Institut Mines-Télécom
Télécom ParisTech - CNRS LTCI - Paris, France

Multichannel audio source separation with probabilistic reverberation modeling

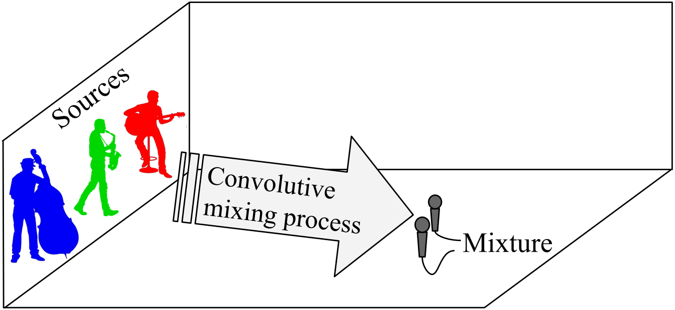
Simon LEGLAIVE, Roland BADEAU, Gaël RICHARD

October 19, 2015

IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), October 18-21, 2015, New Paltz, New York

Introduction

Multichannel audio source separation in reverberant conditions



Objective: Estimate the source and mixing parameters

Outline

Models
- Mixing model
- Source model
- Source separation problem

Reverberation modeling
- Room impulse response
- Early contributions
- Autoregressive model

Parameter estimation
- EM algorithm
- Hyper-parameters

Experiments
- Database and configuration
- Results
- Audio example

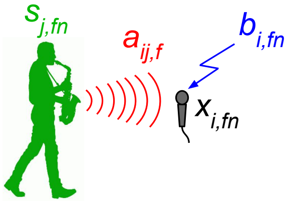
Conclusions

Mixing model

Convolutive noisy mixture of J sources on I channels

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

$$\boxed{x_{i,fn} = \sum_{j=1}^{J} a_{ij,f}\, s_{j,fn} + b_{i,fn}}$$



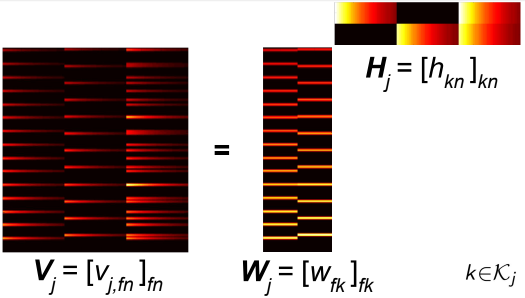
- Frequency response of mixing filters: $a_{ij,f}$
- Sources: $s_{j,fn}$
- Additive noise: $b_{i,fn} \sim \mathcal{N}_c(0, \sigma_f^2)$

Source model

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

$$s_{j,fn} = \sum_{k \in \mathcal{K}_j} c_{k,fn} \quad \text{with} \quad c_{k,fn} \sim \mathcal{N}_c(0, w_{fk} h_{kn})$$

$$s_{j,fn} \sim \mathcal{N}_c\left(0, v_{j,fn} = \sum_{k \in \mathcal{K}_j} w_{fk} h_{kn}\right)$$



Baseline approach

[Ozerov and Févotte, 2010]:

- Estimate $\boldsymbol{\theta}$ from $\mathbf{X} = \{x_{i,fn}\}$

$$\boldsymbol{\theta} = \left\{\mathbf{A} = \{a_{ij,f}\}, \mathbf{W} = \{w_{fk}\}, \mathbf{H} = \{h_{kn}\}, \boldsymbol{\Sigma}_{\mathbf{b}} = \{\sigma_f^2\}\right\}$$

- Maximum Likelihood (ML) estimation

$$\boldsymbol{\theta}_{ML} = \arg\max_{\boldsymbol{\theta}} \; p(\mathbf{X}; \boldsymbol{\theta})$$

- Expectation-Maximization (EM) algorithm

Our approach

- Use the structure of the mixing filters to constrain their estimation
- Early contributions of Room Impulse Responses $\Rightarrow p(\mathbf{A})$
- Maximum A Posteriori (MAP) estimation

$$\boldsymbol{\theta}_{MAP} = \arg\max_{\boldsymbol{\theta}} \; p(\mathbf{X}|\boldsymbol{\theta})p(\mathbf{A})$$

