Audio source separation

Roland Badeau,
roland.badeau@telecom-paris.fr

TSIA 206 - Speech and audio processing

Part I

Introduction

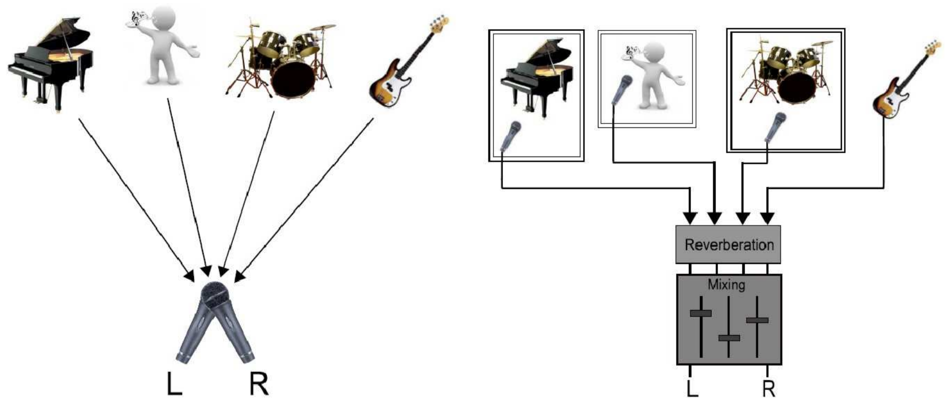
Introduction

- Source separation
  - Art of estimating "source" signals, assumed independent, from the observation of one or several "mixtures" of these sources
- Application examples:
  - Denoising (cocktail party, suppression of vuvuzela, karaoke)
  - Separation of the instruments in polyphonic music
  - Remix, transformations, re-spatialization

Typology of the mixture models

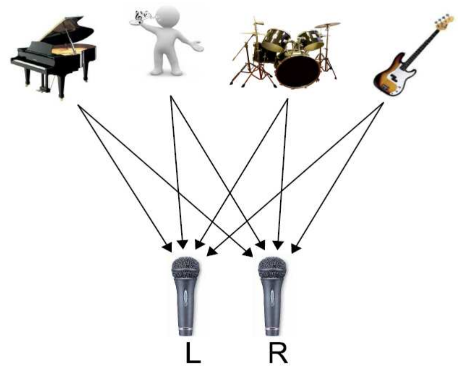
- Definition of the problem
  - Observations: M mixtures $x_m(t)$, concatenated in a vector $\mathbf{x}(t)$
  - Unknowns: K sources $s_k(t)$, concatenated in a vector $\mathbf{s}(t)$
  - General mixture model: function $\mathscr{A}$ which transforms $\mathbf{s}(t)$ into $\mathbf{x}(t)$
- Stationarity: $\mathscr{A}$ is translation invariant
- Linearity: $\mathscr{A}$ is a linear map
- Memory:
  - Convolutive mixtures
  - Instantaneous mixtures: $\mathbf{x}(t) = \mathbf{A}\,\mathbf{s}(t)$
    - $\mathscr{A}$ is defined by the "mixture matrix" $\mathbf{A}$ (of dimension $M \times K$)
- Inversibility:
  - Determined mixtures: $M = K$
  - Over-determined mixtures: $M > K$
  - Under-determined mixtures: $M < K$

