Associação Portuguesa
de
Engenharia de Áudio

Secção Portuguesa da
Audio Engineering Society

MUSIC SIGNAL ANALYSIS
USING
SPECTRAL CLUSTERING

9º Encontro da Secção Portuguesa de
Engenharia de Áudio

20 de Outubro de 2007

Leiria, Portugal

Luis Gustavo Martins
lmartins@inescporto.pt
PhD Student / Researcher
PhD Advisor Prof. Aníbal Ferreira (FEUP)

INESCPORTO
INSTITUTO DE ENGENHARIA DE SISTEMAS
E COMPUTADORES DO PORTO
LABORATÓRIO ASSOCIADO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

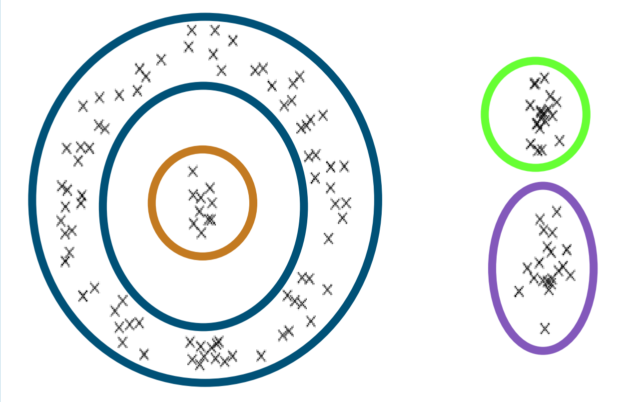
Notice

Presentation Outline

- Summary:
  - Spectral Clustering - Brief Introduction
  - Sound Source Segregation using Spectral Clustering
  - Application Examples:
    - Main Melody Detection
    - Voicing Detection
    - Timbre Recognition
    - Mono to Stereo Up-mixing
  - Conclusions

Spectral Clustering – A brief introduction (1)

- Spectral Clustering



→ How many clusters?

  - Alternative to the *EM* and *k-means* traditional algorithms:
    - Does not assume a convex shaped data representation
    - Does not assume Gaussian distribution of data
    - Does not present multiple minima in log-likelihood

      → Avoids multiple restarts of the iterative process

Spectral Clustering – A brief introduction (2)

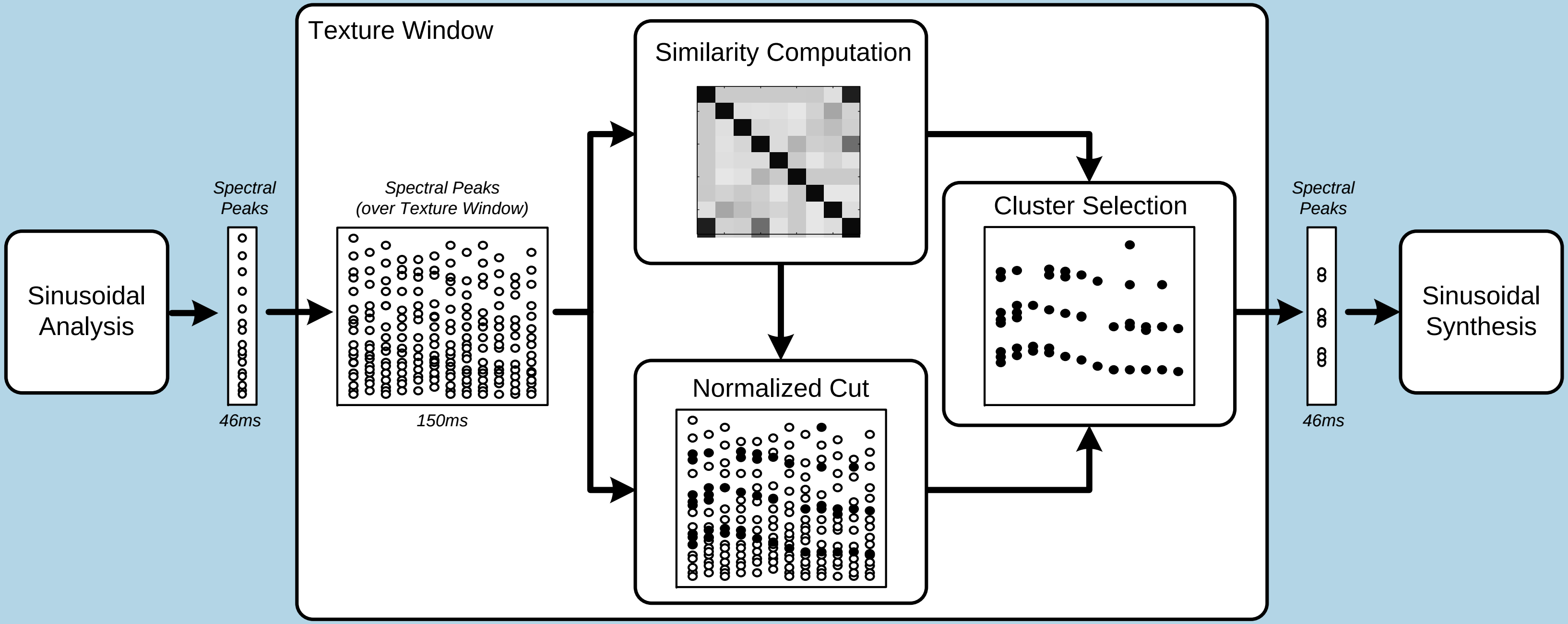
- Spectral Clustering
  - Relies on the *eigenstructure* of a *similarity matrix* to partition points into disjoint clusters
    - Points in the same cluster → high similarity
    - Points in different clusters → low similarity
  - **Normalized Cut**
    - Proposed in the area of ***Computer Vision*** [1]
    - Global criterion for segmenting graphs
    - Uses an *affinity* (i.e. *similarity*) *matrix*