- Expectation-Maximization (EM) algorithm

Outline

Models
- Mixing model
- Source model
- Source separation problem

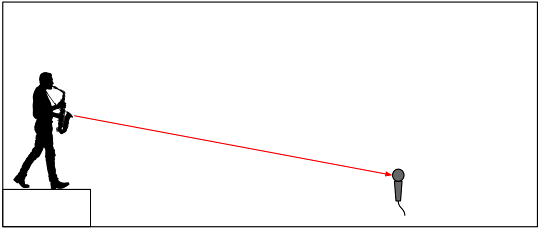
Reverberation modeling
- Room impulse response
- Early contributions
- Autoregressive model

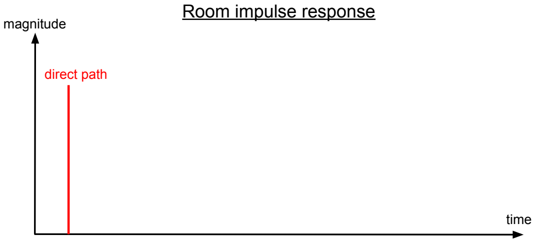
Parameter estimation
- EM algorithm
- Hyper-parameters

Experiments
- Database and configuration
- Results
- Audio example

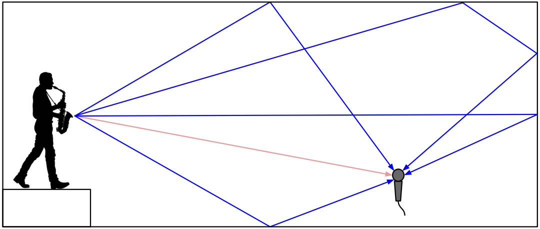
Conclusions

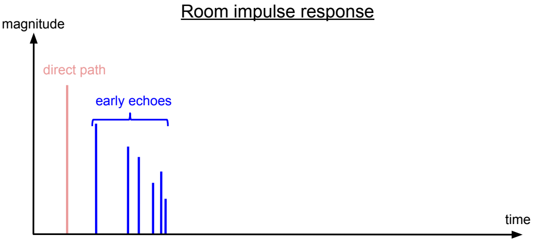
Room impulse response



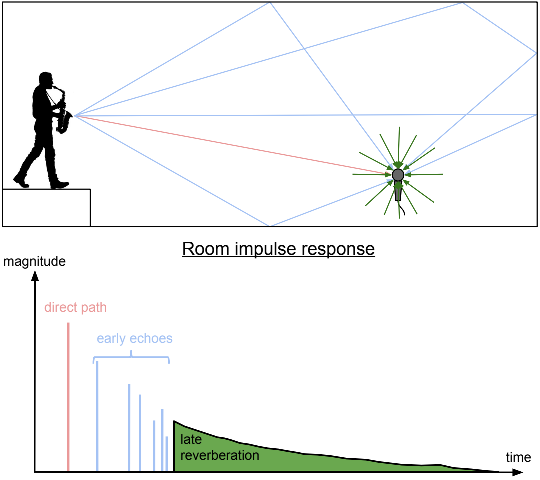


Room impulse response



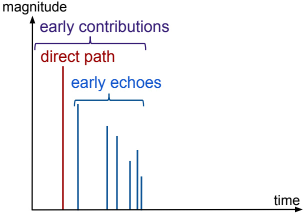


Room impulse response



Early contributions

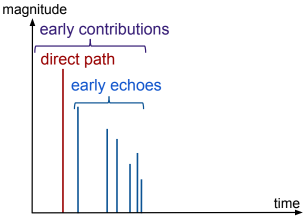
k^{th} early contribution: attenuation ρ_{kij} and delay τ_{kij}



$$a_{ij,f} = \sum_{k=0}^{R-1} \rho_{kij} \delta_{kij}^{f} \quad \text{with} \quad \delta_{kij} = e^{-j2\pi\tau_{kij}}$$

