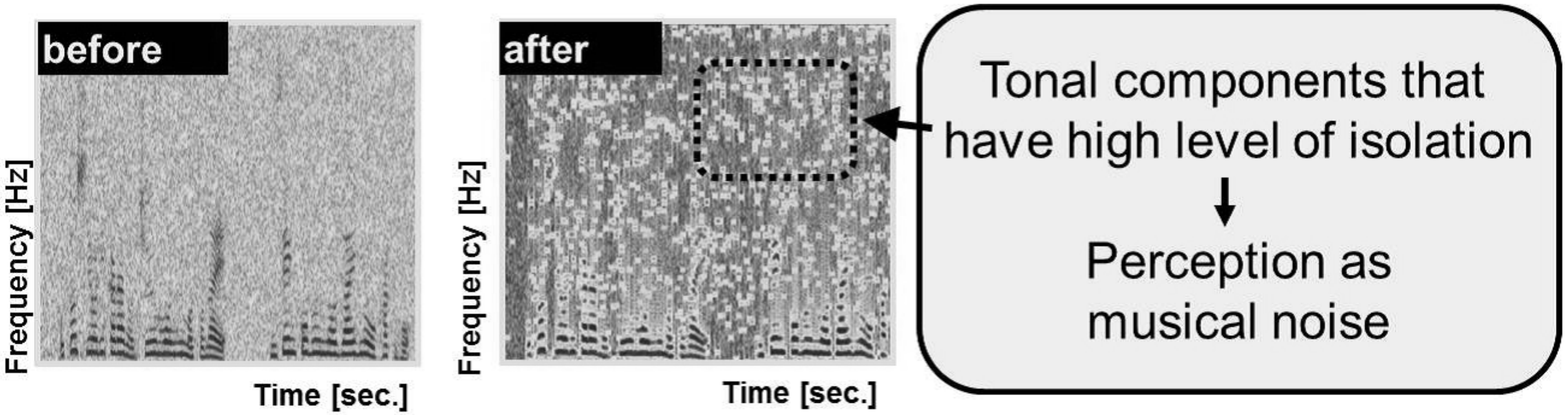
- ICA: Excellent performance if (no or low) reverberation or noise is present in the mixture. For highly reverberant and noisy mixtures, the performance is limited.

- IBM: Excellent performance if both target and background interference are known. Otherwise, the IBM has to be estimated from the acoustic mixture, which however remains an open challenging task!

A multistage approach fusing ICA and IBM



Musical noise



Example of musical noise generation: the input signal on the left plot is corrupted by white Gaussian noise, and the output signal on the right plot is obtained by applying a source separation algorithm to the input. Figure due to Saruwatari and Miyazaki (2014)

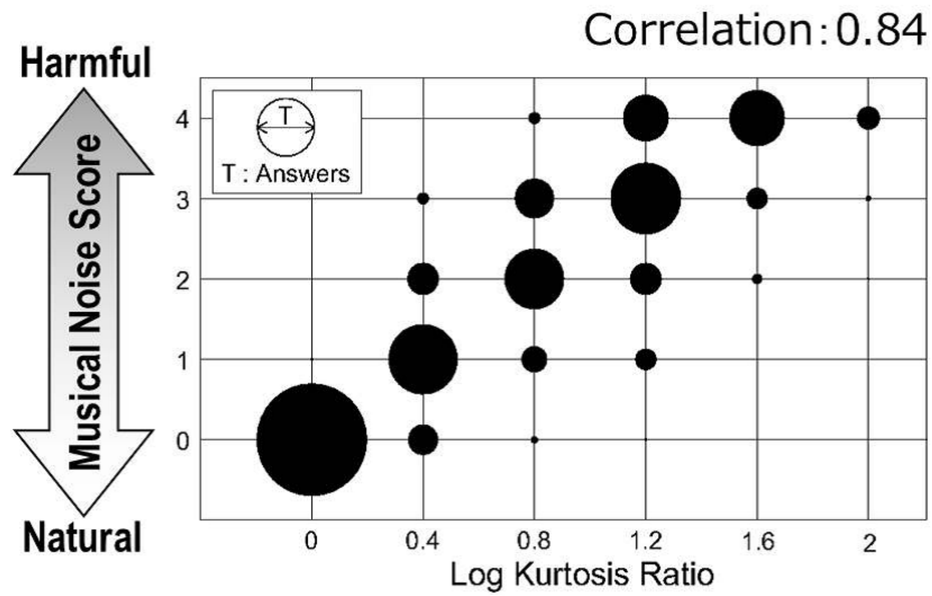
H. Saruwatari and R. Miyazaki, "Statistical analysis and evaluation of blind speech extraction algorithms," in G. Naik and W. Wang (eds), *Blind Source Separation: Advances in Theory, Algorithms and Applications*, Springer, May, 2014

Kurtosis ratio

$$\text{kurtosis ratio} = \frac{\text{kurt}_{\text{proc}}}{\text{kurt}_{\text{org}}},$$

$$\text{kurt} = \frac{\mu_4}{\mu_2^2},$$

$$\mu_m = \int_0^\infty x^m P(x) dx,$$



Relation between kurtosis ratio and human perceptual score of degree of musical noise generation. Figure due to Saruwatari et al. (2014).

H. Saruwatari and R. Miyazaki, “Statistical analysis and evaluation of blind speech extraction algorithms,” in G. Naik and W. Wang (eds), *Blind Source Separation: Advances in Theory, Algorithms and Applications*, Springer, May, 2014

Cepstral smoothing
to mitigate musical noise

Converting mask from spectral domain to cepstral domain:

$$M_i^c(l,m) = DFT^{-1}\{\ln(M_i^f(k,m))|_{k=0,\ldots,K-1}\},$$

Smoothing with various smoothing level to different frequency bands (low smoothing to envelop and pitch band to maintain its structure, more smoothing to other band to remove the artefacts):

$$\overline{M}_i^s(l,m) = \gamma_l \overline{M}_i^s(l,m-1) + (1-\gamma_l)M_i^c(l,m) \quad i = 1,2,$$

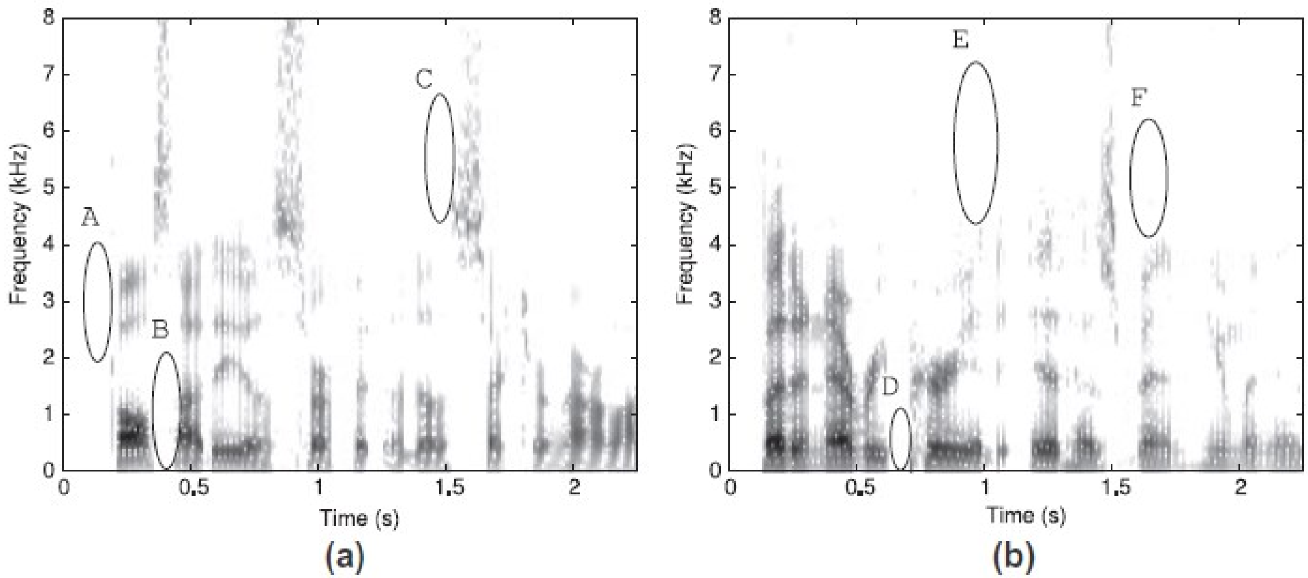
$$\gamma_l = \begin{cases} \gamma_{env} & \text{if } l \in \{0,\ldots,l_{env}\}, \\ \gamma_{pitch} & \text{if } l = l_{pitch}, \\ \gamma_{peak} & \text{if } l \in \{(l_{env}+1),\ldots,K\} \setminus l_{pitch}, \end{cases}$$

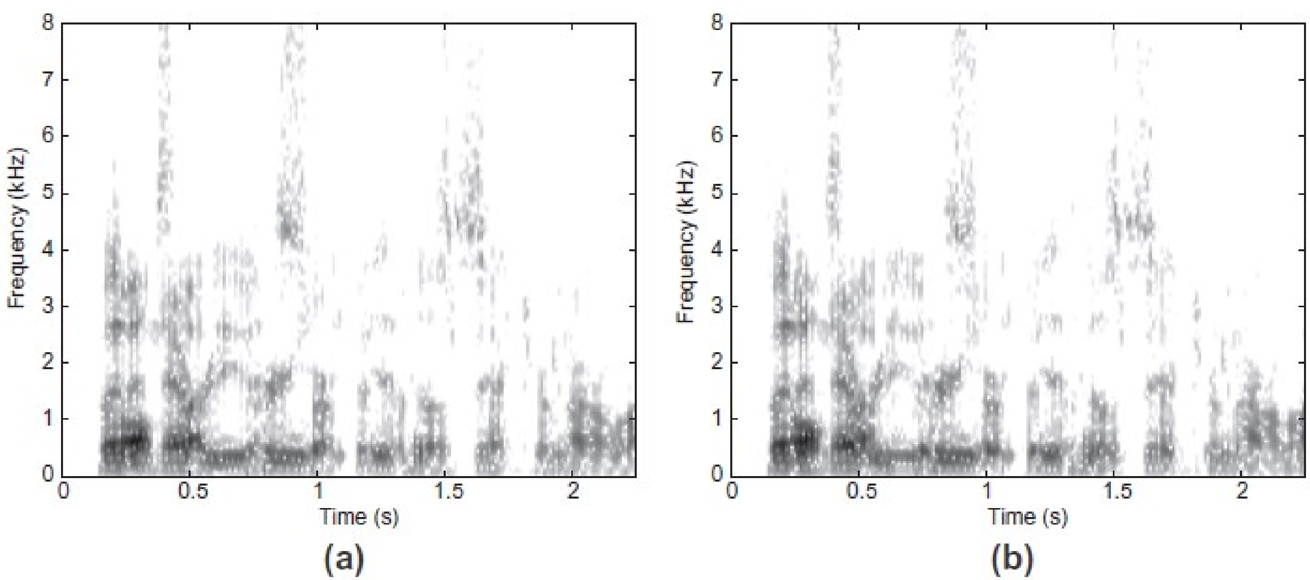
$$l_{pitch} = \mathrm{argmax}_l\{Y^c(l,m) | l_{low} \leqslant l \leqslant l_{high}\},$$