Instantaneous linear mixtures

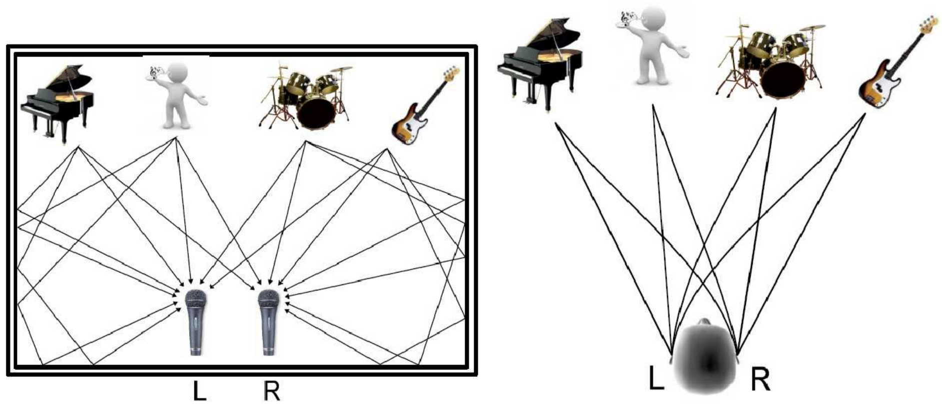


(a) XY Stereo configuration (b) Direct injection to the mixer

Anechoic linear mixtures



Convolutive linear mixtures



(a) Convolutive mixture (b) Binaural mixture

Part II

Mathematical reminders

Real random vectors

- Notation: $\phi[\mathbf{x}]$ denotes a function of $p(\mathbf{x})$
- Mean vector: $\boldsymbol{\mu}_x = \mathbb{E}[\mathbf{x}]$
- Covariance matrix: $\boldsymbol{\Sigma}_{xx} = \mathbb{E}[(\mathbf{x}-\boldsymbol{\mu}_x)(\mathbf{x}-\boldsymbol{\mu}_x)^T]$
- Characteristic function:
  $\phi_x(\mathbf{f}) = \mathbb{E}[e^{-2i\pi\mathbf{f}^T\mathbf{x}}] = \int_{\mathbb{R}} p(\mathbf{x})e^{-2i\pi\mathbf{f}^T\mathbf{x}}d\mathbf{x}$
- Probability distribution: $p(\mathbf{x}) = \int_{\mathbb{R}} \phi_x(\mathbf{f})e^{+2i\pi\mathbf{f}^T\mathbf{x}}d\mathbf{f}$
- Cumulants:
  - Definition: $\ln(\phi_x(\mathbf{f})) = \sum_{n=1}^{+\infty}\frac{(-2i\pi)^n}{n!}\sum_{k_1=1}^{K}\sum_{k_n=1}^{K}\kappa^n_{k_1\dots k_n}[\mathbf{x}]f_{k_1}\dots f_{k_n}$
  - $\kappa^n[\mathbf{x}]$ is an n-th order tensor
  - $\kappa^1[\mathbf{x}]$ is the mean vector, $\kappa^2[\mathbf{x}]$ is the covariance matrix
  - If $p(\mathbf{x})$ is symmetric ($p(-\mathbf{x}) = p(\mathbf{x})$), $\kappa^n[\mathbf{x}] = 0$ for any odd value n
  - the ratio $\kappa^4_{k,k,k,k}[\mathbf{x}]/(\kappa^2_{k,k}[\mathbf{x}])^2$ is called "kurtosis"

Real Gaussian random vectors

- The Gaussian distribution is the one such that all cumulants of order $n > 2$ are zero
- Characteristic function

$$\phi_x(\mathbf{f}) = \exp(-2i\pi\mathbf{f}^T\boldsymbol{\mu}_x - 2\pi^2\mathbf{f}^T\boldsymbol{\Sigma}_{xx}\mathbf{f})$$

- Probability density function (defined if $\boldsymbol{\Sigma}_{xx}$ is invertible)

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{K}{2}}\det(\boldsymbol{\Sigma}_{xx})^{\frac{1}{2}}}\exp\left(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_x)^T\boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{x}-\boldsymbol{\mu}_x)\right)$$

WSS vector processes

- Definition: the cumulants of orders 1 et 2 are translation-invariant
- Covariance matrices of 2 centered WSS processes $\mathbf{x}(t)$ and $\mathbf{y}(t)$:
  - Definition: $\mathbf{R}_{xy}(\tau) = \mathbb{E}\left[\mathbf{x}(t+\tau)\mathbf{y}(t)^T\right]$
  - Property: $\mathbf{R}_{xx}(0) = \boldsymbol{\Sigma}_{xx}$ is Hermitian and positive semi-definite.
- PSD matrices of a WSS process $\mathbf{x}(t)$:
  - Definition: $\mathbf{S}_{xx}(\nu) = \sum_{\tau\in\mathbb{Z}}\mathbf{R}_{xx}(\tau)e^{-2i\pi\nu\tau}$
  - Property: $\forall\nu$, $\mathbf{S}_{xx}(\nu)$ is Hermitian and positive semi-definite

Information theory

- Shannon entropy
  - Definition: $\mathbb{H}[\mathbf{x}] = -\mathbb{E}[\ln(p(\mathbf{x}))]$
  - $\mathbb{H}[\mathbf{x}]$ is not necessarily non-negative for a continuous r.v.
- Kullback-Leibler divergence
  - $D_{KL}(p||q) = \int p(\mathbf{x})\ln\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right)d\mathbf{x}$
  - Property: $D_{KL}(p||q) \geq 0$, $D_{KL}(p||q) = 0$ if and only if $p = q$
- Mutual information
  - Definition:
    $\mathbb{I}[\mathbf{x}] = \mathbb{E}\left[\ln\left(\frac{p(\mathbf{x})}{p(x_1)\dots p(x_K)}\right)\right] = D_{KL}(p(\mathbf{x})||p(x_1)\dots p(x_K))$
  - Property: $\mathbb{I}[\mathbf{x}] = 0$ if and only if $x_1\dots x_K$ are mutually independent
  - Relationship with entropy: $\mathbb{I}[\mathbf{x}] = \sum_{k=1}^{K}\mathbb{H}[x_k] - \mathbb{H}[\mathbf{x}]$

Part III

Linear instantaneous mixtures

Blind source separation (BSS)

- Observation model:
  - $\forall t,\ \mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$ where $\mathbf{A} \in \mathbb{R}^{M\times K}$ is called the "mixture matrix"
  - Sources are assumed IID: $p(\{s_k(t)\}_{k,t}) = \prod_{k=1}^{K}\prod_{t=1}^{T} p_k(s_k(t))$
- Problem: estimate $\mathbf{A}$ and sources $\mathbf{s}(t)$ given $\mathbf{x}(t)$
- Definition: non-mixing matrix
  - a matrix $\mathbf{C}$ of dimension $K \times K$ is non-mixing if and only if it has a unique non-zero entry in each row and each column
- If $\widetilde{\mathbf{s}}(t) = \mathbf{C}\mathbf{s}(t)$ and $\widetilde{\mathbf{A}} = \mathbf{A}\,\mathbf{C}^{-1}$, then $\mathbf{x}(t) = \widetilde{\mathbf{A}}\widetilde{\mathbf{s}}(t)$ is another admissible decomposition of the observations
  - Sources can be recovered up to a permutation and a multiplicative factor