Early contributions

k^{th} early contribution: attenuation ρ_{kij} and delay τ_{kij}



$$a_{ij,f} = \sum_{k=0}^{R-1} \rho_{kij} \delta_{kij}^{f} \quad \text{with} \quad \delta_{kij} = e^{-j2\pi\tau_{kij}}$$

It follows (see, e.g, [Kumaresan, 1983])

$$\{a_{ij,f}\}_{f=R+1,\dots,F} \qquad \text{satisfies} \qquad \sum_{r=0}^{R} \varphi_{rij} a_{ij,f-r} = 0$$

Autoregressive model

Adding an error term

$$\sum_{r=0}^{R} \varphi_{rij} a_{ij,f-r} = \epsilon_{ij,f} \qquad \text{with} \qquad \epsilon_{ij,f} \sim \mathcal{N}_c(0, \sigma_{ij}^2)$$

$\{a_{ij,f}\}_f$ is an autoregressive process of order R

- Autoregressive coefficients: $\boldsymbol{\varphi}_{ij} = [\varphi_{0ij}, \ldots, \varphi_{Rij}]^T$ with $\varphi_{0ij} = 1$
- Error variance: σ_{ij}^2

Prior distribution

$$-\log p(\mathbf{A}) \stackrel{c}{=} \sum_{i,j} \left((F-R) \log \sigma_{ij}^2 + \frac{1}{\sigma_{ij}^2} \sum_{f=R+1}^{F} \left| \sum_{r=0}^{R} \varphi_{rij} a_{ij,f-r} \right|^2 \right)$$

Outline

Models
- Mixing model
- Source model
- Source separation problem

Reverberation modeling
- Room impulse response
- Early contributions
- Autoregressive model

Parameter estimation
- EM algorithm
- Hyper-parameters

Experiments
- Database and configuration
- Results
- Audio example

Conclusions

EM algorithm

Complete data: $\Big\{\mathbf{X} = \{x_{i,fn}\}, \mathbf{C} = \{c_{k,fn}\}\Big\}$

E-step

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{OLD}) = \mathbb{E}_{\mathbf{C}|\mathbf{X},\boldsymbol{\theta}^{OLD}}\Big[-\log p(\mathbf{X},\mathbf{C}|\boldsymbol{\theta})\Big]$$

M-step

ML estimation:

$$\boldsymbol{\theta}_{ML}^{NEW} = \arg\min_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{OLD})$$

MAP estimation:

$$\boldsymbol{\theta}_{MAP}^{NEW} = \arg\min_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{OLD}) - \log p(\mathbf{A})$$

$\Rightarrow$ ML and MAP estimations only differ in the update of **A** at M-step

Hyper-parameters

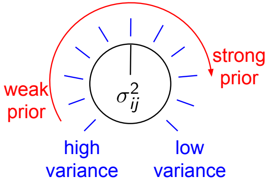
Hyper-parameters of the prior

- Autoregressive coefficients: $\boldsymbol{\varphi}_{ij} = [\varphi_{0ij}, \ldots, \varphi_{Rij}]^T$ with $\varphi_{0ij} = 1$
- Error variance: σ_{ij}^2

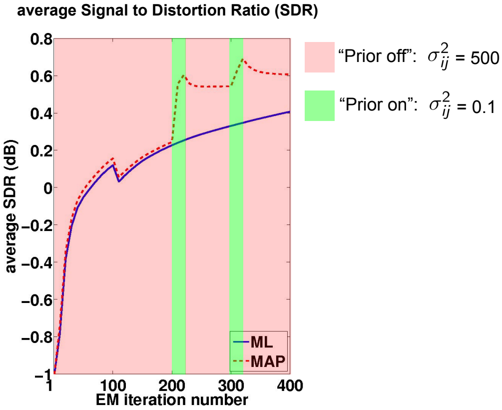
$\rightarrow$ Estimation by maximization of $\log p(\mathbf{A})$

Control of the strength of the prior

By tuning the value of the error variances, we can control the strength of the prior.