Transform back to the spectral domain:

$$\overline{M}_i^f(k,m) = exp(DFT\{\overline{M}_i^s(l,m)|_{l=0,\ldots,K-1}\}).$$

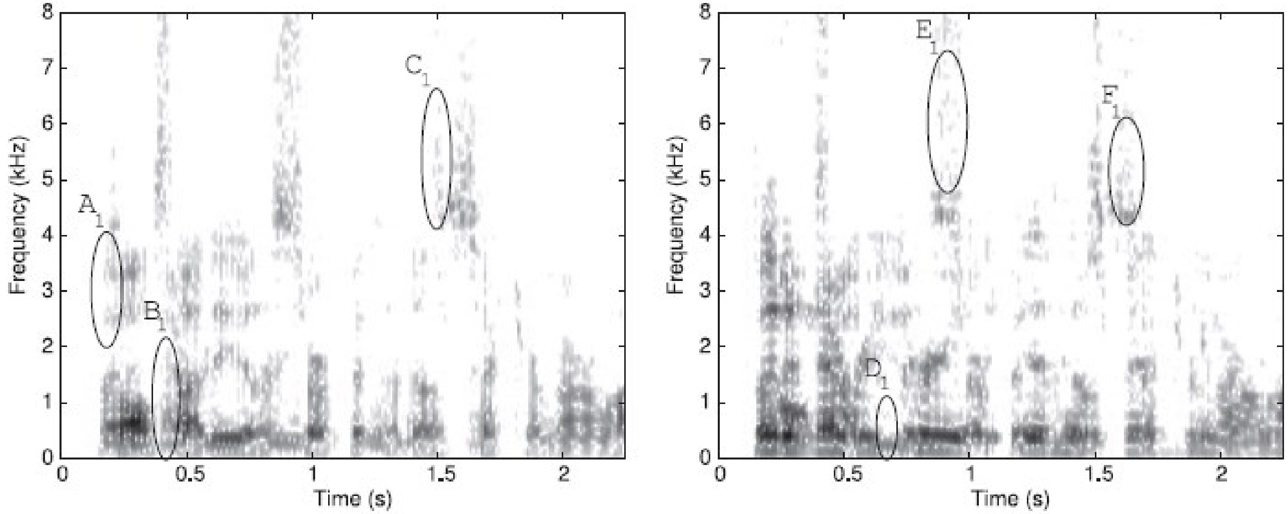
Sources and mixtures



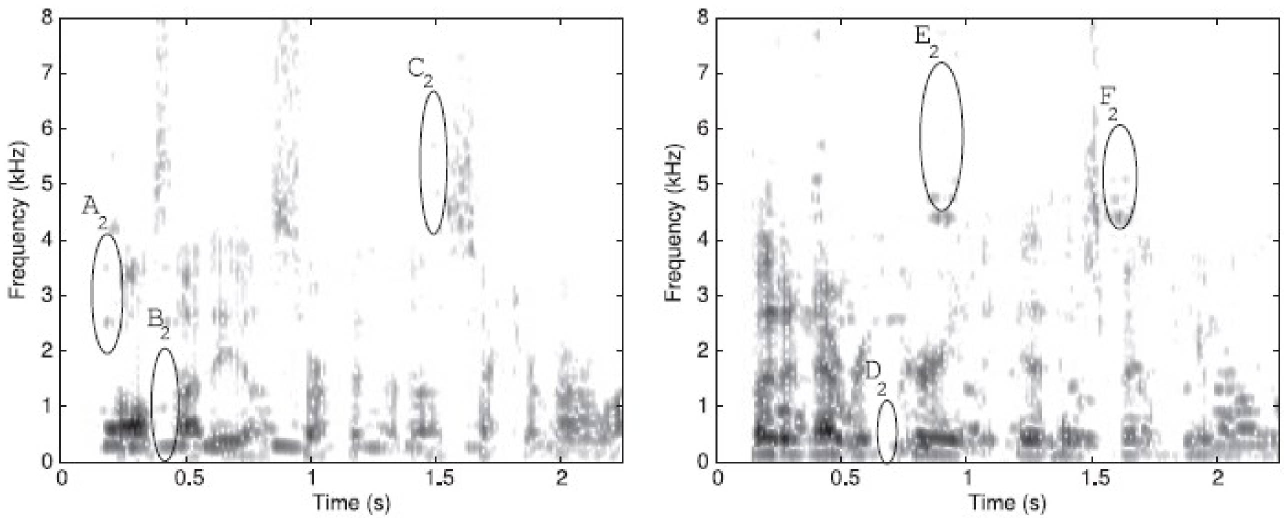


T_{60} = 100ms
Simulated using room image model

Output of convolutive ICA and IBM

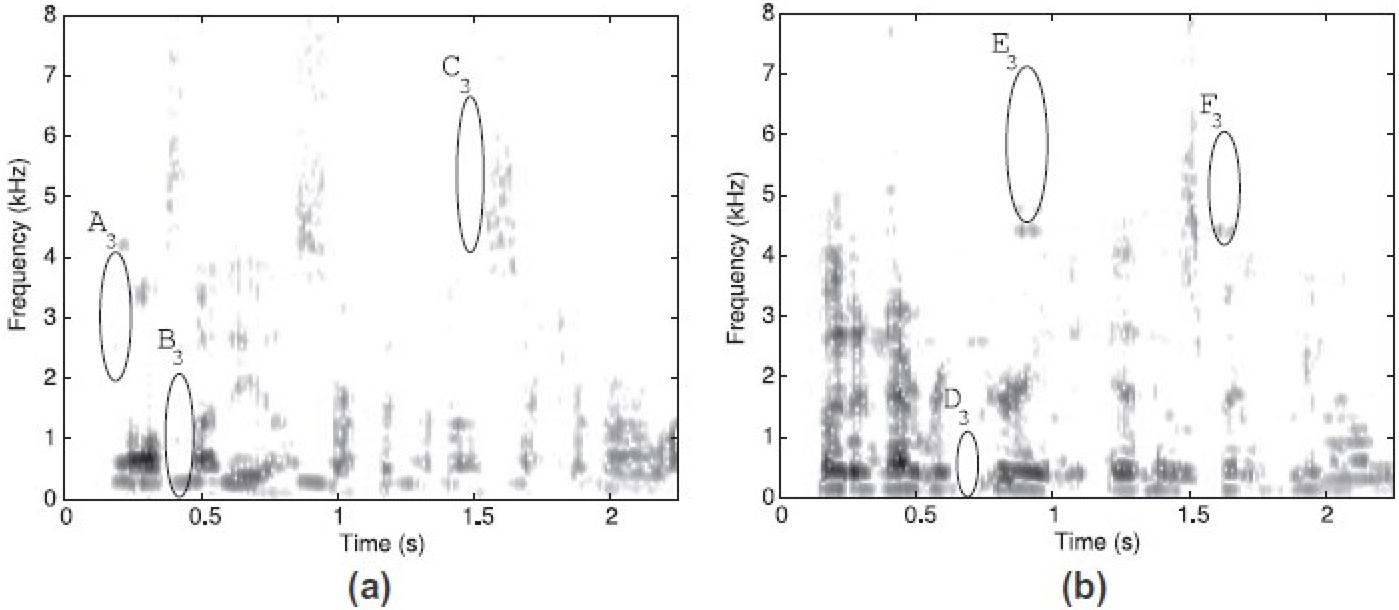


(a) (b)



(a) (b)

Output of cepstral smoothing



Sound demos:
simulated reverberant mixtures

Speakers

Mixtures RT_{60}=30ms	ConvICA	Estimated IBM	Smoothed IBM

Mixtures RT_{60}=150ms	ConvICA	Estimated IBM	Smoothed IBM

Mixtures RT_{60}=400ms	ConvICA	Estimated IBM	Smoothed IBM

Sound demos:
real reverberant mixtures

Separated source signals

Sensor signals

Conv. ICA

Conv. ICA
+IBM

Conv. ICA
+IBM+Cepstral
Smoothing

T. Jan, W. Wang, and D.L. Wang, "A Multistage Approach to Blind Separation of Convolutive Speech Mixtures," *Speech Communication*, vol. 53, pp. 524-539, 2011.

Limitation of the IBM

- Processing artefacts such as musical noise appears to have a deleterious effect on the audio quality of the separated output.
- Not problematic for applications where the output is not auditioned (such as ASR or databasing tasks), but may be problematic for applications (such as speech enhancement or auditory scene reconstruction) where the audio quality is important.
- Recent tests by Hummersone et al. (2014) show that even though the BM gives higher SNR to many other BSS techniques, it gives poorer overall perceptual score (OPS) as compared with these BSS techniques.

C. Hummersone, T. Stokes, and T. Brookes, "On the ideal ratio mask as the goal of computational auditory scene analysis," in *Blind Source Separation: Advances in Theory, Algorithms and Applications*, G. Naik, and W. Wang (ed). , Springer, May, 2014.