Linear separation of sources

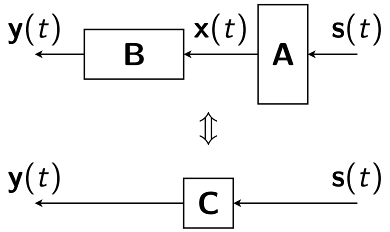
- Let $\mathbf{y}(t) = \mathbf{B}\mathbf{x}(t)$, where $\mathbf{B} \in \mathbb{R}^{K\times M}$ is referred to as the "separation matrix"
- Linear separation is feasible if $\mathbf{A}$ has rank K:
  - We get $\mathbf{y}(t) = \mathbf{s}(t)$ by defining:
    - $\mathbf{B} = \mathbf{A}^{-1}$ in the determined case ($M = K$)
    - $\mathbf{B} = \mathbf{A}^{\dagger}$ in the over-determined case ($M > K$)
  - the pseudo-inverse $\mathbf{A}^{\dagger} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ is such that $\mathbf{A}^{\dagger}\mathbf{A} = \mathbf{I}_K$
- In the under-determined case ($M < K$), separation is not feasible

Part IV

Independent component analysis

Independent component analysis (ICA)

- In practice matrix $\mathbf{A}$ is unknown:
  - We look for a matrix $\mathbf{B}$ that makes the y_k independent (ICA)
  - We then get equation $\mathbf{y}(t) = \mathbf{C}\mathbf{s}(t)$, where $\mathbf{C} = \mathbf{B}\mathbf{A}$
  - The problem is solved if matrix $\mathbf{C}$ is non-mixing



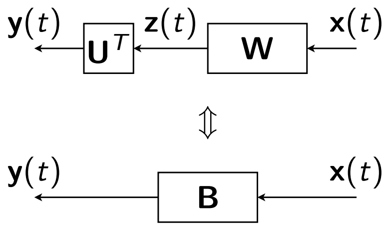
- Theorem (identifiability)
  - Let s_k be K IID sources, among which at most one is Gaussian, and $\mathbf{y}(t) = \mathbf{C}\mathbf{s}(t)$ with $\mathbf{C}$ invertible ((over)-determined case). If signals $y_k(t)$ are independent, then $\mathbf{C}$ is non-mixing.

Whitening

- We now suppose that the sources are centered: $\mathbb{E}[\mathbf{s}(t)] = \mathbf{0}$ and that the mixture is (over-)determined
- Canonical problem: we can assume without loss of generality that $\mathbf{s}(t)$ is spatially white ($\boldsymbol{\Sigma}_{ss} = \mathbb{E}[\mathbf{s}(t)\mathbf{s}(t)^T] = \mathbf{I}_K$)
- Then $\boldsymbol{\Sigma}_{xx} = \mathbf{A}\boldsymbol{\Sigma}_{ss}\mathbf{A}^T = \mathbf{A}\,\mathbf{A}^T$: $\mathbf{A}$ is a matrix square root of $\boldsymbol{\Sigma}_{xx}$
- We first aim to whiten (decorrelate) the mixture:
  - $\boldsymbol{\Sigma}_{xx}$ is diagonalizable in an orthonormal basis: $\boldsymbol{\Sigma}_{xx} = \mathbf{Q}\boldsymbol{\Lambda}^2\mathbf{Q}^T$ where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1 \dots \lambda_M)$ with $\lambda_1 \geq \lambda_K > \lambda_{K+1} = \lambda_M = 0$ (the rank of $\boldsymbol{\Sigma}_{xx}$ is equal to K)
  - Let $\mathbf{S} = \mathbf{Q}_{(:,1:K)}\boldsymbol{\Lambda}_{(1:K,1:K)} \in \mathbb{R}^{M \times K}$
  - $\mathbf{S}$ is a matrix square root of $\boldsymbol{\Sigma}_{xx}$: $\boldsymbol{\Sigma}_{xx} = \mathbf{S}\,\mathbf{S}^T$
  - Let $\mathbf{W} = \mathbf{S}^\dagger$ and $\mathbf{z}(t) = \mathbf{W}\mathbf{x}(t)$
  - Then $\mathbf{z}(t)$ is white ($\mathbb{E}[\mathbf{z}(t)] = \mathbf{0}$ and $\boldsymbol{\Sigma}_{zz} = \mathbf{W}\boldsymbol{\Sigma}_{xx}\mathbf{W}^T = \mathbf{I}$)

Whitening

- We conclude without loss of generality that $\mathbf{U} \triangleq \mathbf{W}\mathbf{A}$ is a rotation matrix ($\mathbf{U}\mathbf{U}^T = \mathbf{I}$).
- Then $\mathbf{y}(t) = \mathbf{U}^T\mathbf{z}(t) = \mathbf{U}^T\mathbf{W}\mathbf{x}(t) = (\mathbf{W}\mathbf{A})^{-1}(\mathbf{W}\mathbf{A})\mathbf{s}(t) = \mathbf{s}(t)$.
- We can thus assume $\mathbf{B} = \mathbf{U}^T\mathbf{W}$ where $\mathbf{U}$ is a rotation matrix.



