Mixing models for multichannel audio source separation in reverberant conditions

Simon LEGLAIVE

Mid-term PhD presentation
March 31, 2016, Télécom ParisTech
Introduction

Multichannel audio source separation in reverberant conditions

sources: latent random variables
mixing and source parameters
observed data
Outline

Context

Convolutional mixing model
Source model
Parameters estimation

Reverberation models

Anechoic model
Early contributions model
Late reverberation model

Future works
Convolutive mixing model

Convolutive noisy mixture of $J$ sources on $I$ channels

STFT domain:
$\forall (f, n) \in [0, F - 1] \times [0, N - 1]$

$\mathbf{x}_{i,fn} = \sum_{j=1}^{J} a_{ij,f} \mathbf{s}_{j,fn} + b_{i,fn}$

- Frequency response of mixing filters: $a_{ij,f}$
- Source STFTs: $s_{j,fn}$
- Additive noise: $b_{i,fn} \sim \mathcal{N}(0, \sigma_{b,f}^2)$
Source model

NMF source model [Févotte, Bertin and Durrieu, 2009]

\[ s_{j,fn} = \sum_{k \in K_j} c_{k,fn} \quad \text{avec} \quad c_{k,fn} \sim \mathcal{N}_c(0, w_{fk} h_{kn}) \]

\[ s_{j,fn} \sim \mathcal{N}_c(0, v_{j,fn} = \sum_{k \in K_j} w_{fk} h_{kn}) \]

\[ V_j = [v_{j,fn}]_{fn} \quad W_j = [w_{fk}]_{fk} \quad H_j = [h_{kn}]_{kn} \]
Parameters estimation

We want to estimate the set of parameters:

\[ \eta = \left\{ \mathbf{A} = \{a_{ij,f}\}, \mathbf{W} = \{w_{fk}\}, \mathbf{H} = \{h_{kn}\}, \mathbf{\Sigma}_b = \{\sigma^2_{b,f}\} \right\} \]

- Maximum Likelihood (ML) estimation [Ozerov et Févotte, 2010]
  \[ \eta^{ML} = \arg \max_{\eta} \ p(\mathbf{X}|\eta) \]

- Maximum A Posteriori (MAP) estimation
  \[ \eta^{MAP} = \arg \max_{\eta} \ p(\mathbf{X}|\eta)p(\mathbf{A}) \]

- Expectation-Maximization (EM) algorithm
Outline

Context
  - Convolutive mixing model
  - Source model
  - Parameters estimation

Reverberation models
  - Anechoic model
  - Early contributions model
  - Late reverberation model

Future works
Room Impulse response
Anechoic model
Anechoic model

Direct path: attenuation $\rho_{ij}$ and delay $\tau_{ij}$

$$a_{ij,f} = \rho_{ij} \delta^f_{ij} \quad \text{with} \quad \delta_{ij} = e^{-j2\pi \tau_{ij}}, \quad (1)$$

so \( \{a_{ij,f}\}_f \) satisfies

$$a_{ij,f} = \delta_{ij} a_{ij,f-1} \quad (2)$$

Adding an error term $\rightarrow$ Markov chain model

$$a_{ij,f} = \delta_{ij} a_{ij,f-1} + \epsilon_f \quad \text{with} \quad \epsilon_f \sim \mathcal{N}_c(0, \sigma^2_a) \quad (3)$$

Prior distribution

$$\ln p(A) = -IJ(F - 1) \ln(\pi \sigma^2_a) - \frac{1}{\sigma^2_a} \sum_{f=1}^{F-1} \| A_f - \Delta \cdot A_{f-1} \|^2_F \quad (4)$$

with $\Delta \in \mathbb{C}^{I \times J}$ the matrix of entries $\delta_{ij}$
**EM algorithm**

Complete data: \( \{ \mathbf{X} = \{x_i, fn\}, \mathbf{C} = \{c_k, fn\} \} \)

**E-step**

\[
Q(\eta | \eta') = \mathbb{E}_{\mathbf{C} | \mathbf{X}, \eta'} \left[ \ln p(\mathbf{X}, \mathbf{C} | \eta) \right]
\]

**M-step**

**ML estimation:**

\[
\eta^{ML} = \arg \max_{\eta} Q(\eta | \eta')
\]