      → encode topological knowledge about a problem

[1] J. Shi and J. Malik, "Normalized cuts and image segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, pp. 888-905, 2000.

Spectral Clustering → Sound Source Segregation (1)

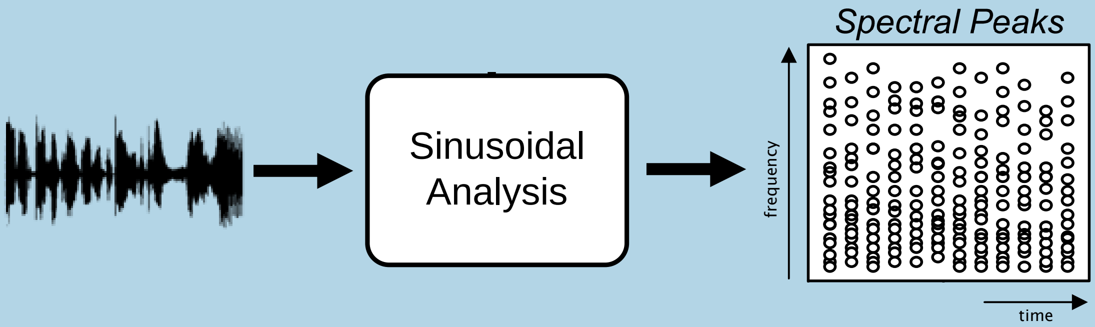
- Overall view



Spectral Clustering → Sound Source Segregation (2)

- Sinusoidal Modeling
  - Sum of most prominent sinusoids
    - Maximum of 20 sinusoids/frame
    - Window = 46ms ; hop = 11ms
    - Amplitude, Frequency, Phase

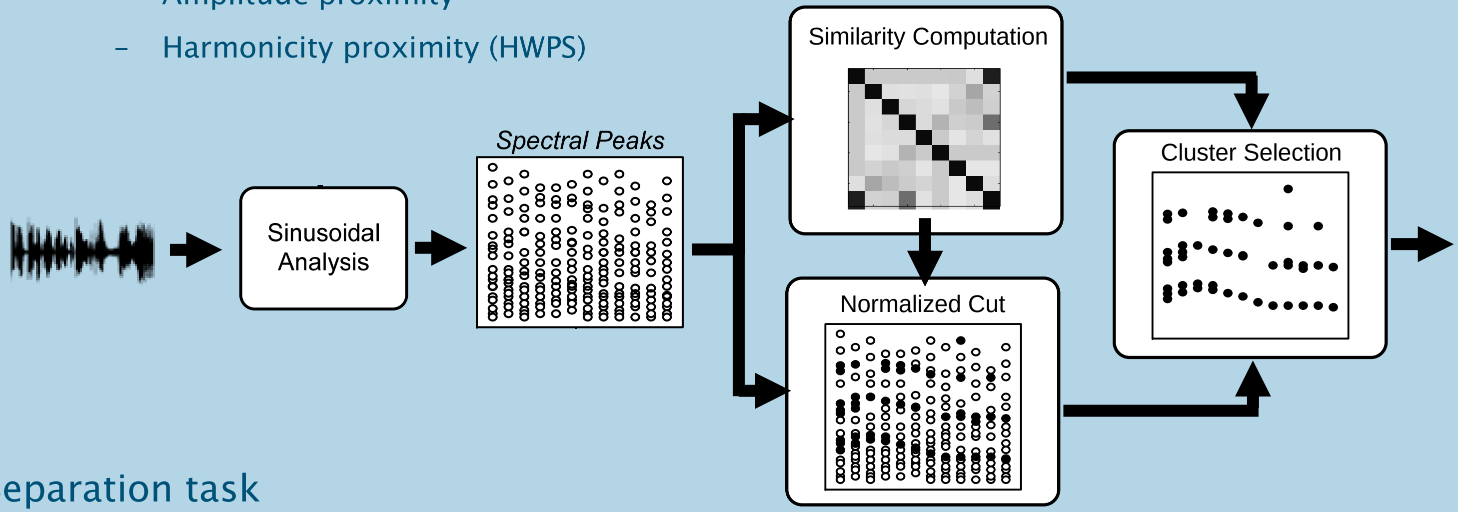
$$x_k(n) = \sum_{l=1}^{L_k} a_{lk} \cos\left(\frac{2\pi}{F_s} f_{lk} \cdot n + \phi_{lk}\right)$$