Higher order statistics

- One can estimate $\boldsymbol{\Sigma}_{xx}$ from the observations and get $\mathbf{W}$
- The whiteness property (second order cumulants) determines $\mathbf{W}$ and leaves $\mathbf{U}$ unknown.
- If sources are Gaussian, the z_k are independent and $\mathbf{U}$ cannot be determined.
- In order to determine rotation $\mathbf{U}$, we need to exploit the non-Gaussianity of sources and characterize the independence property by using cumulants of order greater than 2.

Contrast functions

- Definition: ϕ is a "contrast function" if and only if $\phi[\mathbf{Cs}(t)] \geq \phi[\mathbf{s}(t)]\ \forall \mathbf{C}$ and if $\phi[\mathbf{Cs}(t)] = \phi[\mathbf{s}(t)] \Leftrightarrow \mathbf{C}$ is non-mixing.
- Separation is performed by minimizing $\phi[\mathbf{y}(t) = \mathbf{Cs}(t)]$ with respect to $\mathbf{U}$ (or $\mathbf{B}$)
- "Canonical" contrast function: $\phi_{IM}[\mathbf{y}(t)] = \mathbb{I}[\mathbf{y}(t)]$
- Orthogonal contrasts: to be minimized under the constraint $\mathbb{E}[\mathbf{y}(t)\mathbf{y}(t)^T] = \mathbf{I}$. For instance, $\phi^\circ_{IM}[\mathbf{y}(t)] = \sum_{k=1}^{K} \mathbb{H}(y_k(t))$
- Order 4 approximation of ϕ°_{IM}: $\phi^\circ_{ICA}[\mathbf{y}(t)] = \sum_{ijkl \neq iiii} (\kappa^4_{ijkl}[\mathbf{y}(t)])^2$
- Descent algorithms for minimizing ϕ with respect to $\mathbf{B}$ or $\mathbf{U}$:
  - Gradient algorithm applied to matrix $\mathbf{B}$
  - Parameterization of $\mathbf{U}$ with Givens rotations and coordinate descent

TELECOM Paris

Algorithm

1. Estimation of the covariance matrix $\mathbf{\Sigma}_{xx}$
2. Diagonalization of $\mathbf{\Sigma}_{xx}$: $\mathbf{\Sigma}_{xx} = \mathbf{Q}\mathbf{\Lambda}^2\mathbf{Q}^T$ where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1 \dots \lambda_M)$ with $\lambda_1 \geq \dots \geq \lambda_M \geq 0$
3. Computation of $\mathbf{S} = \mathbf{Q}_{(:,1:K)}\mathbf{\Lambda}_{(1:K,1:K)}$
4. Computation of the whitening matrix $\mathbf{W} = \mathbf{S}^\dagger$
5. Data whitening: $\mathbf{z}(t) = \mathbf{Wx}(t)$
6. Estimation of $\mathbf{U}$ by minimizing the contrast function ϕ°
7. Estimation of source signals via $\mathbf{y}(t) = \mathbf{U}^T\mathbf{z}(t)$

TELECOM Paris

Part V

Second order methods

TELECOM Paris

Temporal coherence of sources

- Model: $\mathbb{E}(\mathbf{s}(t)) = \mathbf{0}$, $\mathbf{R}_{ss}(\tau) = \mathbb{E}\left(\mathbf{s}(t+\tau)\mathbf{s}(t)^T\right) = \mathrm{diag}(r_{s_k}(\tau))$
- Canonical problem: we assume that $\mathbf{\Sigma}_{ss} = \mathbf{R}_{ss}(0) = \mathbf{I}$
- We first aim to spatially whiten the mixture:
  - Let $\mathbf{S}$ be a matrix square root of $\mathbf{\Sigma}_{xx}$
  - Let $\mathbf{W} = \mathbf{S}^\dagger$ and $\mathbf{z}(t) = \mathbf{Wx}(t)$
- Since $\mathbf{\Sigma}_{xx} = \mathbf{A}\mathbf{A}^T$, $\mathbf{U} \triangleq \mathbf{WA}$ is a rotation matrix
- However, $\forall \tau \in \mathbb{Z}$, $\mathbf{R}_{zz}(\tau) = \mathbf{U}\mathbf{R}_{ss}(\tau)\mathbf{U}^T$
- The joint diagonalization of matrices $\mathbf{R}_{zz}(\tau)$ for various values of τ permits us to identify rotation $\mathbf{U}$