**MAP estimation:**

\[
\eta^{MAP} = \arg \max_{\eta} Q(\eta | \eta') + \ln p(\mathbf{A})
\]

⇒ ML and MAP estimations only differ in the mixing filters update at the M-step
5 stereo mixtures of 3 sources: 3 synthetic and 2 live recordings

- **SIR (Signal to Interference Ratio)**
- **SAR (Signal to Artifact Ratio)**
- **ISR (Image to Source Spatial Distortion Ratio)**
- **SDR (Signal to Distortion Ratio)**

Published in *Colloque GRETSI*, 2015.
Early contributions model
Early contributions model

$k^{th}$ contribution: attenuation $\rho_{kij}$ and delay $\tau_{kij}$

\[ a_{ij,f} = \sum_{k=0}^{R-1} \rho_{kij} \delta_{kij}^f \quad \text{with} \quad \delta_{kij} = e^{-j2\pi \tau_{kij}} \]  \hspace{1cm} (5)

It follows [Kumaresan, 1983]

\[ \{a_{ij,f}\}_f \quad \text{satisfies} \quad \sum_{r=0}^{R} \varphi_{rij} a_{ij,f-r} = 0 \]  \hspace{1cm} (6)

such that $\{\varphi_{Rij}, ..., \varphi_{0ij}\}$ and $\{\delta_{kij}\}_{k=0}^{R-1}$ are the coefficients and roots of a polynomial of order $R$. 
Early contributions model

Adding an error term → autoregressive model

\[
\sum_{r=0}^{R} \varphi_{rij} a_{ij,f-r} = \epsilon_f \quad \text{with} \quad \epsilon_f \sim \mathcal{N}_c(0, \sigma_{ij}^2)
\] (7)

Prior distribution

\[
\ln p(A) = - \sum_{i=1}^{I} \sum_{j=1}^{J} \left( (F - R) \ln(\pi \sigma_{ij}^2) + \frac{1}{\sigma_{ij}^2} \sum_{f=R}^{F-1} \left| \sum_{r=0}^{R} \varphi_{rij} a_{ij,f-r} \right|^2 \right)
\] (8)

Parameters estimation → same principle as before.
## Results

8 stereo mixtures of 3 sources: 4 synthetic and 4 live recordings

<table>
<thead>
<tr>
<th></th>
<th>Synthetic mixtures</th>
<th>Live recordings</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIR (in dB)</td>
<td>0.66</td>
<td>-0.31</td>
</tr>
<tr>
<td>SAR (in dB)</td>
<td>2.71</td>
<td>0.45</td>
</tr>
<tr>
<td>ISR (in dB)</td>
<td>9.13</td>
<td>0.40</td>
</tr>
<tr>
<td>SDR (in dB)</td>
<td>4.25</td>
<td>6.77</td>
</tr>
</tbody>
</table>

Signal to Interference Ratio

Signal to Artifact Ratio

Source Image to Spatial distortion Ratio

Signal to Distortion Ratio

Published in WASPAA, 2015.
Late reverberation model

- Contributions précoces
- Trajet direct
- Premiers échos
- Réverbération tardive

Graph showing the late reverberation model with amplitude on the y-axis and time on the x-axis.
RIR and RFR

Room impulse and frequency responses

For $t, k \in [0, T - 1]$:

$$h(t) = h_e(t) + h_l(t)$$

Room impulse response (RIR)

$$\mathcal{F}_T: \text{Discrete Fourier Transform (DFT)}$$

$$H(k) = H_e(k) + H_l(k)$$

Room frequency response (RFR)

Mixing time:

$$t_0 = \left\lfloor C_0 \sqrt{V f_s} \right\rfloor \text{ samples}$$

- $C_0 = 2 \times 10^{-3}$
- $V$: volume of the room
- $f_s$: sampling rate
Statistical model of late reverberation

Theory of statistical room acoustics

\( \{H_i(k)\}_k \) is a centered and WSS complex Gaussian random process

- **Autocovariance function (ACVF):**
  \[
  \gamma(m) = \mathbb{E}[H_i(k)H_i(k - m)^*]
  \]

- **Power Spectral Density (PSD):**
  \[
  \phi(t) = \frac{1}{T} \mathbb{E}[|\mathcal{F}_T \{H_i(k)\}|^2]
  \]