  - Construct a graph over a texture window of the sound mixture (e.g.150ms)
    - Provides time integration
      - Approaches partial tracking and source separation jointly, which have been traditionally two separated, consecutive stages

Spectral Clustering → Sound Source Segregation (3)

- Sound Source Segregation
  - Use of a flexible framework for representation of perceptual cues, from ASA [2]
    - expressed in terms of similarity between time-frequency components → *similarity space*
      - Frequency proximity
      - Amplitude proximity
      - Harmonicity proximity (HWPS)

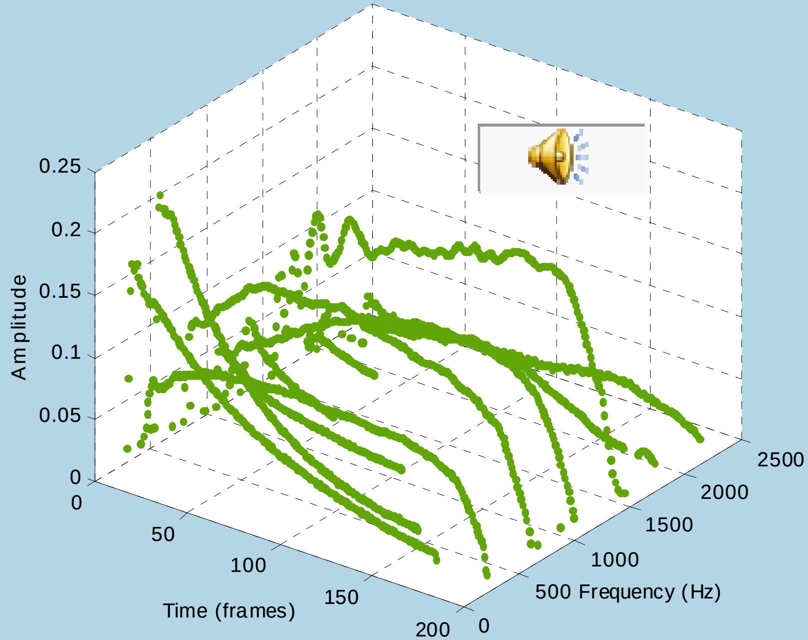


  - Separation task
    - Carried out by clustering components that are close in the similarity space
    - Use global ***Normalized Cut*** criterion
      - partition the graph into clusters (i.e. sources), using perceptual similarity cues

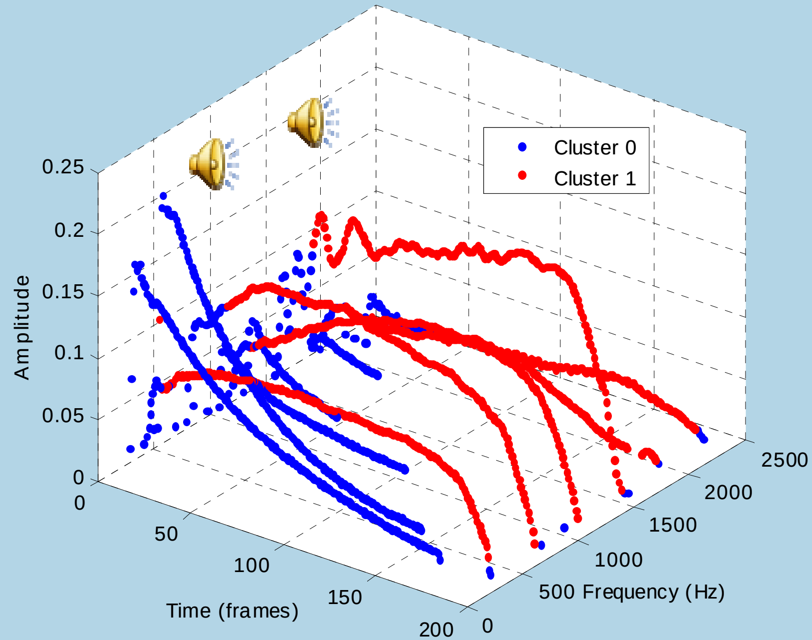
[2] A. Bregman, Auditory Scene Analysis – The Perceptual Organization of Sound: MIT Press, 1990.