TELECOM Paris

Joint diagonalization

- Unicity theorem :
  - Let a set of matrices $\mathbf{R}_{zz}(\tau)$ of dimension $K \times K$ and of the form $\mathbf{R}_{zz}(\tau) = \mathbf{U}\mathbf{R}_{ss}(\tau)\mathbf{U}^T$ with $\mathbf{U}$ unitary and $\mathbf{R}_{ss}(\tau) = \mathrm{diag}(r_{s_k}(\tau))$. Then $\mathbf{U}$ is unique (up to a non-mixing matrix) if and only if $\forall 1 \leq k \neq l \leq K$, there is τ such that $r_{s_k}(\tau) \neq r_{s_l}(\tau)$
- Joint diagonalization methods: minimize the criterion

$$J(\mathbf{U}) = \sum_\tau \|\mathbf{U}^T\mathbf{R}_{zz}(\tau)\mathbf{U} - \mathrm{diag}(\mathbf{U}^T\mathbf{R}_{zz}(\tau)\mathbf{U})\|_F^2$$

  - Parameterization of $\mathbf{U}$ with Givens rotations and coordinate descent

SOBI algorithm

- *Second Order Blind Identification (SOBI)*
  1. Estimation and diagonalization of $\mathbf{\Sigma}_{xx}$: $\mathbf{\Sigma}_{xx} = \mathbf{Q}\mathbf{\Lambda}^2\mathbf{Q}^T$ where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1 \dots \lambda_M)$ with $\lambda_1 \geq \dots \geq \lambda_M \geq 0$
  2. Computation of $\mathbf{S} = \mathbf{Q}_{(:,1:K)}\mathbf{\Lambda}_{(1:K,1:K)}$
  3. Computation of the whitening matrix $\mathbf{W} = \mathbf{S}^\dagger$
  4. Data whitening: $\mathbf{z}(t) = \mathbf{W}\mathbf{x}(t)$
  5. Estimation of covariance matrices $\mathbf{R}_{zz}(\tau)$ for various delays τ
  6. Approximate joint diagonalization of matrices $\mathbf{R}_{zz}(\tau)$ in a common basis $\mathbf{U}$
  7. Estimation of source signals via $\mathbf{y}(t) = \mathbf{U}^T\mathbf{z}(t)$

Non-stationarity of sources

- Model: $\mathbb{E}(\mathbf{s}(t)) = \mathbf{0}$, $\mathbf{\Sigma}_{ss}(t) \triangleq \mathbb{E}\left(\mathbf{s}(t)\mathbf{s}(t)^T\right) = \mathrm{diag}(\sigma_k^2(t))$
- Then $\forall t \in \mathbb{Z}$, $\mathbf{\Sigma}_{xx}(t) = \mathbf{A}\mathbf{\Sigma}_{ss}(t)\mathbf{A}^T$
- Joint diagonalization methods: minimize the criterion

$$J(\mathbf{B}) = \sum_t \|\mathbf{B}\mathbf{\Sigma}_{xx}(t)\mathbf{B}^T - \mathrm{diag}(\mathbf{B}\mathbf{\Sigma}_{xx}(t)\mathbf{B}^T)\|_F^2$$

  - Gradient descent algorithm applied to matrix $\mathbf{B}$
  - In the over-determined case, $\mathbf{B}$ must be constrained to span the principal subspace of all matrices $\mathbf{\Sigma}_{xx}(t)$
- Variant of the SOBI algorithm:
  1. Segmentation of source signals and estimation of covariance matrices $\mathbf{\Sigma}_{xx}(t)$ on windows centered at different times t
  2. Joint diagonalization of matrices $\mathbf{\Sigma}_{xx}(t)$ in a common basis $\mathbf{B}$
  3. Estimation of source signals via $\mathbf{y}(t) = \mathbf{B}\mathbf{x}(t)$

Conclusion of the first part

- The use of higher order cumulants is only necessary for the non-Gaussian IID source model
- Second order statistics are sufficient for sources that are:
  - stationary but not IID ($\rightarrow$ spectral dynamics)
  - non stationary ($\rightarrow$ temporal dynamics)
- Remember that classical tools (based on second order statistics) are appropriate for blind separation of independent (and possibly Gaussian) sources, on condition that the spectral / temporal source dynamics is taken into account.

Part VI

Time-frequency methods

Time-frequency representations

- Motivations
  - Spectral and temporal dynamics are highlighted by a time-frequency (TF) representation of signals
  - TF representations are appropriate to process convolutive and/or under-determined mixtures
- Use of a filter bank (examples: STFT, MDCT):
  - Decomposition in F sub-bands and decimation of factor $T \leq F$
  - Analysis filters h_f and synthesis filters g_f
  - TF representation of sources: $s_k(f,n) = (h_f * s_k)(nT)$
  - TF representation of mixtures: $x_m(f,n) = (h_f * x_m)(nT)$
  - Perfect reconstruction: $s_k(t) = \sum_{f=1}^{F} \sum_{n \in \mathbb{Z}} g_f(t - nT) s_k(f,n)$
- Then $\forall f, n$, $\mathbf{x}(f,n) = \mathbf{A}\,\mathbf{s}(f,n)$ (same linear instantaneous mixture)