Ideal ratio mask (IRM)

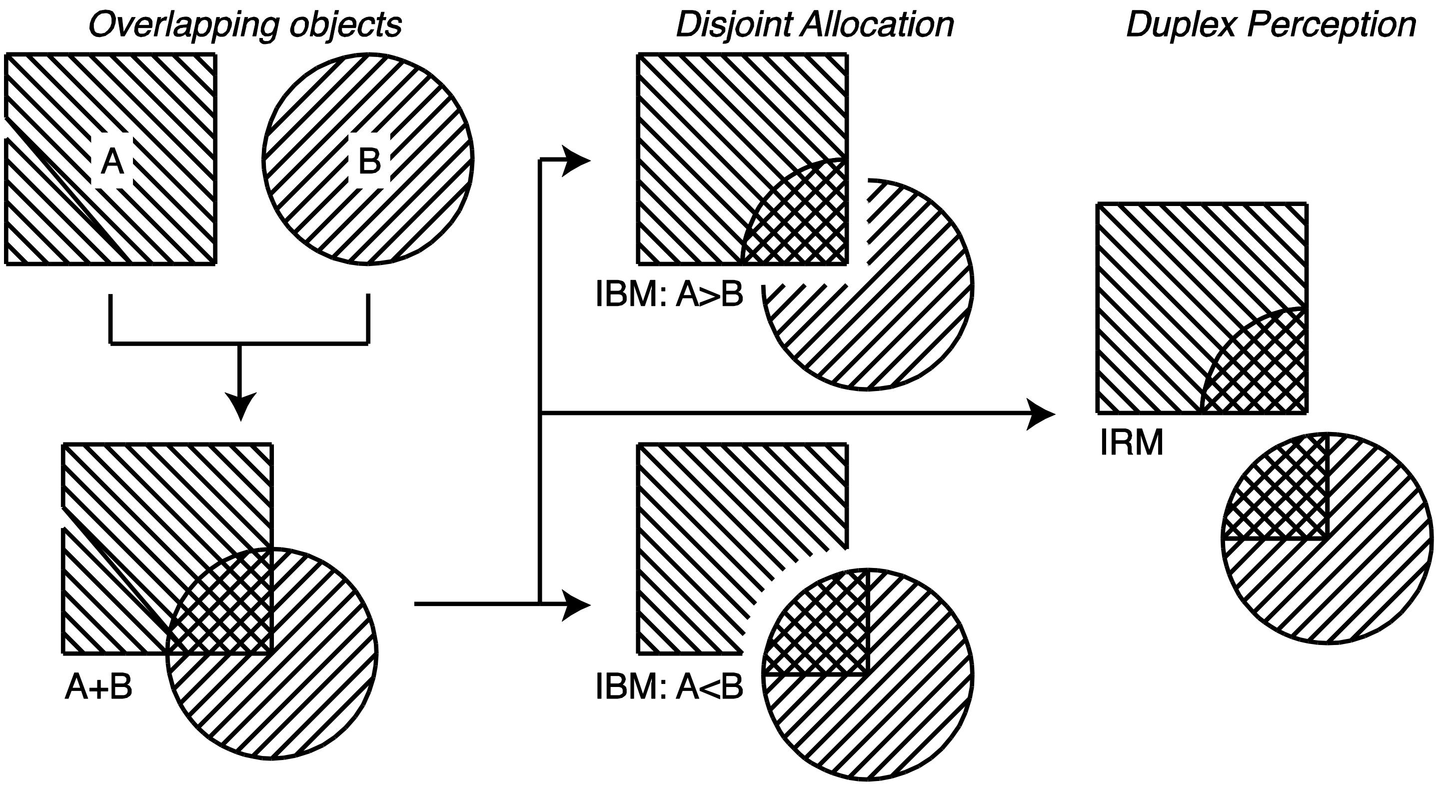
$$m(t,f) = \frac{s(t,f)}{s(t,f) + n(t,f)}$$

S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," Speech Commun., vol. 48, no. 11, pp. 1486-1501, 2006.

In terms of Hummersone et al. (2014), IRM has the following properties:

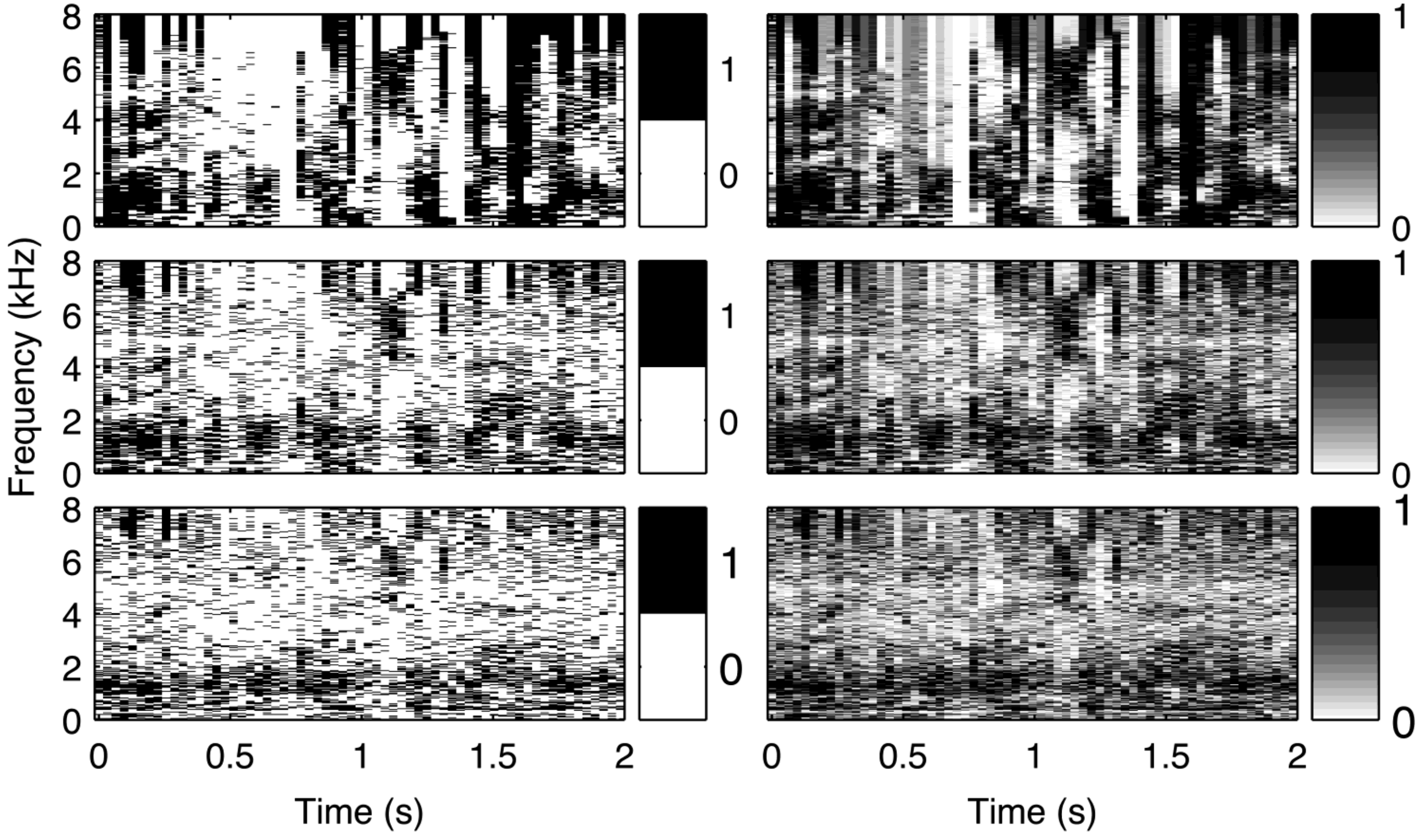
- Flexible: any source can be designated as the target, and the sum of remaining sources is typically designated as the interference.
- Well-defined: the interference component may constitute any number of sources.
- Optimality: closely related to the ideal Wiener filter, which is the optimal linear filter with respect to MMSE.
- Psychoacoustic principles: IRM is perhaps a better approximation of auditory masking and ASA principles than the IBM.

IRM v.s. IBM



Visual analogies of disjoint allocation and duplex perception when objects overlap (left), the disjoint allocation case (middle) is analogous to IBM, while the duplex perception case is analogous to the IRM (right). Plots taken from Hummersone et al. (2014).

IRM v.s. IBM

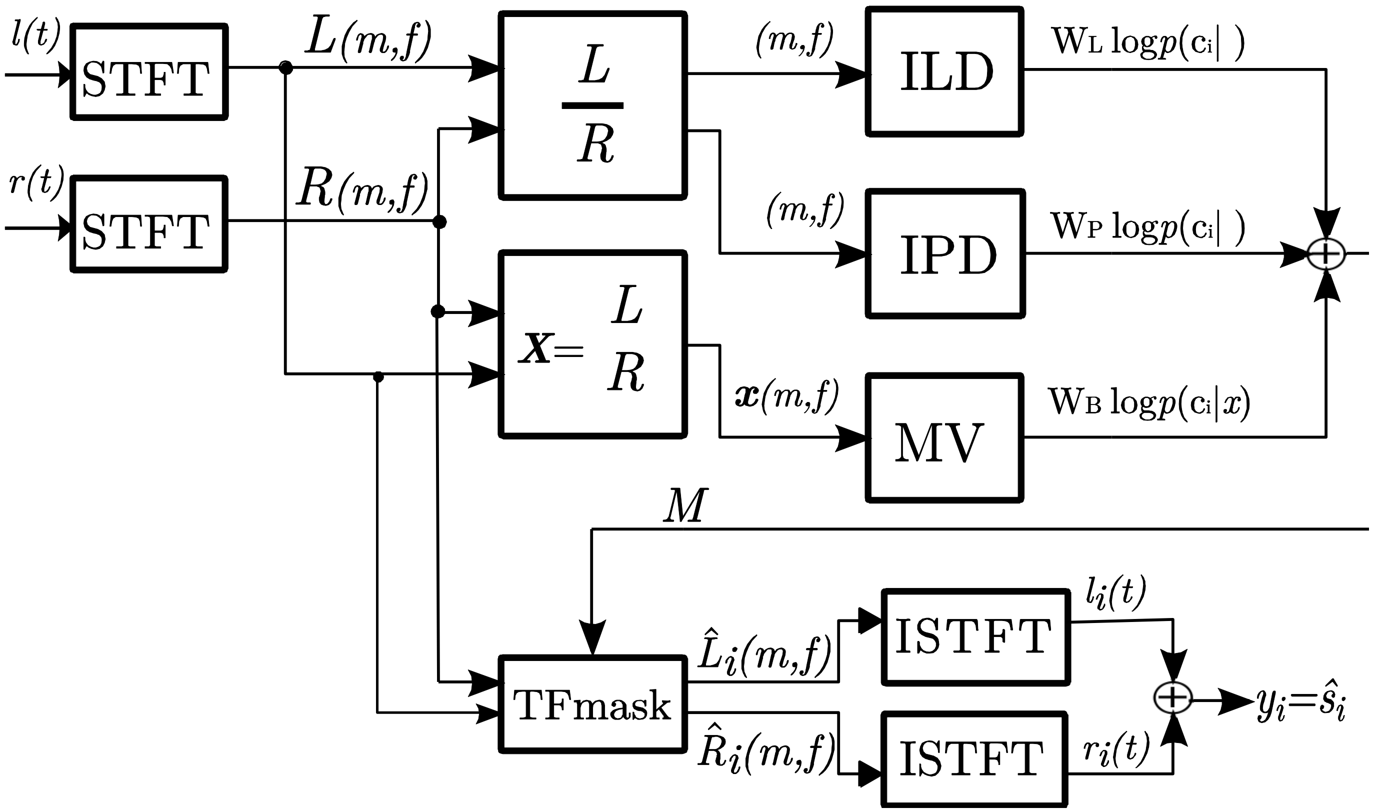


Examples of ideal (top row) and "estimated" masks (middle and bottom rows, with error puturbations). Binary masks (left column) and ratio masks (right column). Plots due to Hummersone et al. (2014).