Spectral Clustering → Sound Source Segregation (4)

Spectral Peaks



Clustered Spectral Peaks



Segregating the most prominent voice

→Jazz examples

→U2's Helter Skelter [live]

More real-world examples at: http://opihi.cs.uvic.ca/NormCutAudio/index.php?page=data

Spectral Clustering → Sound Source Segregation (5)

- Want to give it a try? ☺

MARSYAS

MUSIC ANALYSIS, RETRIEVAL AND SYNTHESIS FOR AUDIO SIGNALS

http://marsyas.sourceforge.net

```
> peakClustering myAudio.wav
```

[3] M. Lagrange, L. G. Martins, J. Murdoch, and G. Tzanetakis, "Normalized Cuts for Predominant Melodic Source Separation," IEEE Transactions on Audio, Speech, and Language Processing (in press), 2007.

Main Melody Detection

Spectral Clustering → Main Melody Detection (1)

- Main melody detection in **real-world polyphonic music signals**:
  - **Melody is one of the key musical descriptors of a song**
    - Monophonic pitch estimation techniques perform poorly on polyphonic signals
      - Too complex spectra from simultaneously sounding sources (too much spectral overlapping occurs)
    - Common approach for main melody estimation

      → Start with multipitch extraction followed by predominant pitch estimation [3, 4]

    - **Spectral Clustering** allows segregating the most prominent clusters over time

      → Resynthesize the **segregated main voice clusters**

      → *(Even nicer: estimate pitch of each cluster directly in feature domain → future work)*

      → Easier to perform pitch estimation using well known monophonic pitch estimation techniques

[3] R. P. Paiva, T. Mendes, and A. Cardoso, "Melody detection in polyphonic musical signals: Exploiting perceptual rules, note salience, and melodic smoothness," Computer Music Journal, vol. 30, pp. 80-98, Win 2006.

[4] A. Klapuri and M. Davy, "Signal Processing Methods for Music Transcription," Springer-Verlag, 2006.

Spectral Clustering → Main Melody Detection (2)

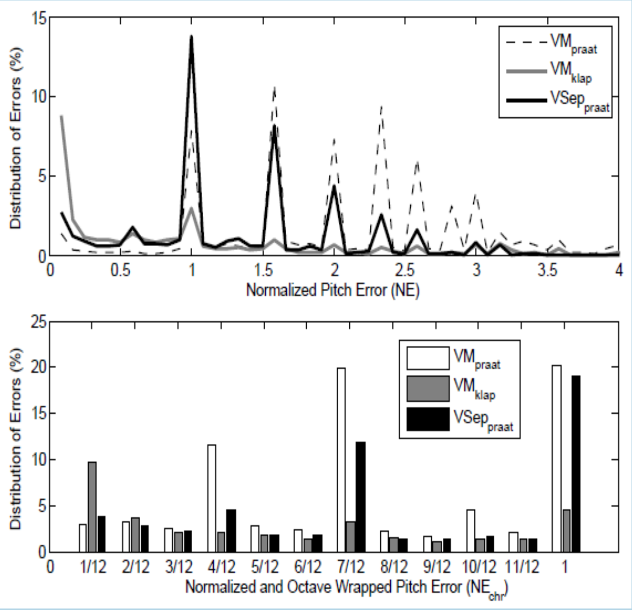
- Some experimental results [3]:
  - ***MIREX 2005*** *automatic melody extraction evaluation exchange* dataset
    - Included the pitch contour ground-truth for each song
    - http://www.music-ir.org/mirex2005/index.php/Main_Page
  - Dataset of **10 real-world polyphonic music recordings**
    - Availability of the original isolated tracks

      → Allowed to generate ground-truth and perform evaluations
    - http://opihi.cs.uvic.ca/NormCutAudio/index.php?page=data
  - Comparison with two techniques:
    - Monophonic pitch estimation (from *Praat*)
    - State-of-the-Art multipitch and main melody estimation algorithm [5]

[3] M. Lagrange, L. G. Martins, J. Murdoch, and G. Tzanetakis, "Normalized Cuts for Predominant Melodic Source Separation," IEEE Transactions on Audio, Speech, and Language Processing (in press), 2007.

[5] A. Klapuri, "Multiple fundamental frequency estimation by summing harmonic amplitudes," in International Conference on Music Information Retrieval (ISMIR) Victoria, BC, Canada, 2006.

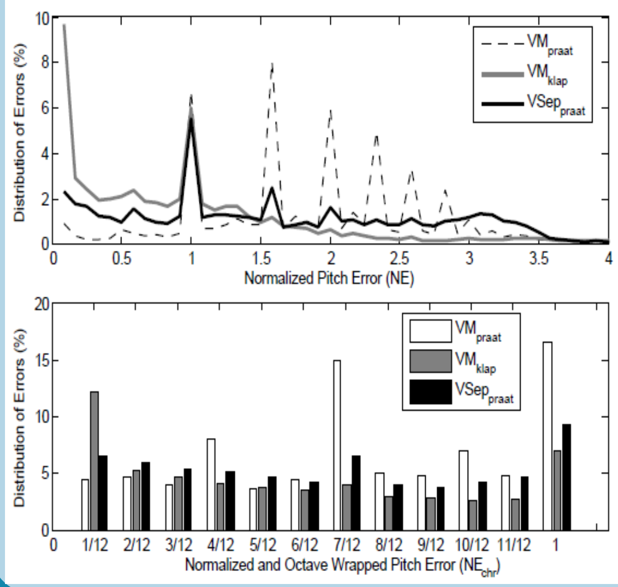
Spectral Clustering → Main Melody Detection (3)