Non-stationary source model

- Assumption: independent and centered second order processes
  - Model of **non-stationary sources**:
    - if the time frames n_1 and n_2 are disjoint, then $s_k(.,n_1)$ and $s_k(.,n_2)$ are uncorrelated and of distinct variances
  - Model of **WSS sources**:
    - if sub-bands f_1 and f_2 are disjoint ($h_{f_1} * \widetilde{h}_{f_2} = 0$), then $s_k(f_1,.)$ and $s_k(f_2,.)$ are WSS, uncorrelated and of distinct variances $\sigma_k^2(f_1) = (h_{f_1} * \widetilde{h}_{f_1} * r_{s_k})(0)$ and $\sigma_k^2(f_2) = (h_{f_2} * \widetilde{h}_{f_2} * r_{s_k})(0)$
  - **Time-frequency source model**:
    - all $s_k(f,n)$ are uncorrelated for all n and f, of distinct variances $\sigma_k^2(f,n)$ ($\Rightarrow$ time-frequency dynamics)

Separation method

- Separation by joint matrix diagonalization:
  - Let $\mathbf{\Sigma}_{ss}(f,n) = \mathbb{E}[\mathbf{s}(f,n)\mathbf{s}(f,n)^T]$ and $\mathbf{\Sigma}_{xx}(f,n) = \mathbb{E}[\mathbf{x}(f,n)\mathbf{x}(f,n)^T]$
  - Then $\mathbf{\Sigma}_{xx}(f,n) = \mathbf{A}\mathbf{\Sigma}_{ss}(f,n)\mathbf{A}^T$ where $\mathbf{\Sigma}_{ss}(f,n) = \mathrm{diag}(\sigma_k^2(f,n))$
- Variant of the SOBI algorithm:
  1. TF analysis of the mixtures: $x_k(f,n) = (h_f * x_k)(nT)$
  2. Estimation of covariance matrices $\mathbf{\Sigma}_{xx}(f,n)$
  3. Joint diagonalization of matrices $\mathbf{\Sigma}_{xx}(f,n)$ in a basis $\mathbf{B}$
  4. Estimation of the source signals via $\mathbf{y}(f,n) = \mathbf{B}\,\mathbf{x}(f,n)$
  5. TF synthesis of the sources: $y_k(t) = \sum_{f=1}^{F} \sum_{n \in \mathbb{Z}} g_f(t - nT) y_k(f,n)$

Part VII

Convolutive mixtures

Convolutive linear mixture

- Instantaneous mixture model unsuitable for real acoustic mixtures
- Let $\mathbf{x}_k(f,n) \in \mathbb{R}^M$ be the **image** of source $s_k(f,n)$
  - received multichannel signal if only source $s_k(f,n)$ was active
- Mixture model: $\mathbf{x}(f,n) = \sum_{k=1}^{K} \mathbf{x}_k(f,n)$
- Decomposition of the source separation problem
  - **separation**: estimate $\mathbf{x}_k(f,n)$ from the mixture $\mathbf{x}(f,n)$
  - **deconvolution**: estimate $s_k(f,n)$ from $\mathbf{x}_k(f,n)$
- Mixture model: $x_m(t) = \sum_{k=1}^{K} (a_{mk} * s_k)(t)$, i.e. $\mathbf{x}(t) = \mathbf{A} * \mathbf{s}(t)$
- Theorem (identifiability)
  - Let s_k be K IID sources, among which at most one is Gaussian, and $\mathbf{y}(t) = \mathbf{C} * \mathbf{s}(t)$ with $\mathbf{C}$ invertible ((over)-determined case). If signals $y_k(t)$ are independent, then $\mathbf{C}$ is non-mixing.

Time-frequency approach

- Mixture model: $x_m(t) = \sum_{k=1}^{K} (a_{mk} * s_k)(t)$
- Assumptions:
  - the filter bank corresponds to an STFT
  - the IR of a_{mk} is short compared with the window length
  - $\forall m, k, f,\ a_{mk}(\nu)$ varies slowly compared with $h_f(\nu)$
  - $\Rightarrow (h_f * a_{mk})(t) \approx a_{mk}(f)\, h_f(t)$
- Approximation of the convolutive mixture model:

$$x_m(f,n) = \sum_{k=1}^{K} a_{mk}(f) s_k(f,n)$$

- Matrix form: $\mathbf{x}(f,n) = \mathbf{A}(f)\mathbf{s}(f,n)$
  - F instantaneous mixture models in every sub-band
  - $\Rightarrow$ we can use any ICA method in every sub-band

Independent component analysis

- Let $\mathbf{y}(f,n) = \mathbf{B}(f)\mathbf{x}(f,n)$, where $\mathbf{B}(f) \in \mathbb{C}^{K \times M}$
- Linear separation is feasible if $\mathbf{A}(f)$ has rank K:
  - We get $\mathbf{y}(f,n) = \mathbf{s}(f,n)$ by defining:
    - $\mathbf{B}(f) = \mathbf{A}(f)^{-1}$ in the determined case ($M = K$)
    - $\mathbf{B}(f) = \mathbf{A}(f)^{\dagger}$ in the over-determined case ($M > K$)
- In the under-determined case ($M < K$), separation remains impossible
- In practice matrix $\mathbf{A}(f)$ is unknown:
  - We look for $\mathbf{B}(f)$ that makes the $y_k(f,n)$ independent (ICA)
  - We then get $\mathbf{y}(f,n) = \mathbf{C}(f)\mathbf{s}(f,n)$, where $\mathbf{C}(f) = \mathbf{B}(f)\mathbf{A}(f)$
  - $\mathbf{C}(f)$ is non-mixing