Soft time-frequency mask:
a model based approach for binaural source separation

- Information considered: mixing vector (MV), binaural cues (interaural level difference (ILD), interaural phase difference (IPD))
- Model and algorithm used:
  - For each time-frequency point, the cues are modelled as Gaussian distributed, and a mixture of Gaussians are therefore used to model the joint distribution of the cues.
  - The model parameters estimated and refined using the expectation maximisation (EM) algorithm
- Soft mask generation: the probability that each source present at each time-frequency point of the mixtures is therefore estimated by the EM which leads to a soft mask that can be used to separate the sources.

Soft mask



A. Alinaghi, P. Jackson, Q. Liu, and W. Wang, "Joint Mixing Vector and Binaural Model Based Stereo Source Separation", *IEEE Transactions on Audio Speech and Language Processing*, 2014. (in press)

Signal model

$$x_k(t) = \sum_{i=1}^{N} s_i(t) * h_{ik}(t) * n_k^c(t) + n_k^a(t),$$

Sparsifying the mixtures with a time-frequency transform, such as STFT

$$X_k(m,f) = \sum_{i=1}^{N} S_i(m,f) \cdot H_{ik}(f) \cdot N_k^c(m,f) + N_k^a(m,f),$$

Assuming the sparsity, each time-frequency point will be dominated by one source

$$X_k(m,f) \approx X_{k|i}(m,f) \cdot N_k^c(m,f) + N_k^a(m,f)$$

where $X_{k|i}(m,f) = S_i(m,f) \cdot H_{ik}(f)$

Estimating cues from mixtures: mixing vector

$$\mathbf{x}(m,f) \approx S_i(m,f)\mathbf{h}_i(f) + \mathbf{n}^a(m,f),$$

$$\tilde{\mathbf{x}}(m,f) = \frac{\mathbf{x}(m,f)}{||\mathbf{x}(m,f)||},$$

$$\approx \tilde{S}_i(m,f)\cdot\tilde{\mathbf{h}}_i(f) + \tilde{\mathbf{n}}^a(m,f).$$

$$\mathbf{z}(m,f) = \frac{\mathbf{W}(f)\tilde{\mathbf{x}}(m,f)}{||\mathbf{W}(f)\tilde{\mathbf{x}}(m,f)||},$$

$$p_B^i(m,f) = \frac{\exp\left(-\frac{||\mathbf{z}-(\mathbf{a}_i^H\mathbf{z}).\mathbf{a}_i||^2}{\gamma_i^2}\right)}{(\pi\gamma_i^2)^{M-1}},$$

$$\mathbf{a}_i(f) \approx \frac{\mathbf{W}(f)\tilde{\mathbf{h}}_i(f)}{||\mathbf{W}(f)\tilde{\mathbf{h}}_i(f)||}.$$

H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Audio, Speech, Lang. Process., vol. 19,* no. 3, pp. 516–527, March 2011.

Estimating cues from mixtures: ILD/IPD cues

$$
\begin{aligned}
\alpha(m,f) &= dB\left(\frac{|X_1(m,f)|}{|X_2(m,f)|}\right) \\
&\approx dB\left(\frac{|H_{i1}(f)|}{|H_{i2}(f)|}\right) + dB\left(\frac{|N_1^c(m,f)|}{|N_2^c(m,f)|}\right), \\
\phi(m,f) &= \angle\left(\frac{X_1(m,f)}{X_2(m,f)}\right) \\
&\approx \angle\left(\frac{H_{i1}(f)}{H_{i2}(f)}\right) + \angle\left(\frac{N_1^c(m,f)}{N_2^c(m,f)}\right)
\end{aligned}
$$

$$
\begin{aligned}
p_L^i(m,f) &= \mathcal{N}(\alpha(m,f)|\mu_i(f), \eta_i^2(f)), \\
p_P^{i,\tau}(m,f) &= \mathcal{N}(\hat{\phi}(m,f;\tau(f))|\xi_{i\tau}(f), \sigma_{i\tau}^2(f))
\end{aligned}
$$

M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, “Model-based expectationmaximization source separation and localization,” *IEEE Trans. Audio, Speech, Lang. Process., vol. 18, no. 2, pp. 382–394, February 2010.*

GMM model

Log-likelihood of the observations:

$$\mathcal{L}(\Theta) = \sum_{m,f} \log p(\phi(m,f), \alpha(m,f), \mathbf{z}(m,f)|\Theta)$$

$$= \sum_{m,f} \log \sum_{i,\tau} \Big\{ \psi_{i\tau} \cdot p_P^{i,\tau}(m,f) \cdot$$

$$p_L^i(m,f) \cdot p_B^i(m,f) \Big\},$$

Model parameters:

$$\Theta = \{\xi_{i\tau}, \sigma_{i\tau}, \mu_i, \eta_i, \mathbf{a}_i, \gamma_i \, \psi_{i\tau}\}$$

Parameter estimation via expectation maximization

UNIVERSITY OF SURREY

- The E-step calculates the expected value of the log-likelihood function with respect to the observations of the IPD, ILD, and MV, under the current estimates of the parameters.

  - In other words, given the estimated parameters Θ and the observations, and assuming the statistical independence of the cues, the probability of each source occupying at each time-frequency point of the mixture is calculated:

$$\nu_{i\tau}(m,f) = K \cdot \psi_{i\tau} \cdot p_P^{i,\tau}(m,f) \cdot p_L^{i}(m,f) \cdot p_B^{i}(m,f)$$

Parameter estimation via expectation maximization

- The M-step calculates the model parameters (mean and variance):

ILD:

$$\mu_i(f) = \frac{\sum_{m,\tau} \alpha(m,f)\nu_{i\tau}(m,f)}{\sum_{m,\tau} \nu_{i\tau}(m,f)},$$

$$\eta_i^2(f) = \frac{\sum_m (\alpha(m,f) - \mu_i(f))^2 \sum_\tau \nu_{i\tau}(m,f)}{\sum_{m,\tau} \nu_{i\tau}(m,f)}.$$

IPD:

$$\xi_{i\tau}(f) = \frac{\sum_m \hat{\phi}(m,f;\tau)\nu_{i\tau}(m,f)}{\sum_m \nu_{i\tau}(m,f)},$$

$$\sigma_{i\tau}^2(f) = \frac{\sum_m (\hat{\phi}(m,f;\tau) - \xi_{i\tau}(f))^2 \nu_{i\tau}(m,f)}{\sum_m \nu_{i\tau}(m,f)}.$$

MV:

$$\mathbf{R}_i(f) = \sum_{m,\tau} \nu_{i\tau}(m,f)\mathbf{z}(m,f)\mathbf{z}^H(m,f),$$

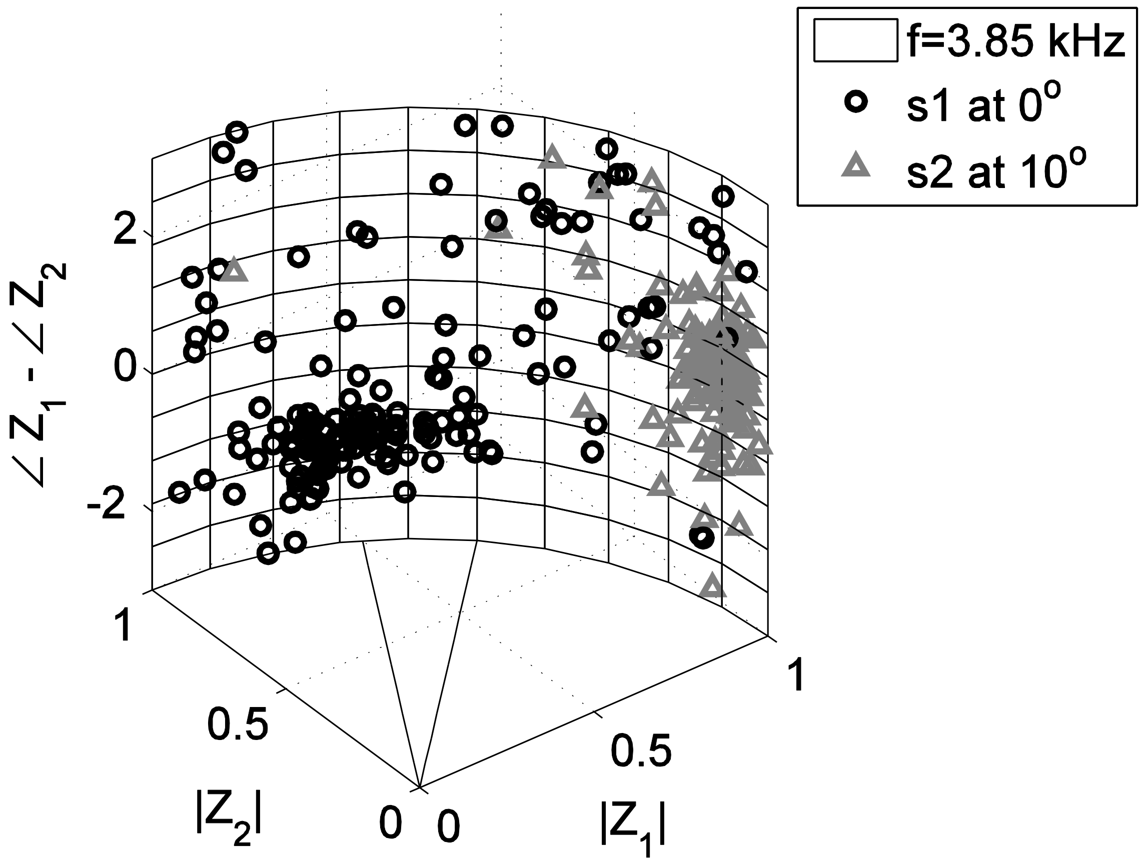
$$\mathbf{a}_i(f) = \text{eigenvector}(\mathbf{R}_i(f))_{\max(\lambda)},$$

$$\gamma_i^2(f) = \frac{\sum_{m,\tau} \nu_{i\tau}(m,f)||\mathbf{z} - (\mathbf{a}_i^H\mathbf{z}).\mathbf{a}_i||^2}{(M-1)\sum_{m,\tau} \nu_{i\tau}(m,f)},$$