Results on the **MIREX 2005 dataset**

NORMALIZED PITCH ERRORS AND GROSS ERRORS FOR MIREX DATASET

	NE	NE_{chr}	$GE(\%)$	$GE-8^{ve}(\%)$
VM_{praat}	3.29	0.48	76.02	55.87
$VSep_{praat}$	1.34	0.36	54.12	34.97
VM_{klap}	0.34	0.15	34.27	29.77



Results on the **10 real-world recordings dataset**

NORMALIZED PITCH ERRORS AND GROSS ERRORS ACROSS CORPUS

	NE	NE_{chr}	$GE(\%)$	$GE-8^{ve}(\%)$
VM_{praat}	8.62	0.51	82.44	66.00
$VSep_{praat}$	3.89	0.35	64.45	55.23
VM_{klap}	0.55	0.26	55.70	48.68

Voicing Detection

Spectral Clustering → Voicing Detection (1)

- Identifying **where the melody pitches occur in a song**
  - Evaluation performed on the same 10 real-world songs dataset
    - http://opihi.cs.uvic.ca/NormCutAudio/index.php?page=data
    - Ground truth was created manually from the isolated melody tracks
  - Evaluated three feature sets:
    - **MFCC** features extracted from the **mixed signal** of each song
    - **MFCC** features extracted from the **segregated main voice signal** using Spectral Clustering
    - **Cluster Peak Ratio (CPR)** feature [3] extracted from the segregated main voice clusters using Spectral Clustering

$$CPR = \frac{\max(A^k)}{\mathrm{mean}(A^k)}$$

[3] M. Lagrange, L. G. Martins, J. Murdoch, and G. Tzanetakis, "Normalized Cuts for Predominant Melodic Source Separation," IEEE Transactions on Audio, Speech, and Language Processing (in press), 2007.

Spectral Clustering → Voicing Detection (1)

- Machine Learning framework
  - Training of two classifiers on three feature sets:
    - *ZeroR → baseline (i.e. random classifier)*
    - *Naive Bayes* classifier (NB)
    - *Support Vector Machine* (SVM)
  - Results [3]:

VOICING DETECTION PERCENTAGE ACCURACY

	ZeroR	*NB*	*SVM*
VM_{MFCC}	55	69	69
$VSep_{MFCC}$	55	77	86
$VSep_{CPR}$	55	73	74

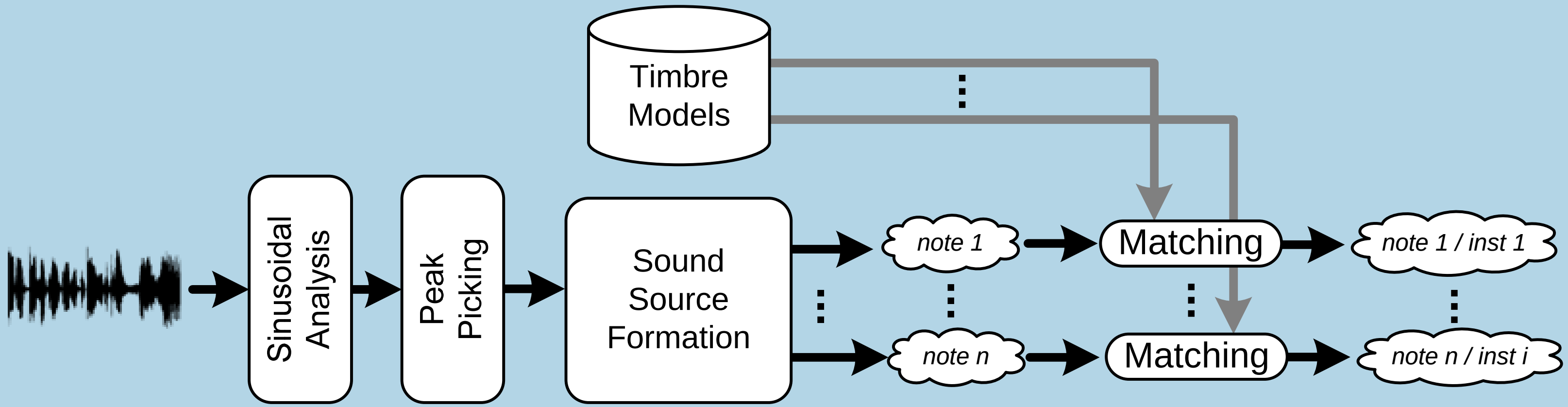
[3] M. Lagrange, L. G. Martins, J. Murdoch, and G. Tzanetakis, "Normalized Cuts for Predominant Melodic Source Separation," IEEE Transactions on Audio, Speech, and Language Processing (in press), 2007.

