Phase reconstruction of spectrograms based on a model of repeated audio events

Paul MAGRON, Roland BADEAU, Bertrand DAVID
Institut Mines-Télécom - Télécom ParisTech - CNRS LTCI - Paris, France

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

Introduction

Problem

Many audio source separation techniques act on spectrogram-like representations in the Time-Frequency (TF) domain.

- We need to reconstruct the phase to synthesize time-domain signals.
- Issues appear when sources overlap in the TF domain.
- Most approaches do not use a phase model of repeated audio events.

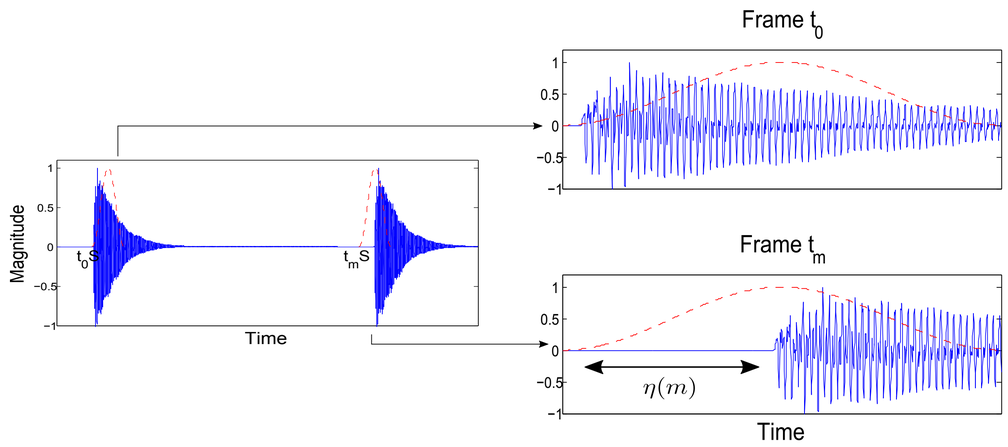
Our approach:

- Exploiting a phase model of repeated audio events.
- Defining a complex mixture model which includes this phase model.
- Application to source separation.

Phase model of repeated audio events

Main concept

- x: time-domain source that is activated M times.
- $X(f,t)$: STFT of x with hop size S, with $f \in [\![0; F-1]\!]$ and $t \in \mathbb{Z}$.
- The corresponding onset frames begin at $t_m S$, $m \in [\![0; M-1]\!]$.



Key idea: $X(f,t_m)$ and $X(f,t_0)$ are the Fourier transforms of two signals approximately equal up to a gain factor $\rho(m) \in \mathbb{R}_+$ and a delay of $\eta(m) \in \mathbb{R}$ samples:

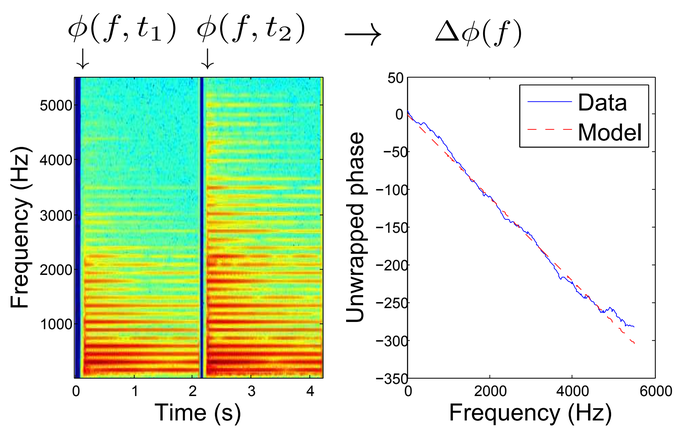
$$X(f,t_m) \approx X(f,t_0)\rho e^{i\lambda(m)f}, \text{ with } \lambda(m) = \frac{2\pi\eta(m)}{F}. \quad (1)$$

Phase model:

$$\underbrace{\angle X(f,t_m) = \phi(f,t_m)}_{\text{phase within an onset frame}} \approx \underbrace{\psi(f)}_{\text{reference phase}} + \underbrace{\lambda(m)f}_{\text{offset}}. \quad (2)$$

Example

Two occurrences of the same piano note with different nuances.
$\rightarrow$ Phase difference between onset frames.



$\rightarrow$ Accuracy of the phase model.

Mixture model

- $X \in \mathbb{C}^{F \times T}$ is a mixture of K sources X^k,
- $Y \in \mathbb{C}^{F \times M}$ is a submatrix of X corresponding to the onset frames:

$$Y(f,m) = X(f,t_m) = \sum_{k=1}^{K} A^k(f,t_m) e^{i\phi^k(f,t_m)}. \quad (3)$$

- Complex mixture with phase model:

$$\hat{Y}(f,m) = \sum_{k=1}^{K} A^k(f,t_m) e^{i\psi^k(f)} e^{i\lambda^k(m)f}. \quad (4)$$

Reduction of the dimensionality of the data: $K(F+M) < KFM$.