Weights:

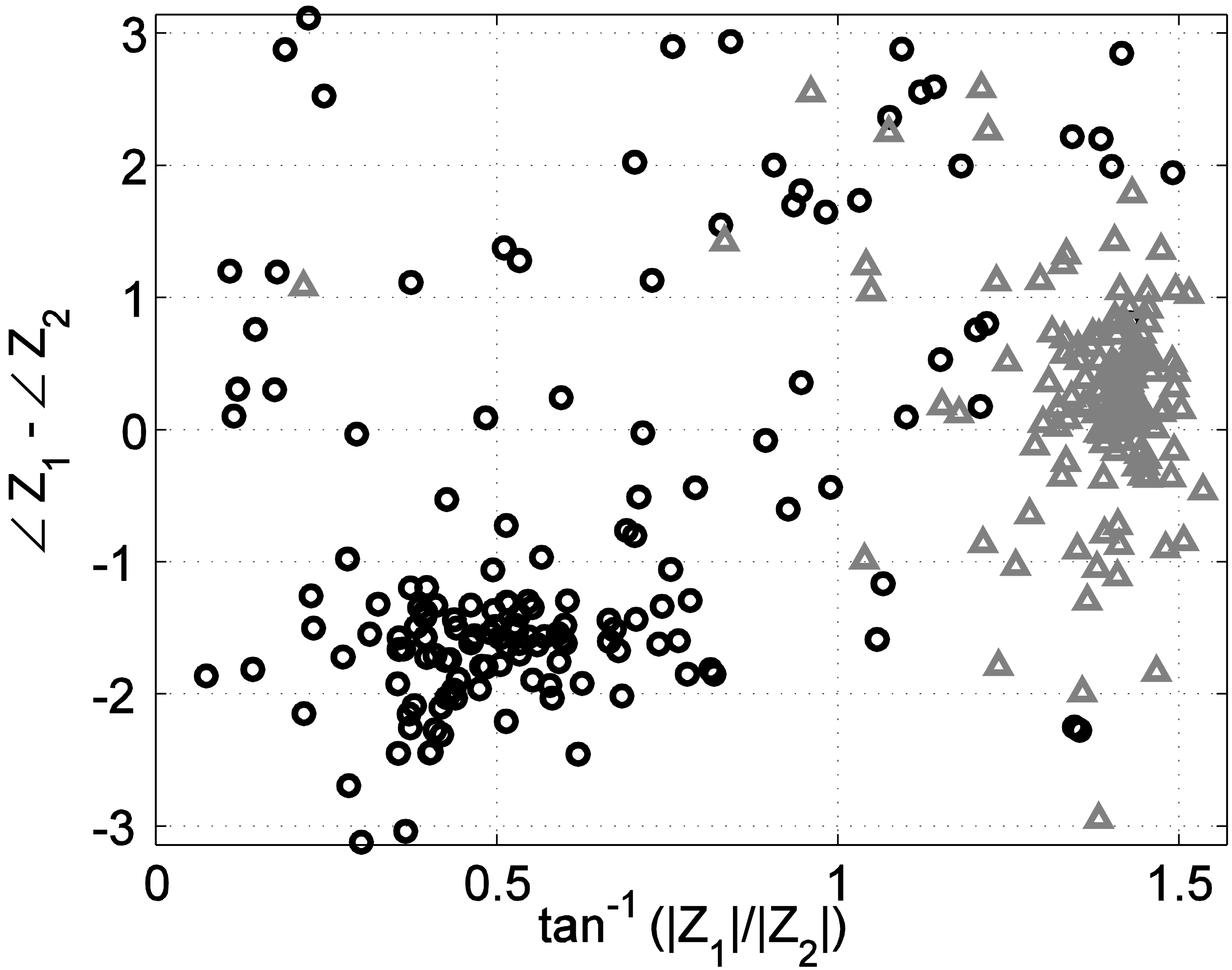
$$\psi_{i\tau} = \frac{1}{TF}\sum_{m,f} \nu_{i\tau}(m,f)$$

2D representation of the observation vectors

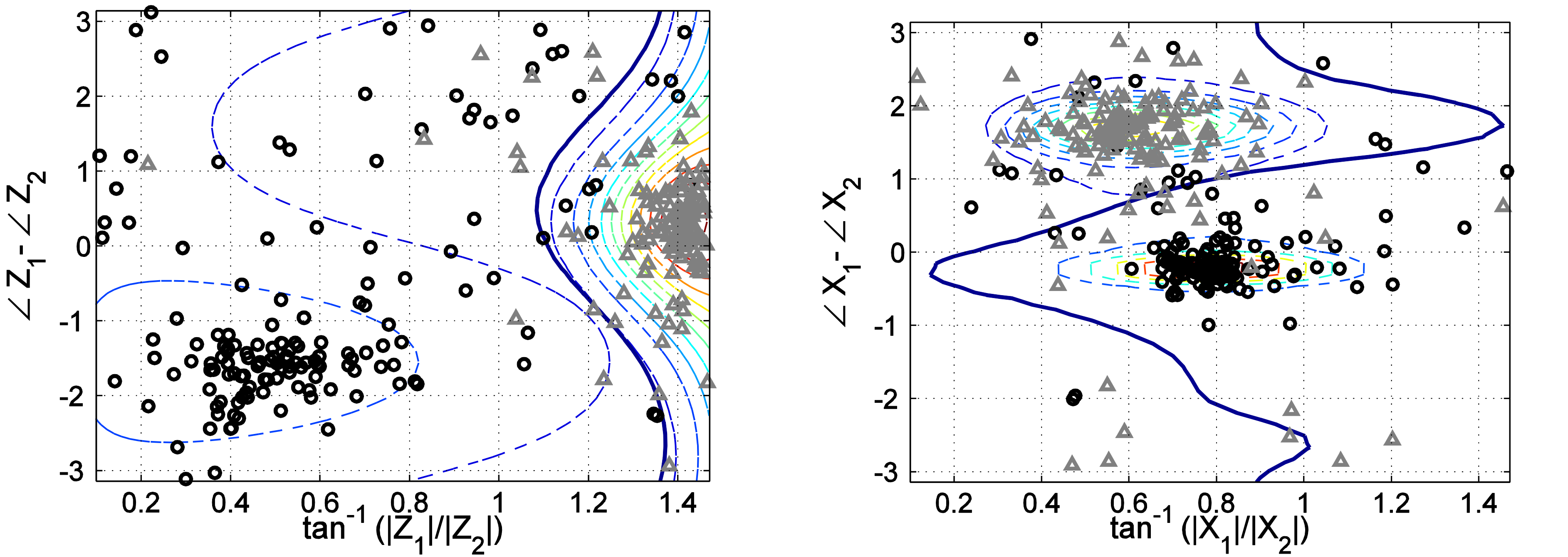


A unit cylinder wall is used to visualise the observation vectors after normalisation and whitening, in frequency channel 3.85 kHz, for two different sources that are close to each other.

Unwrapped 2D plane

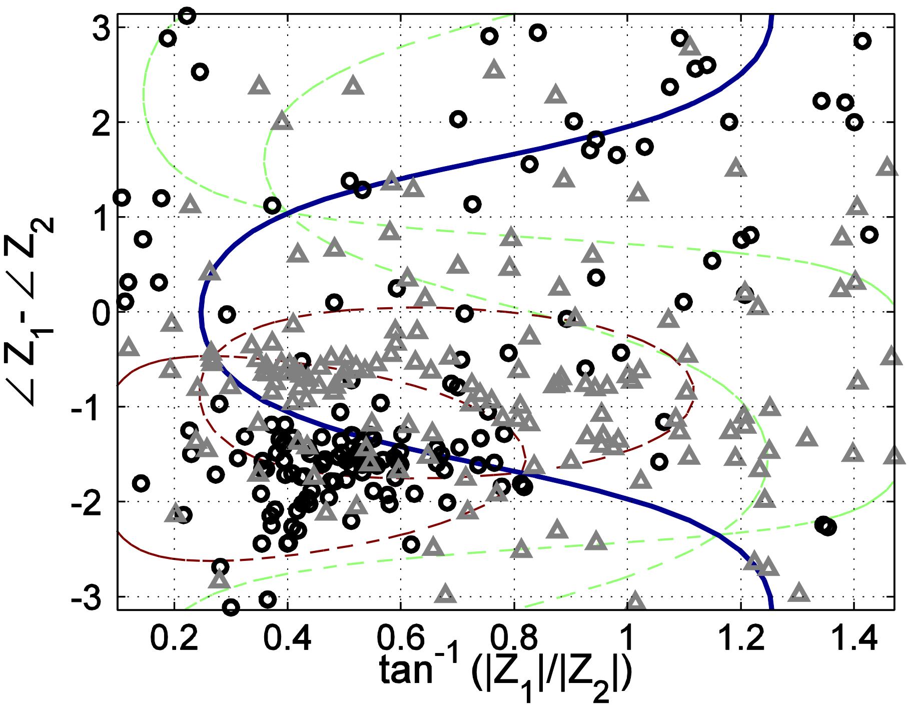


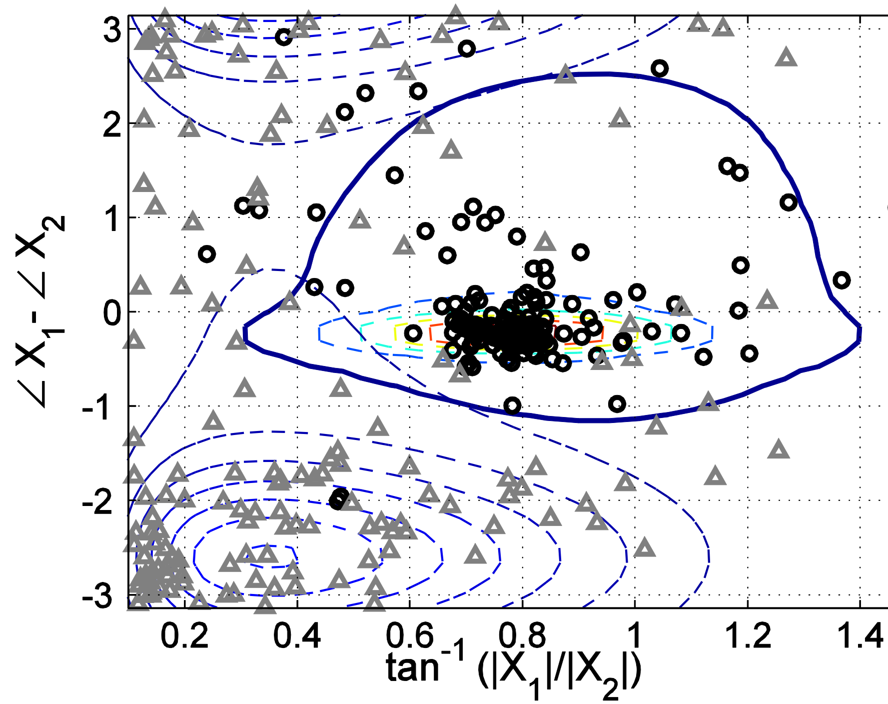
MV v.s. binaural cues: closely spaced sources