Application Example

Timbre Recognition

Spectral Clustering → Timbre Recognition (1)

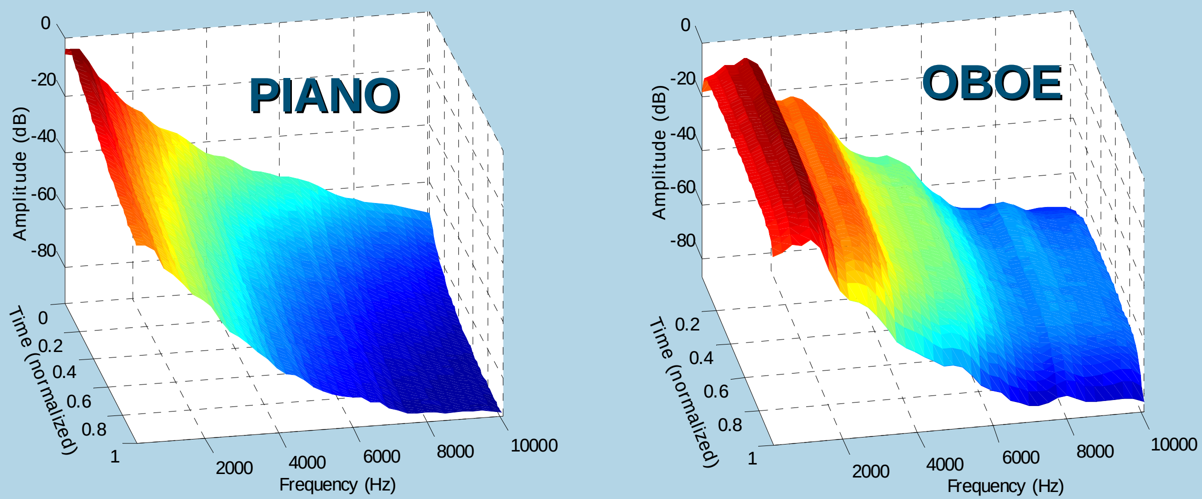
- Framework for timbre classification



  - polyphonic, multi-instrumental audio signals
    - Artificial mixtures of 2-, 3- and 4-notes from real instruments
  - Automatic separation of the sound sources
    - Sound sources and events are reasonably captured, corresponding in most cases to played notes
  - Matching of the separated events to a collection of 6 timbre models

Spectral Clustering → Timbre Recognition (2)

- 6 instruments modeled [10]:
  - *Piano*, *violin*, *oboe*, *clarinet*, *trumpet* and *alto sax*
  - Modeled as a set of time-frequency templates



- Describe the typical evolution in time of the spectral envelope of a note
  - Matches the salient peaks of the spectrum

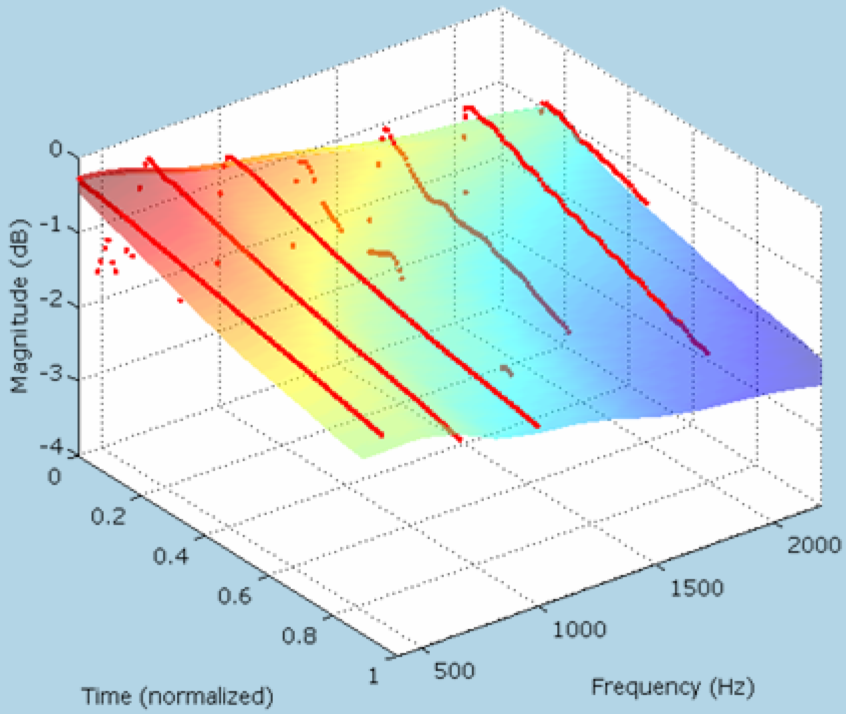
[10] J. J. Burred, A. Röbel, and X. Rodet, "An Accurate Timbre Model for Musical Instruments and its Application to Classification," in *First Workshop on Learning the Semantics of Audio Signals*, Athens, Greece, 2006.