Indeterminacies

- Problem: indeterminacies (permutations and multiplicative factors) in matrices $\mathbf{C}(f)$
  - $\forall k$, identify indexes k_f such that $\forall f$, $y_{k_f}(f,n) = c_{k_f,k} s_k(f,n)$
  - identify the multiplicative factors $c_{k_f,k}$
- Infinitely many solutions $\Rightarrow$ need to constrain the model:
  - Assumptions on the mixture
    - continuity of the frequency responses $a_{mk}(f)$ with respect to f
    - $\rightarrow$ beamforming model and anechoic model
  - Assumptions on the sources
    - similarity of the temporal dynamics of $\sigma_k^2(f,n)$

Convolutive mixture models

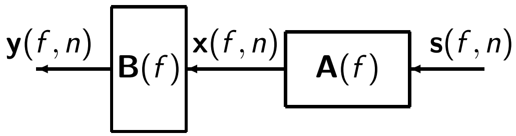
- **Beamforming** model:
  - Assumptions: plane waves, far field, no reverberation, linear antenna
  - Model: $a_{mk}(f) = e^{-2i\pi f \tau_{mk}}$ where $\tau_{mk} = \frac{d_m}{c} \sin(\theta_k)$
  - Parameters: positions d_m of the sensors and angles θ_k of the sources
- **Anechoic** model:
  - Assumptions: punctual sources, no reverberation
  - Model: $a_{mk}(f) = \alpha_{mk} e^{-2i\pi f \tau_{mk}}$ where $\alpha_{mk} = \frac{1}{\sqrt{4\pi} r_{mk}}$ and $\tau_{mk} = \frac{r_{mk}}{c}$
  - Parameters: distances r_{mk} between the sensors and sources

Part VIII

Under-determined mixtures

Under-determined convolutive mixtures

- Usual case in audio: monophonic ($M = 1$) or stereophonic ($M = 2$) mixtures, with a number of sources $K > M$
- Convolutive mixture model: $\mathbf{x}(f,n) = \mathbf{A}(f)\mathbf{s}(f,n)$ with $M < K$
- Assumption: the mixture model $\mathbf{A}(f)$ and the source model $\mathbf{\Sigma}_{ss}(f,n)$ are known
- Even in this case, the exact separation is not feasible, because there is no matrix $\mathbf{B}(f)$ such that $\mathbf{B}(f)\mathbf{A}(f) = \mathbf{I}_K$



Separation via non-stationary filtering

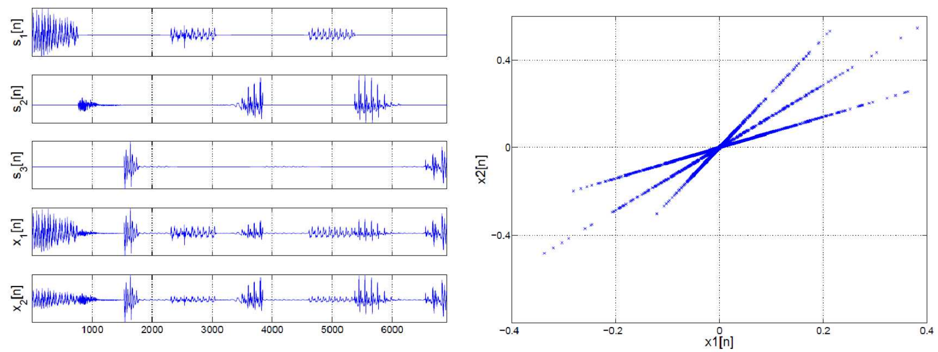
- Let $\mathbf{y}(f,n) = \mathbf{B}(f,n)\mathbf{x}(f,n)$ where $\mathbf{B}(f,n) \in \mathbb{C}^{K \times M}$ depends on n
- *Minimum Mean Square Error* (MMSE) estimator: we look for $\mathbf{B}(f,n)$ which minimizes $\mathbb{E}[\|\mathbf{y}(f,n) - \mathbf{s}(f,n)\|_2^2]$
- Solution: $\mathbf{B}(f,n) = \mathbf{\Sigma}_{sx}(f,n)\mathbf{\Sigma}_{xx}(f,n)^{-1}$ where $\mathbf{\Sigma}_{xx}(f,n) = \mathbf{A}(f)\mathbf{\Sigma}_{ss}(f,n)\mathbf{A}(f)^H$ and $\mathbf{\Sigma}_{sx}(f,n) = \mathbf{\Sigma}_{ss}(f,n)\mathbf{A}(f)^H$ ($(.)^H$ denotes the Hermitian conjugate)
- Remark: $\mathbf{x}(f,n) = \mathbf{A}(f)\mathbf{y}(f,n)$ (exact reconstruction)
- Particular case: monophonic mixtures
  - Without loss of generality, we define $\mathbf{A}(f) = [1,\ldots,1]$
  - We get $y_k(f,n) = \frac{\sigma_k^2(f,n)}{\sum_{l=1}^K \sigma_l^2(f,n)} x(f,n)$
  - $\Rightarrow$ similar to Wiener filtering