Scatter plot and probability contours (dashed lines) for sources in room A at 0º in circles, and 10º in triangles with decision boundaries by solid lines based on mixing vectors and binaural cues in the frequency band of 3.85 kHz.

MV v.s. binaural cues: sources placed far from each other





Scatter plot and probability contours (dashed lines) for sources in room A at 0° in circles, and 80° in triangles with decision boundaries by solid lines based on mixing vectors and binaural cues in the frequency band of 3.85 kHz.

MV v.s. binaural cues: KL divergence measure

$$\kappa^{\mathrm{MV}}(f) = \sum_m \Big\{ p(\mathbf{z}(m,f)|\mathbf{a}_1(f),\gamma_1(f)) \cdot \log \frac{p(\mathbf{z}(m,f)|\mathbf{a}_1(f),\gamma_1(f))}{p(\mathbf{z}(m,f)|\mathbf{a}_2(f),\gamma_2(f))} \Big\},$$

$$p(\mathbf{z}(m,f)|\mathbf{a}_i(f),\gamma_i(f)) = \frac{\exp\left(-\frac{||\mathbf{z}-(\mathbf{a}_i^H\mathbf{z}).\mathbf{a}_i||^2}{\gamma_i^2}\right)}{(\pi\gamma_i^2)^{M-1}},$$

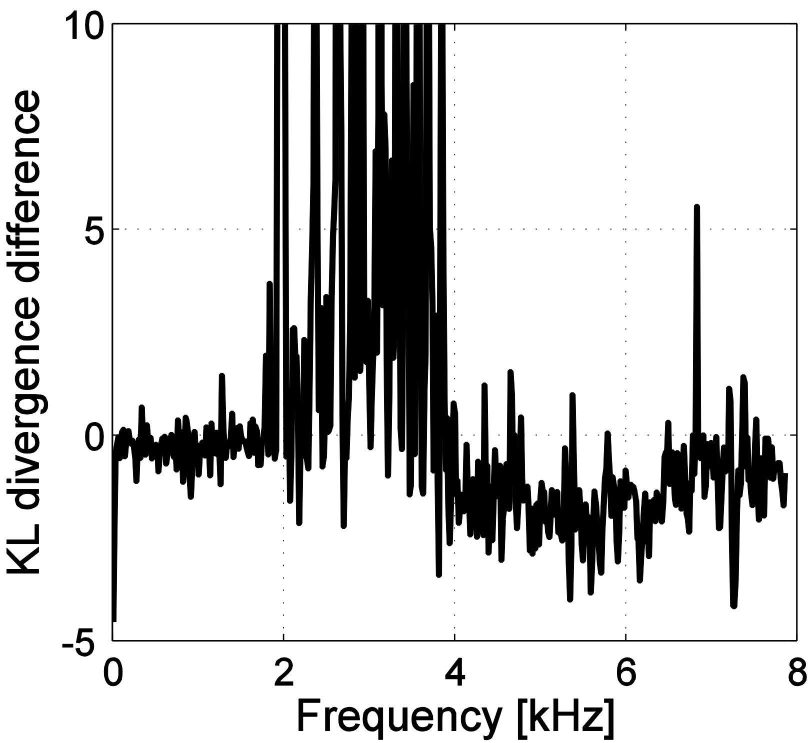
$$\kappa^{\mathrm{Binaural}}(f) = \sum_m p\left(\mathbf{x}(m,f)|1\right) \log \frac{p\left(\mathbf{x}(m,f)|1\right)}{p\left(\mathbf{x}(m,f)|2\right)},$$

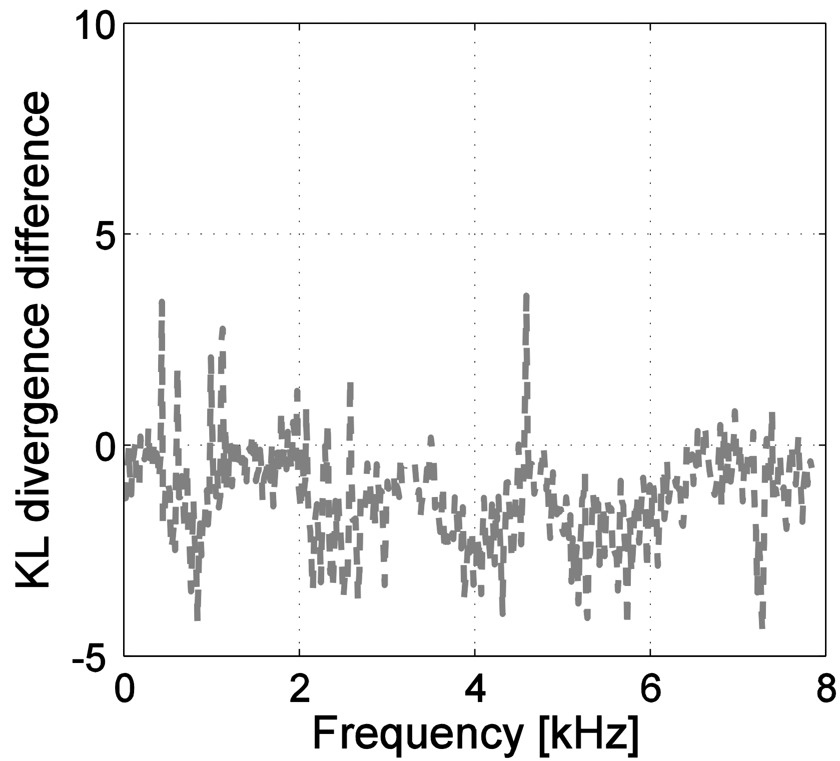
where $p\left(\mathbf{x}(m,f)|i\right) = p_P^i(m,f) \cdot p_L^i(m,f)$

$$p(\alpha(m,f)|i) = \mathcal{N}(\alpha(m,f)|\mu_i(f),\eta_i^2(f))$$

$$p(\hat{\phi}(m,f)|i,\tau) = \mathcal{N}(\hat{\phi}(m,f;\tau(f))|\xi_{i\tau}(f),\sigma_{i\tau}^2(f))$$

MV v.s. binaural cues:
KL divergence difference ($KL^{MV} – KL^{Binaural}$)





The difference between the KL divergences obtained respectively from the MV and the binaural models. The KL divergence between the two source models is calculated based on binaural cues and MV cues in room A (RT=0.32s), where one source is placed at 0º and the other at 10º (left plot), and 80º (right plot) respectively.

MV v.s. binaural cues:
High reverberations

UNIVERSITY OF SURREY

$$\kappa_B^{n}(f) = \sum_m \Big\{ p(\mathbf{z}(m,f)|\mathbf{a}(f),\gamma(f)) \cdot \log \frac{p(\mathbf{z}(m,f)|\mathbf{a}(f),\gamma(f))}{p(\mathbf{z}(m,f)|\mathbf{a}^{n}(f),\gamma^{n}(f))} \Big\}.$$

-	additive noise	convolutive noise
mixing vector (MV)	**2.10**	2.31
IPD	2.70	**2.01**
ILD	3.39	**3.29**

KL divergence between the clean and noisy signal models for three different cues and two types of noise averaged over all frequencies.