Spectral Clustering → Timbre Recognition (3)

- Matching Examples

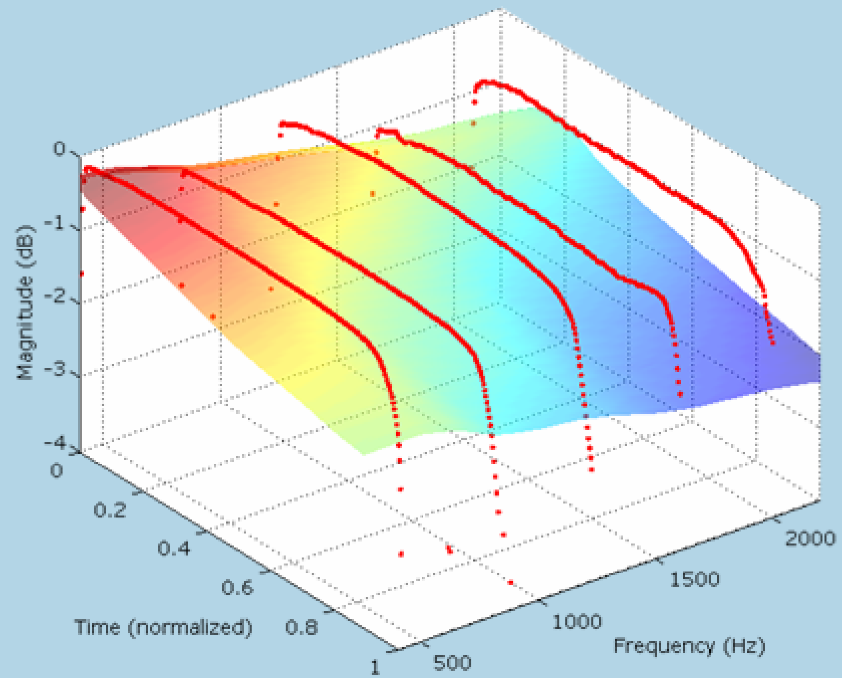
Strong Matching

Piano cluster ←→ piano prototype



Weak Matching

Alto sax cluster ←→ piano prototype



[3] L. G. Martins, J. J. Burred, G. Tzanetakis, and M. Lagrange, "Polyphonic Instrument Recognition using Spectral Clustering," in 8th International Conference on Music Information Retrieval (ISMIR 2007) Vienna, Austria, 2007.

Spectral Clustering → Timbre Recognition (4)

- Instrument presence detection in mixtures of notes
  - 54 different combinations of instruments and notes
    - 2-, 3- and 4-note mixtures
      - 18 audio files x 3 = 54 audio examples in the dataset
  - **56%** of instruments occurrences correctly detected, with a precision of **64%**
    - Oboe and alto sax as a good examples of good detections
    - Piano as the most difficult instrument (mainly in 4-note mixtures)

Performance degrades...

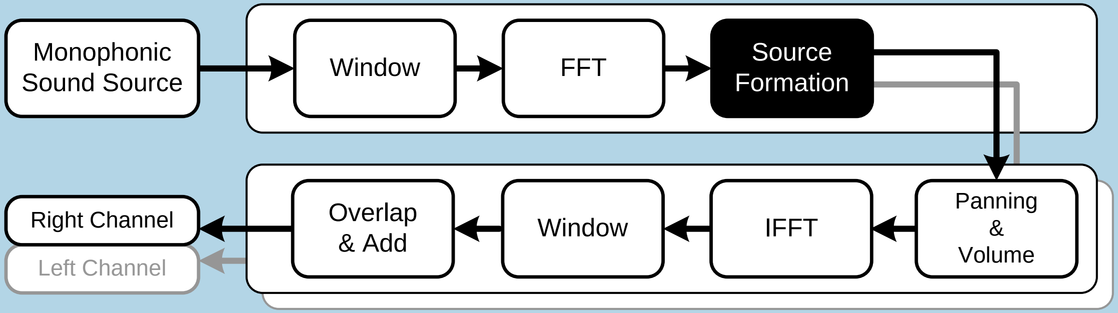
	2-note			3-note			4-note			total		
	RCL	**PRC**	**F1**	**RCL**	**PRC**	**F1**	**RCL**	**PRC**	**F1**	**RCL**	**PRC**	**F1**
p	83	100	91	22	100	36	0	0	0	23	100	38
o	100	75	86	100	46	63	67	40	50	86	50	63
c	33	100	50	33	100	50	40	86	55	36	93	52
t	89	100	94	58	100	74	58	64	61	67	85	75
v	67	67	67	83	45	59	83	36	50	80	43	56
s	100	43	60	67	60	63	60	75	67	67	62	64
total	**75**	**79**	**77**	**56**	**64**	**59**	**46**	**56**	**50**	**56**	**64**	**60**