- **Objective**: To theoretically define these quantities

- **Starting point**: The exponential decay of the reverberation power induces specific frequency correlations [Schroeder 1962]
  \Rightarrow exponential decay only valid for late reverberation
**Temporal dynamics and total power**

Power Temporal Profil (PTP) → exponential decay

\[
\tilde{h}_l(t) = \mathbb{E}[|h_l(t)|^2] = P_0^2 e^{-2t/\tau_1} \mathbb{1}_{t \geq t_0}(t)
\]

- \(P_0^2\) is related to the total power of late reverberation
- \(\tau = \frac{T_{60} f_s}{3 \ln(10)}\) samples
- \(\mathbb{E}[\cdot]\): spatial averaging

**Variance of the late part of the RFR**

\[
\sigma_{rev}^2 := \mathbb{E}[|H_l(k)|^2] = C_{rev} \frac{1 - \alpha}{\pi \alpha S}
\]

- \(\alpha\): average absorption coefficient (without dimension)
- \(S\): total wall area (m²)
- theoretically \(C_{rev} = 1\), empirically \(C_{rev} = 75\)
Statistical characterization of \( \{ H_l(k) \} \): 

**Definition of the PSD according to the PTP**

We can show that:

\[
\phi(t) = T \bar{h}_l(T - t)
\]  

(9)

**Wiener-Khinchin theorem → theoretical ACVF**

It follows:

\[
\gamma(m) = \mathcal{F}_T^{-1} \{ \phi(t) \}
= \sigma_{rev}^2 \frac{1 - e^{2/\tau}}{1 - e^{2(T-t_0+1)/\tau}} \frac{1 - e^{(j2\pi m/T+2/\tau)(T-t_0+1)}}{1 - e^{j2\pi m/T+2/\tau}}
\]  

(10)

\( \Rightarrow \) theoretical expression depending on some room parameters
Experimental validation

Empirical autocovariance functions computed from a Monte-Carlo simulation on synthesized and real room responses

(a) 196 simulated RIRs
$T_{60} = 0.25 \text{ s}$
$10 \times 6.6 \times 3.3 \text{ m}$

(b) 130 real RIRs
$T_{60} = 1.8 \text{ s}$
$7.5 \times 9 \times 3.5 \text{ m}$

Figure 1: Theoretical and empirical autocovariance functions
ARMA parametrization

ARMA representation of late reverberation in the frequency domain

We assume that \( \{H_l(k)\} \) follows an ARMA\((P, Q)\) model:

\[
\Phi(L)H_l(k) = \Theta(L)\epsilon(k)
\]

\(\Phi(L) = \sum_{p=0}^{P} \varphi_p L^p\) and \(\Theta(L) = \sum_{q=0}^{Q} \theta_q L^q\) with \(\varphi_0 = \theta_0 = 1\)

\(L\) is the lag operator, i.e. \(LH_l(k) = H_l(k-1)\)

\(\epsilon(k)\): centered complex white Gaussian noise

ARMA parametrization of the PSD and ACVF

We can compute the ARMA parameters from the theoretical ACVF
Experimental validation

Same room parameters as used before for simulated RIRs

---

**Figure 2:** ARMA(7,2) parametrization

**Figure 3:** Synthesized late RIR

▶ Submitted to *EUSIPCO*, 2016.
## Outline

### Context
- Convolutive mixing model
- Source model
- Parameters estimation

### Reverberation models
- Anechoic model
- Early contributions model
- Late reverberation model

### Future works
Future works

Global reverberation model for source separation

- Early reverberation: AR prior
  - as presented before

- Late reverberation: ARMA prior
  - fixed hyper-parameters according to some room parameters (reverberation time, room dimensions)
  - informed source separation
  - conjugate gradient method in M-step
**Future works**

### Time domain mixing model

Write the mixture in the time domain...

\[
x_i(t) = \sum_{j=1}^{J} [a_{ij} \ast s_j](t) + b_i(t)
\]

while keeping TF modeling of the sources...

\[
s_j(t) = \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} s_{j,fn} \tilde{\psi}_{fn}(t) \quad \text{with} \quad \tilde{\psi}_{fn}(t) \quad \text{a TF synthesis atom}
\]

in order to release the approximation of a convolutive mixing by an instantaneous one in each frequency band.
Thank you