ISMIR 2007

Technische Universität Berlin

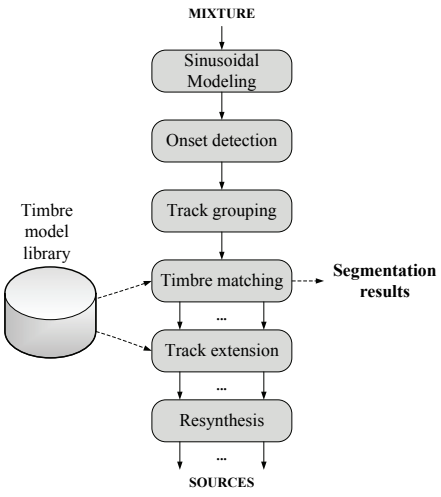
Juan José Burred, Thomas Sikora

Monaural Source Separation from Musical Mixtures based on Time-Frequency Timbre Models

Introduction

We present a system for source separation from monaural musical mixtures based on sinusoidal modeling and on a library of timbre models trained a priori. The models, which rely on Principal Component Analysis, serve as time-frequency probabilistic templates of the spectral envelope. They are used to match groups of sinusoidal tracks and assign them to a source, as well as to reconstruct overlapping partials. The proposed method does not make any assumptions on the harmonicity of the sources, and does not require a previous multipitch estimation stage. Since the timbre matching stage detects the instruments present on the mixture, the system can also be used for classification and segmentation.

1. System Overview



Pre-processing:

1. Sinusoidal modeling (spectral peak picking and partial tracking)

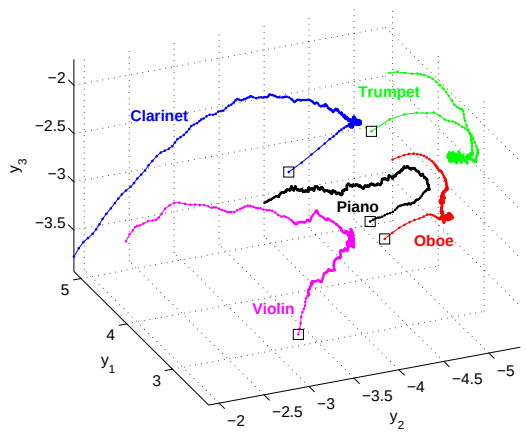
2. Onset detection (peaks of moving average of the number of new tracks)

3. Track grouping (common-onset)

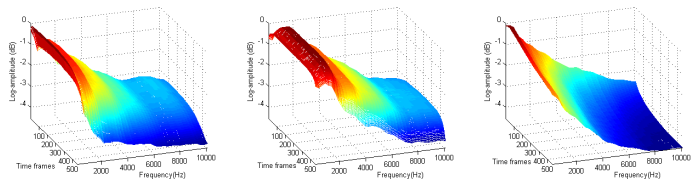
2. Timbre models

- Time-frequency templates describing the **spectral envelope** and its evolution in time. Based on performing **PCA** on a training database of spectral envelopes and modeling the projected coefficients as **Gaussian Processes**.
- Representation in PCA space (**prototype curves**) or projected back to the t-f domain (**prototype envelopes**).

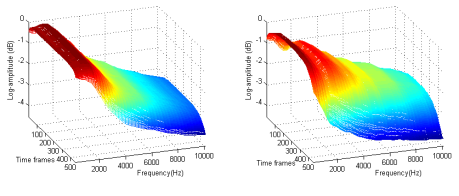
Prototype curves (mean trajectories)



Prototype envelopes (mean surfaces)



Clarinet *Oboe* *Piano*



Trumpet *Violin*

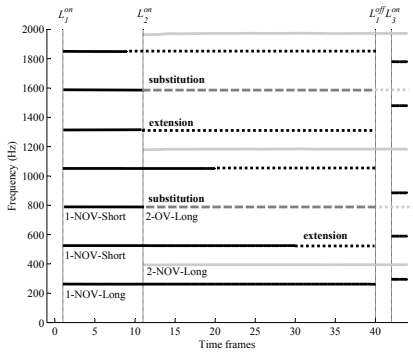
3. Timbre detection

- Matching of each common-onset track group $\mathbf{T}_o$ with each prototype envelope parametrized by $\boldsymbol{\theta}_i = (\mathbf{M}_i, \boldsymbol{\Sigma}_i)$
- Maximisation of the following likelihood:

$$L(\mathbf{T}_o|\boldsymbol{\theta}_i) = \max_{\alpha,N} \prod_{t,r} p\left(A^N_{tr} + \alpha|\mathbf{M}_i(f^N_{tr}), \boldsymbol{\Sigma}_i(f^N_{tr})\right) \quad (1)$$

α: amplitude scaling parameter.
A^N_{tr}, f^N_{tr}: amplitude and frequency values for a track belonging to a group that has been time-stretched so that its last frame is N.

4. Track extension and substitution



- The offsets are defined at the last frame of the longest track of group $\mathbf{T}_o$.
- **Non-overlapping tracks** are extended towards the offset, retrieving the missing amplitude from the corresponding timbre model.
- The amplitude of the **overlapping tracks** is retrieved from the model at their frequency support by linear interpolation.

5. Experiments and results

- Spectral Signal-to-Residual Ratio as global quality measure:

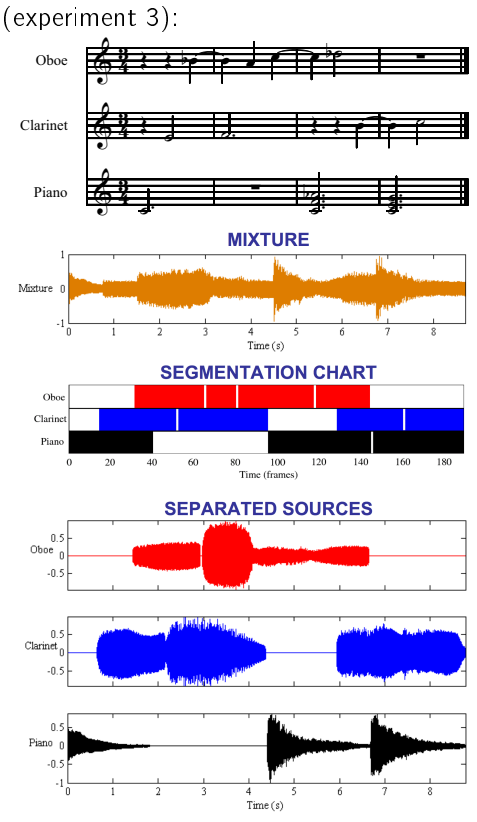
$$SSRR = 10\log \frac{\sum_{k,r} |S(k,r)|^2}{\sum_{k,r}(|S(k,r)| - |\hat{S}(k,r)|)^2} \quad (2)$$

Experiment 1: Intervals and arpeggios
2-note intervals and 3-note arpeggios of single notes, unknown instruments, 5 timbre models.

Experiment 2: Note sequences
Up to 4 consecutive notes, up to 3 simultaneous, known instruments.

Experiment 3: Sequences with chords
More complex sequences with same-instrument chords, up to 3 simultaneous, known instruments.

Example of separation including chords (experiment 3):



Polyphony	2	3
Exp. 1: Intervals / arpeggios	8.95 dB SSRR	5.38 dB SSRR
Exp. 2: Sequences	3.17 dB SSRR	2.26 dB SSRR

Acknowledgements

- Part of this research was performed at the Analysis/Synthesis team, IRCAM.
- The research work leading to this paper has been supported by the European Commission under the IST research network of excellence K-SPACE.