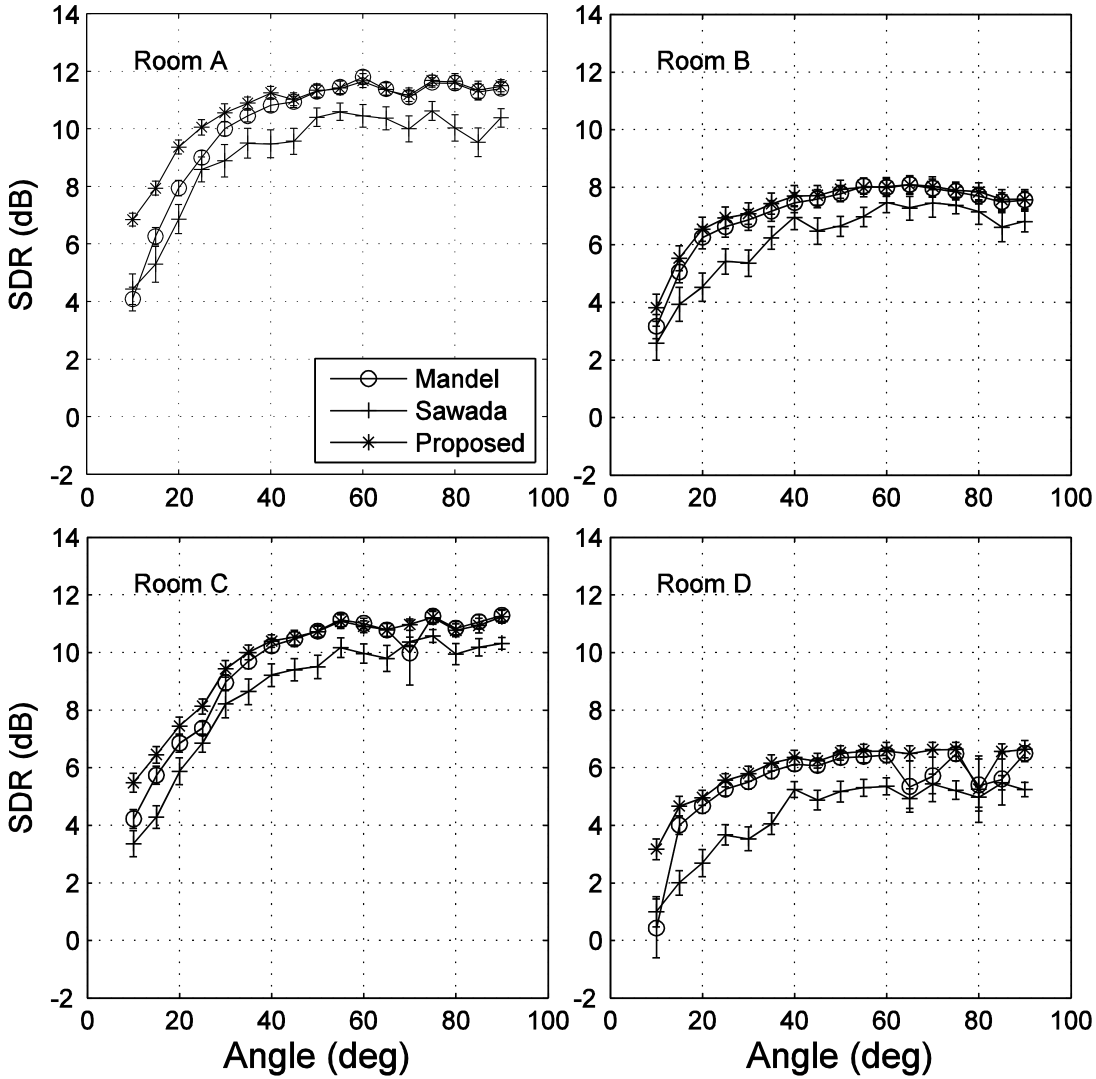
Experimental set-up

- 15 utterances, sampled at 16 kHz, spoken by both male and female are selected from TIMIT database.
- BRIRs measured by Hummersone at University of Surrey was used.
- For each T60, and angular configuration, 15 pairs from those 15 selected utterances were chosen, and then convolved with the BRIRs to generate the binaural mixtures.

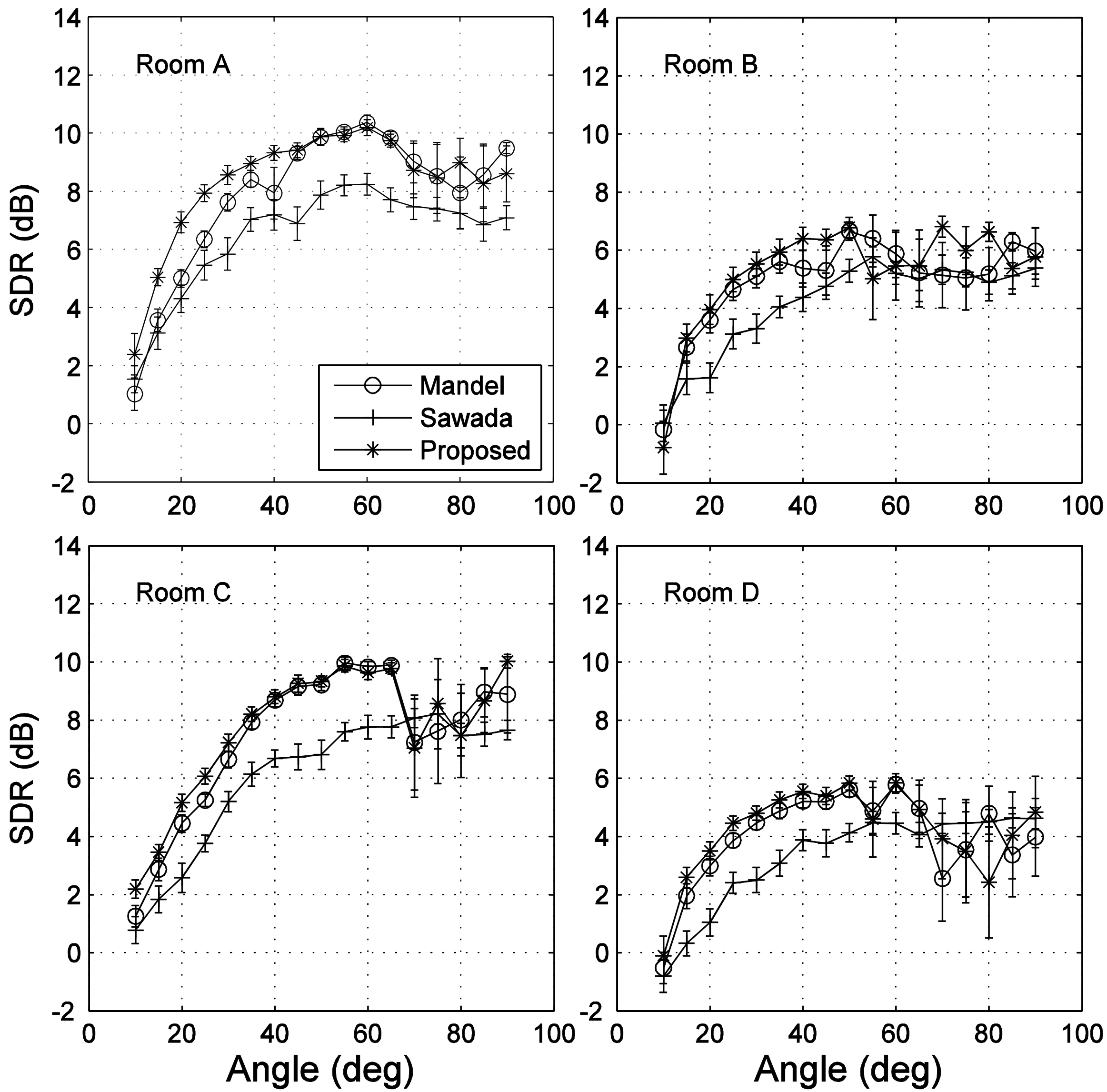
ROOM ACOUSTICAL PROPERTIES IN INITIAL TIME DELAY GAP (ITDG), DIRECT-TO-REVERBERANT RATIO IN TERMS OF (DRR) AND REVERBERATION TIME T_{60} [32].

Room	Type	ITDG [ms]	DRR [dB]	T_{60} [s]
A	a medium office	8.72	6.09	0.32
B	a small class room	9.66	5.31	0.47
C	a large lecture theatre	11.9	8.82	0.68
D	a large seminar room	21.6	6.12	0.89

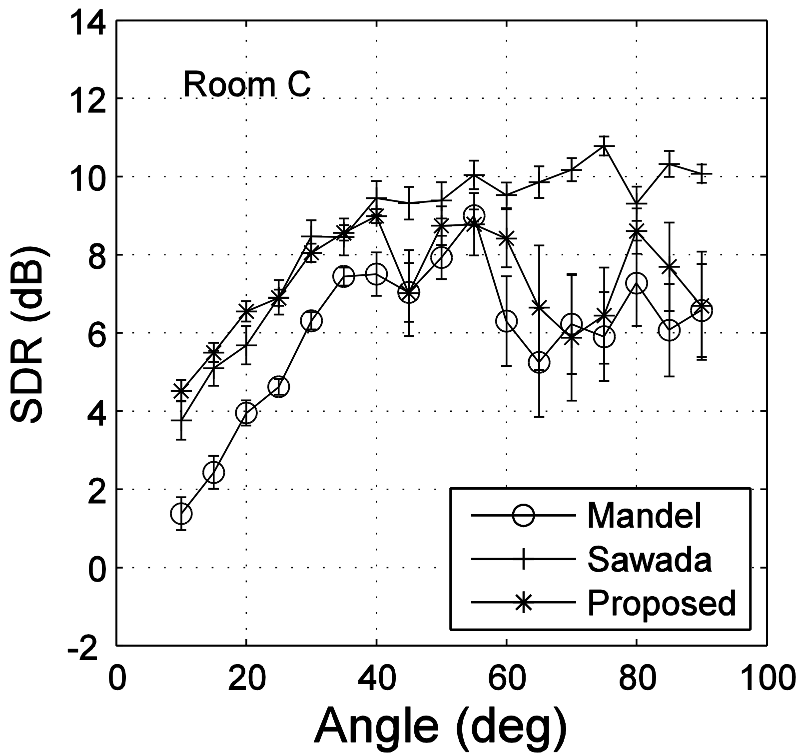
Results on real room mixtures: two sources case