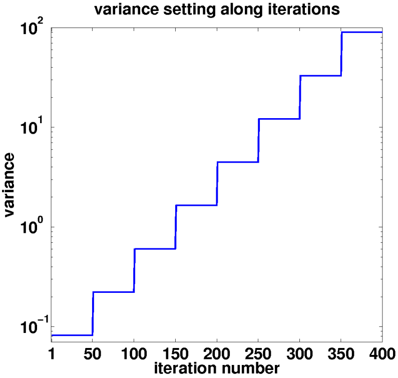
Illustrative example



We have to constrain the variances in order to have a significant influence of the prior, but which setting ?

Strategy

Empirically, a good strategy is to force a strong prior at the beginning of the EM algorithm and to weaken it over the iterations.



Outline

Models
- Mixing model
- Source model
- Source separation problem

Reverberation modeling
- Room impulse response
- Early contributions
- Autoregressive model

Parameter estimation
- EM algorithm
- Hyper-parameters

Experiments
- Database and configuration
- Results
- Audio example

Conclusions

Database and configuration

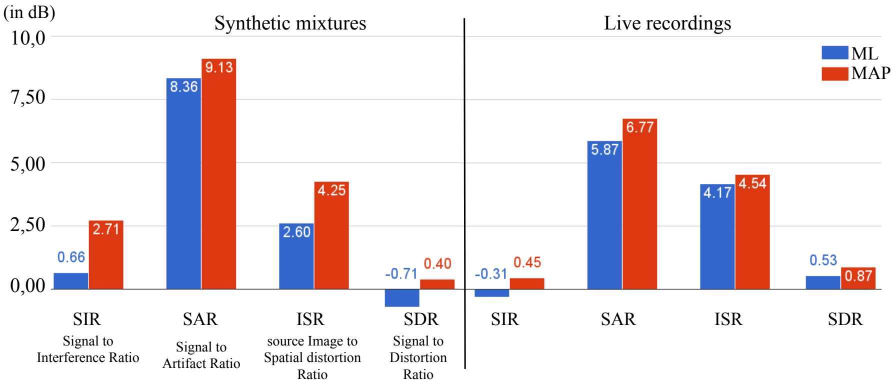
SiSEC: "Under-determined speech and music mixtures" task

- 8 stereo mixtures, each containing 3 musical sources
- Live recordings and synthetic mixtures ($T_{60} = 250$ ms)

Configuration

- Blind setting
- $\#\mathcal{K}_j = 4$ latent components for each source
- $R = 6$ early contributions

Results


(in dB)
Synthetic mixtures
Live recordings
ML
MAP
10,0
7,50
5,00
2,50
0,00
0.66
2.71
8.36
9.13
2.60
4.25
-0.71
0.40
-0.31
0.45
5.87
6.77
4.17
4.54
0.53
0.87
SIR
Signal to Interference Ratio
SAR
Signal to Artifact Ratio
ISR
source Image to Spatial distortion Ratio
SDR
Signal to Distortion Ratio
SIR
SAR
ISR
SDR

Audio example

Blind separation of 3 sources from a stereo mixture

source	piano		drums		voice	
method	ML	MAP	ML	MAP	ML	MAP
SDR	3.0	**8.6**	**0.9**	0.8	2.5	**4.1**
SIR	4.1	**10.9**	0.1	**0.4**	5.2	**13.1**
SAR	11.4	**13.0**	5.9	**9.5**	6.7	**11.0**
ISR	11.0	**18.0**	4.1	**8.5**	4.0	**5.0**

Sunrise by Shannon Hurley

Outline

Models
- Mixing model
- Source model
- Source separation problem

Reverberation modeling
- Room impulse response
- Early contributions
- Autoregressive model

Parameter estimation
- EM algorithm
- Hyper-parameters

Experiments
- Database and configuration
- Results
- Audio example

Conclusions

Conclusions

Conclusions

Prior distribution of the frequency response of mixing filters:

- Early contributions of room impulse responses
- AR model in frequency
- Better separation results than ML estimation

Future work

AR modeling in frequency $\Leftrightarrow$ time domain envelope modeling

- AR model for late reverberation
- Fixed AR parameters knowing some room characteristics

Thank you