Semi-automatic Mono to Stereo Up-mixing

Spectral Clustering → Mono to Stereo Up-mixing (1)

- Convert monophonic recordings to stereo
  - Spectral Clustering for Sound Source Formation
    - build a middle level representation of the sound using a perceptually motivated clustering of spectral components
  - include spatial panning information when converting from mono to stereo
    - allows the user to define panning information for major sound sources

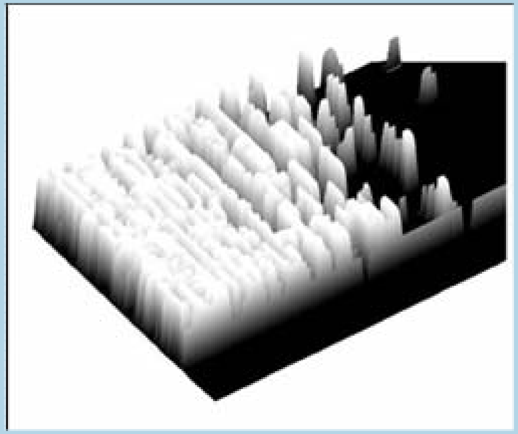
      → enables enhancing the stereophonic immersion quality of the resulting sound



Spectral Clustering → Mono to Stereo Up-mixing (2)

- FFT Resynthesis
  - A Fourier based approach is considered
    - A mask is assigned to each peak
    - The amplitude of each frequency bin is weighted accordingly:

$$
\begin{aligned}
m_l(k,t) &= g \cdot (v \cdot (1-p)) + (1-g)m_l(k,t-1) \\
m_r(k,t) &= g \cdot (v \cdot (1+p)) + (1-g)m_r(k,t-1)
\end{aligned}
$$



A piano source spectral mask

  - Spectral components of each source may be panned to different azimuths

DEMO [6]

[6] M. Lagrange, L. G. Martins, and G. Tzanetakis, "Semi-Automatic Mono to Stereo Up-mixing using Sound Source Formation," in 122nd Convention of the Audio Engineering Society, Vienna, Austria, 2007.

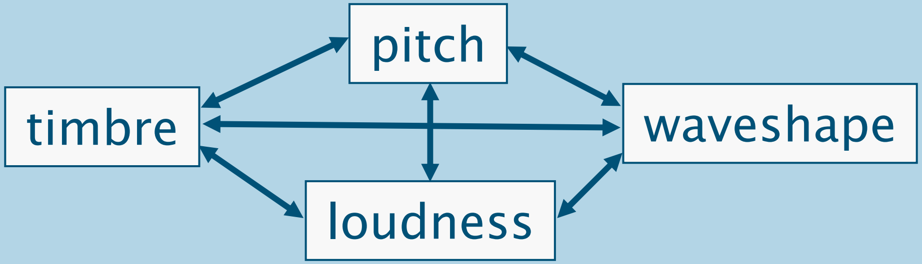
Conclusions

Discussion (1)

- Proposal of a framework for sound source segregation
  - Based on a Spectral Clustering technique
  - Approaches partial tracking and source separation jointly, using a flexible framework for including new perceptually motivated auditory cues
  - does not require any a priori information about pitch of sources
  - Shows good potential for applications in:
    - source segregation/separation,
    - monophonic or polyphonic instrument classification,
    - Main melody estimation
    - pre-processing for polyphonic transcription, ...
  - *Sources* VS *Events*
    - Weak matching of separated clusters to actual sources...
      - What are we segregating? Original Sources or sound events?

Discussion (2)

- Future work:
  - Inclusion of new perceptually motivated auditory cues
    - Time and frequency masking
    - Stereo placement of spectral components [7]
    - Timbre models as a priori information
  - Analysis of time events as side information for Sound Source Formation
    - Prior time segmentation of music notes/events
      → Automatically define the duration of the analysis texture window
  - Extraction of new descriptors directly from segregated cluster parameters:
    - Pitch, spectral features, frequency tracks, timing information
  - Models of attention of the human auditory system when performing auditory scene analysis



[7] G. Tzanetakis, L. G. Martins, "Stereo Panning Information for Music Information Retrieval Tasks", submitted to the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, USA.

Acknowledgments

- This work is the result of the collaboration with:
  - *University of Victoria, BC, Canada*
    - George Tzanetakis
    - Mathieu Lagrange (now with the McGill Music Technology Group, Canada)
    - Jennifer Murdock
    - All the Marsyas team
  - *Technical University of Berlin, Germany*
    - Juan José Burred (now with the IRCAM, Paris, France)
  - *INESC Porto, Portugal*
    - Luis Filipe Teixeira
    - Jaime Cardoso
    - Fabien Gouyon
- Supporting entities
  - *Fundação para a Ciência e Tecnologia - FCT*
  - *Fundação Calouste Gulbenkian*
  - *VISNET II, NoE European Project*

THANK YOU!

Questions?

lmartins@inescporto.pt