Parameters estimation

Minimization of a cost function:

- Strict:

$$\mathcal{C}_s = \sum_{f,m} \left| Y(f,m) - \sum_{k=1}^{K} A^k(f,t_m) e^{i\psi^k(f)} e^{i\lambda^k(m)f} \right|^2. \quad (5)$$

- Relaxed:

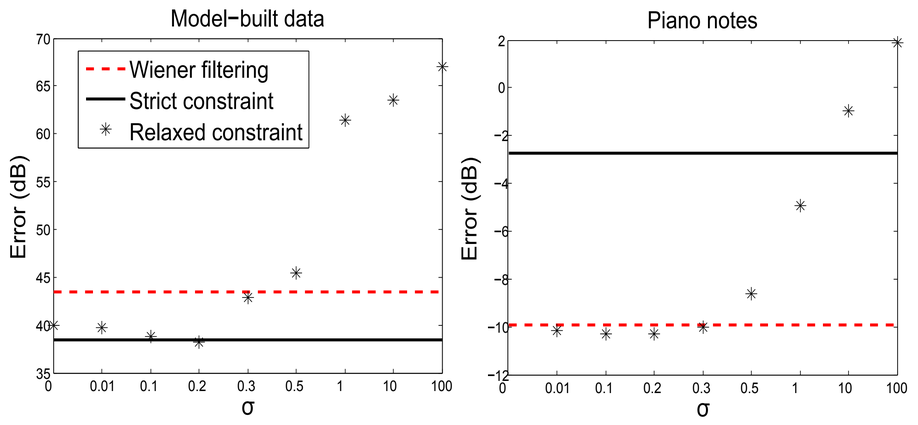
$$\mathcal{C}_r = \sum_{f,m} \left| Y(f,m) - \sum_{k=1}^{K} A^k(f,t_m) e^{i\phi^k(f,t_m)} \right|^2 + \sigma \sum_{f,m,k} A^k(f,t_m)^2 \left| e^{i\phi^k(f,t_m)} - e^{i\psi^k(f)} e^{i\lambda^k(m)f} \right|^2, \quad (6)$$

σ: prior weight which promotes the phase constraint.
$\partial_\psi \mathcal{C}$ (resp. $\partial_\phi \mathcal{C}$)$= 0 \rightarrow$ Estimation of ψ (resp. ϕ).
Adaptation of the ESPRIT algorithm $\rightarrow$ Estimation of λ.

Experimental results

Onset phase estimation

- Mixtures of two sources: model-built data and piano notes with different nuances. Magnitude values A^k are assumed to be known.
- Comparison with the traditional Wiener filtering approach.
- Estimation error: $\frac{1}{K}\sum_k ||Y^k - \hat{Y}^k||$.



Model-built data: the strict algorithm outperforms Wiener filtering.
Realistic data: slightly better phase estimation with the relaxed algorithm. Drawback: it requires that σ be carefully chosen.

Source separation

RePU (**Re**peating **P**hase & **U**nwrapping): onset phase estimation with the relaxed algorithm ($\sigma = 0.2$) and linear unwrapping over time [1].
Separation quality measured with the **S**ignal to **D**istortion, **I**nterference and **A**rtifact **R**atios (SDR, SIR and SAR).

Dataset	Method	SDR	SIR	SAR
Synthetic sinusoids (without TF overlap)	Wiener	**29.3**	**20.8**	**58.6**
	RePU	3.2	9.6	26.1
Synthetic sinusoids (with TF overlap)	Wiener	**10.5**	7.9	20.9
	RePU	3.2	**8.8**	**25.2**
MIDI audio excerpt (bass, guitar and piano)	Wiener	**−2.3**	−20.6	**20.8**
	RePU	−3.2	**−16.7**	11.6

- Without TF overlap: superiority of Wiener over RePU (expected).
- With TF overlap: RePU leads to an increase in interference and (sometimes) artifact rejection.

$\rightarrow$ **Potential of a phase model of repeated audio events for separating overlapping sources in the TF domain.**

Future research

- Refine the phase model: complex signals, different nuances.
- Phase correlations between frequency channels within an onset frame.
- Joint estimation of magnitude and phase: combine RePU and NMF models in a complex NMF framework.

References:

[1] P. Magron, R. Badeau and B. David, Phase reconstruction of spectrograms with linear unwrapping: application to audio signal restoration, EUSIPCO, 2015.