Separation method

1. TF analysis of the mixtures: $x_k(f,n) = (h_f * x_k)(f,nT)$
2. Estimation of $\mathbf{A}(f)$ and $\sigma_k^2(f,n)$
   - **instantaneous mixture model**
   - **sparse source model**
3. Computation of $\mathbf{B}(f,n) = \mathbf{\Sigma}_{ss}(f,n)\mathbf{A}(f)^H \left(\mathbf{A}(f)\mathbf{\Sigma}_{ss}(f,n)\mathbf{A}(f)^H\right)^{-1}$
4. Estimation of the source signals via $\mathbf{y}(f,n) = \mathbf{B}(f,n)\mathbf{x}(f,n)$
5. TF synthesis of the sources: $y_k(t) = \sum_{f=1}^F \sum_{n\in\mathbb{Z}} g_f(t-nT)y_k(f,n)$

Stereophonic mixtures: temporal sparsity

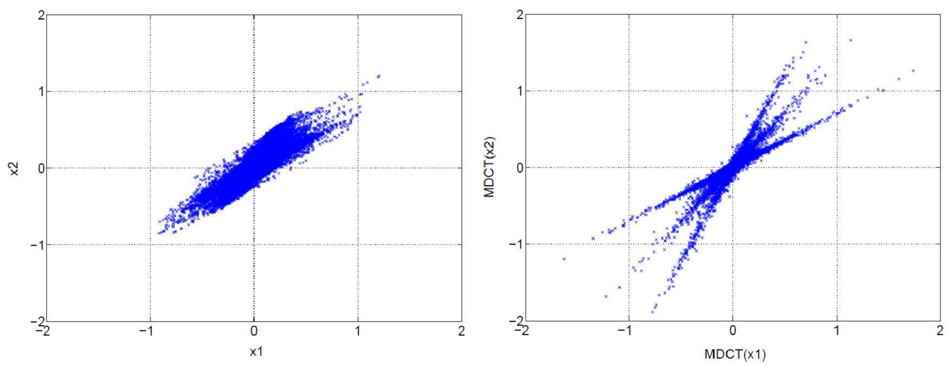
Case of a linear instantaneous stereophonic mixture: $\mathbf{x}(t) = \mathbf{A}\,\mathbf{s}(t)$



(a) Temporal source signals and corresponding stereo mixture (b) Dispersion diagrams (x_1, x_2) over time

Sparsity in a transformed domain

Case of a linear instantaneous stereophonic mixture: $\mathbf{x}(f,n) = \mathbf{A}\,\mathbf{s}(f,n)$



(a) Time samples (b) Time-frequency coefficients after MDCT decomposition

DUET method

- *Degenerate Unmixing Estimation Technique* (DUET)
- Linear instantaneous stereophonic mixture model: $\mathbf{x}(f,n) = \mathbf{A}\,\mathbf{s}(f,n)$
  - Without loss of generality, we assume $\mathbf{A}_{(:,k)} = \begin{bmatrix} \cos(\theta_k) \\ \sin(\theta_k) \end{bmatrix}$ $\forall k$
- **Sparse** source model:
  - $\forall f,n$, $\exists!\ k_{(f,n)}$ such that $\sigma^2_{k_{(f,n)}}(f,n) > 0$, and $\forall l \neq k_{(f,n)}$, $\sigma_l^2(f,n) = 0$
- If only source k is active at (f,n), then $\mathbf{x}(f,n) = \mathbf{a}_k s_k(f,n)$

DUET method

1. TF analysis of the mixtures: $x_k(f,n) = (h_f * x_k)(nT)$
2. Estimation of parameters θ_k and of the active source $k_{(f,n)}$
   - computation of the histogram of the angles of vectors $\mathbf{x}(f,n)$
   - peak detection in order to estimate parameters θ_k
   - determination of the active source at (f,n) by proximity with θ_k
3. Source separation: for all k,
   - estimation of source images via binary masking: $\mathbf{y}_k(f,n) = \mathbf{x}(f,n)$ $\forall$ (f,n) such that $k_{(f,n)} = k$ and $\mathbf{y}_k(f,n) = 0$ for the other time-frequency bins (f,n)
   - MMSE estimation of the sources: $y_k(f,n) = \hat{\mathbf{a}}_k(f)^\dagger \mathbf{y}_k(f,n)$
4. TF synthesis of the sources: $y_k(t) = \sum_{f=1}^{F} \sum_{n\in\mathbb{Z}} g_f(t - nT) y_k(f,n)$

Part IX

Conclusion

Conclusion

- Summary
  - Source separation requires to make assumptions about the mixture and sources
  - For an (over-)determined instantaneous linear mixture, the assumption of independent sources is sufficient
  - In all other cases, we need to model the mixture and/or the sources
- Perspectives
  - Non-stationary mixtures (adaptive algorithms)
  - Informed source separation (e.g. from music score)
  - Deep learning techniques
  - Objective assessment of audio source separation