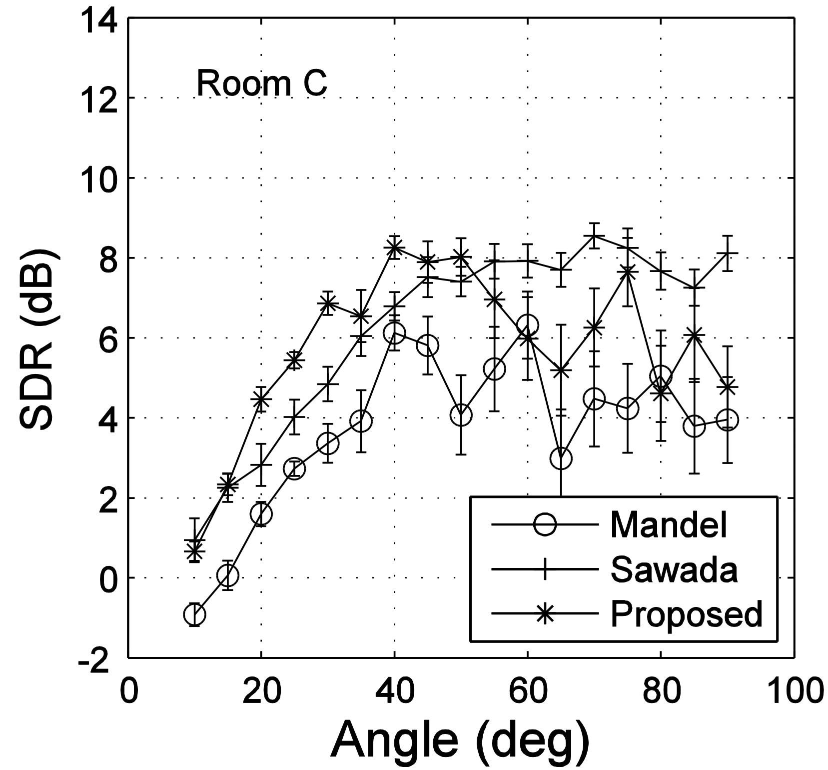


Results on three sources case



Diffuse noise





2-source case

3-source case

Sound demos

UNIVERSITY OF SURREY

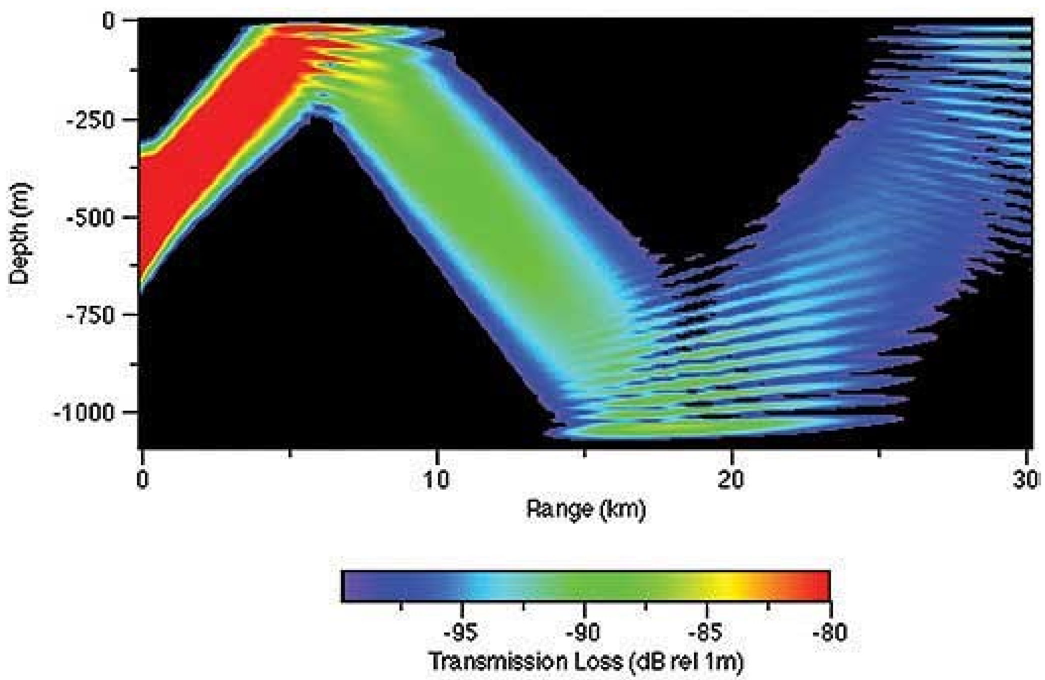
2-source case:

			Mandel et al.	Sawada et al.	Alinaghi et al.
left		es1			
right		es2			

3-source case:

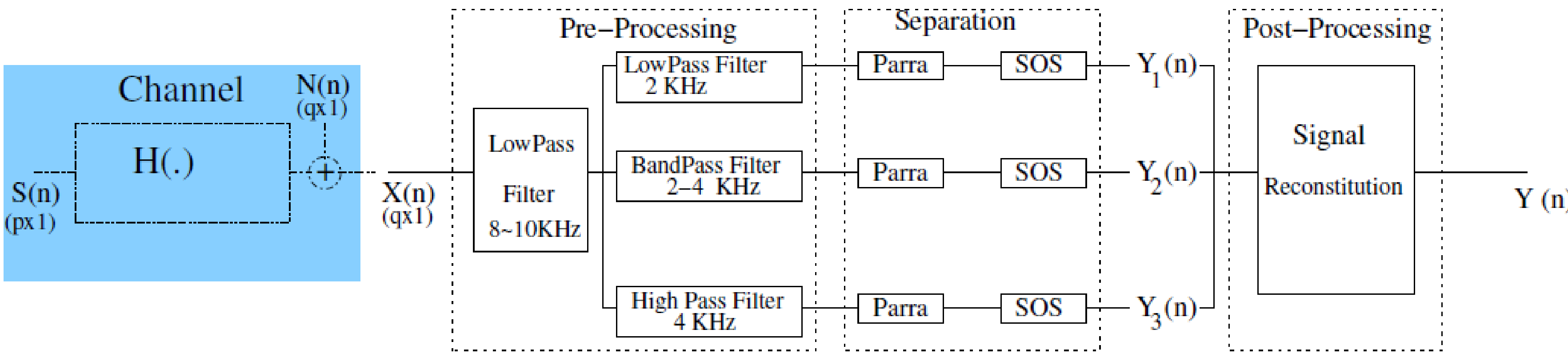
			Mandel et al.	Sawada et al.	Alinaghi et al.
left		es1			
right		es2			
		es3			

Convolutive source separation: from room acoustics to underwater acoustics?



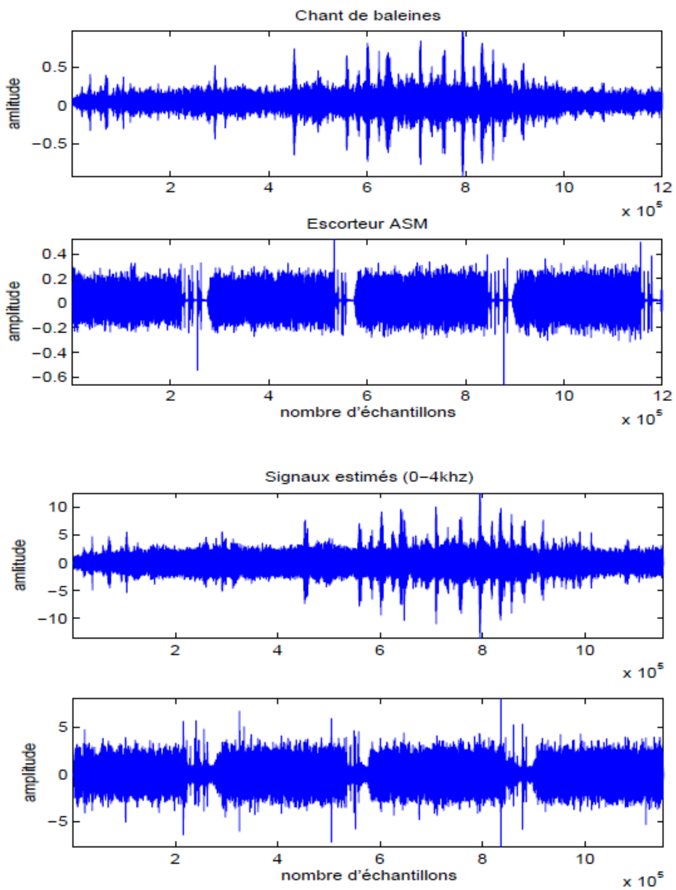
Computer simulated underwater acoustic propagation model in a simplied ocean environment.

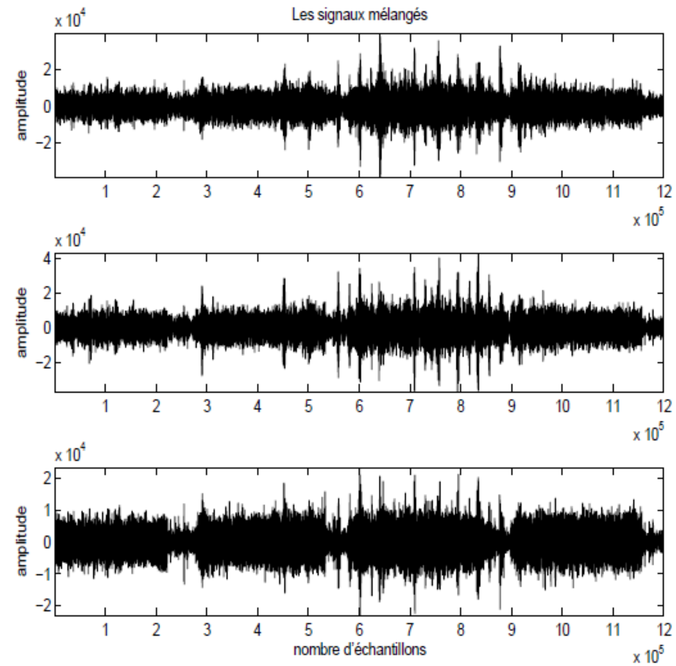
Blind separation of underwater acoustic signals



A. Mansour, N. Benchekroun, and C. Gervaise, "Blind separation of underwater acoustic signals," Proc. ICA, Lecture Notes in Computer Science, Volume 3889, pp 181-188, 2006.

Blind separation of underwater acoustic signals





A. Mansour, N. Benchekroun, and C. Gervaise, "Blind separation of underwater acoustic signals," Proc. ICA, Lecture Notes in Computer Science, Volume 3889, pp 181-188, 2006.

Summary & future work

We have covered the following:

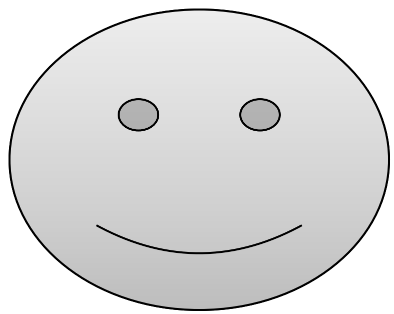
- ❑Concept of convolutive source separation
- ❑Methods for performing convolutive source separation, such as
    - Convolutive ICA and frequency domain ICA (permutation/scaling ambiguities)
    - Time-frequency masking (CASA, IBM, IRM, etc)
    - Integrating ICA/IBM
    - Music noise problem & mitigation
    - Model-based convolutive stereo source separation (ILD/IPD, MV, etc.)
- ❑Applications from room acoustics to underwater acoustics
- ❑Future work include improving source separation performance in highly noisy environment, and/or missing data scenarios.

Acknowledgement

- Collaborators: Mrs Atiyeh Alinaghi, Dr Qingju Liu, and Dr Tariq Jan, Dr Philip Jackson, Prof Josef Kittler, Prof Jonathon Chambers (Loughborough University), Dr Saeid Sanei, Prof DeLiang Wang (Ohio State University).
- Financial support: EPSRC & DSTL, UDRC in Signal Processing

… and finally

Thank you for your attention!



Q & A

w.wang@surrey.ac.uk
http://personal.ee.surrey.ac.uk/Personal/W.Wang/