Supervised Musical Source Separation from Mono and Stereo Mixtures based on Sinusoidal Modeling

Juan José Burred
Équipe Analyse/Synthèse, IRCAM
burred@ircam.fr

Communication Systems Group
Technische Universität Berlin
Prof. Dr.-Ing. Thomas Sikora

Presentation overview

- Motivations, goals
- Timbre modeling of musical instruments
  - Representation stage
  - Prototyping stage
  - Application to instrument classification
- Monaural separation
  - Track grouping
  - Timbre matching
  - Application to polyphonic instrument recognition
  - Track retrieval
  - Evaluation and examples of mono separation
- Stereo separation
  - Blind Source Separation (BSS) stage
  - Extraneous track detection
  - Evaluation and examples of stereo separation
- Conclusions and outlook

Motivation

- Source Separation for Music Information Retrieval
  - Goal: Facilitate feature extraction of complex signals
- The paradigms of Musical Source Separation (based on [Scheirer00])
  - *Understanding without separation*
    - Multipitch estimation, music genre classification
    - "Glass ceiling" of traditional methods (MFCC, GMM) [Aucouturier&Pachet04]
  - *Separation for understanding*
    - First (partially) separate, then feature extraction
    - Source separation as a way to break the glass ceiling!
  - *Separation without understanding*
    - BSS: Blind Source Separation (ICA, ISA, NMF)
  - *Understanding for separation*
    - Supervised source separation

[Scheirer00] E. D. Scheirer. *Music-Listening Systems*. PhD thesis, Massachusetts Institute of Technology, 2000.

[Aucouturier&Pachet04] J.-J. Aucouturier and F. Pachet. Improving Timbre Similarity: How High is the Sky? *Journal of Negative Results in Speech and Audio Sciences*, 1 (1), 2004.

Musical Source Separation Tasks

- Classification according to the nature of the mixtures:

Difficulty: - → +

Source position	Mixing process	Source/mixture ratio	Noise	Musical texture	Harmony
• changing • **static**	• echoic (changing impulse response) • echoic (static impulse response) • delayed • **instantaneous**	• **underdetermined** • overdetermined • even-determined	• noisy • **noiseless**	• monodic (multiple voices) • heterophonic • **homophonic / homorhythmic** • **polyphonic / contrapuntal** • monodic (single voice)	• **tonal** • **atonal**

- Classification according to available a priori information:

A priori knowledge: + → -

Difficulty: - → +

Source position	Source model	Number of sources	Type of sources	Onset times	Pitch knowledge
• **unknown** • statistical model • known mixing matrix	• none • statistical independence • **sparsity** • **advanced/trained source models**	• **unknown** • **known**	• **unknown** • **known**	• **unknown** • known (score/MIDI available)	• **none** • pitch ranges • score/MIDI available

Modeling of Timbre

- Based on the Spectral Envelope and its dynamic evolution
- Requirements on the model
  - Generality
    - Ability to handle unknown, realistic signals.
    - Implemented by statistical learning from sample database.
  - Compactness
    - Together with generality, implies that the model has captured the essential source characteristics.
    - Implemented with spectral basis decomposition via Principal Component Analysis (PCA).
  - Accuracy
    - The model must guide the grouping and unmixing of the partials.
    - Demanding requirement that is not always necessary in other MIR application.
    - Realized by estimating the spectral envelope by Sinusoidal Modeling + Spectral Interpolation.
- Details on design and evaluation: [Burred 06]

[Burred06] J.J. Burred, A. Röbel and X. Rodet. An Accurate Timbre Model for Musical Instruments and its Application to Classification. In *Proc. Workshop on Learning the Semantics of Audio Signals (LSAS)*, Athens, Greece, December 2006.

Representation stage (I)

- Basis decomposition of partial spectra

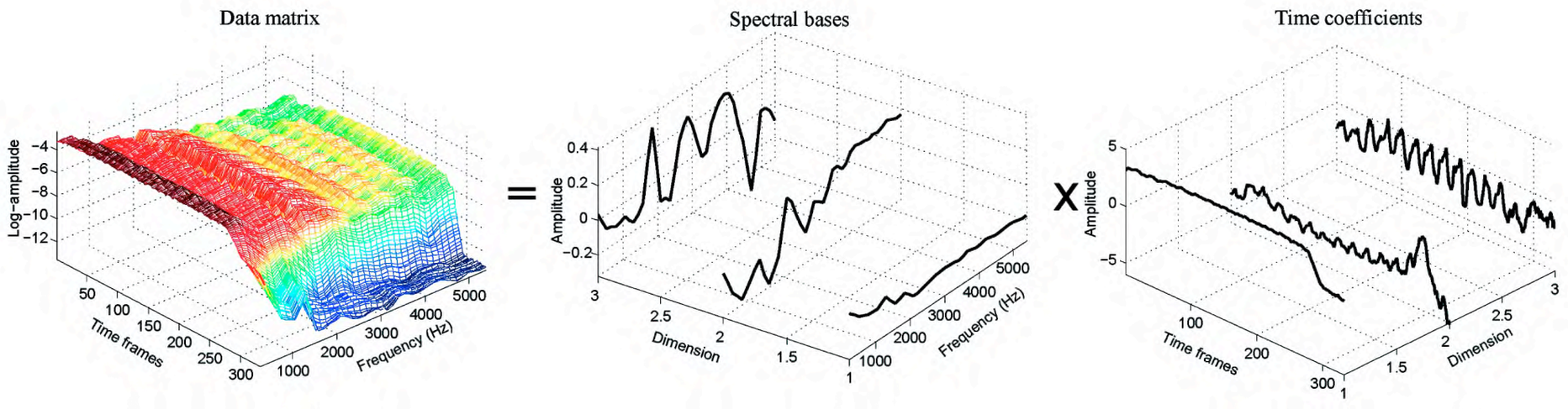
$$\mathbf{X} = \mathbf{PY}$$

$\mathbf{X}$ Data matrix (partial amplitudes)

$\mathbf{P}$ Transformation basis

$\mathbf{Y}$ Projected coefficients

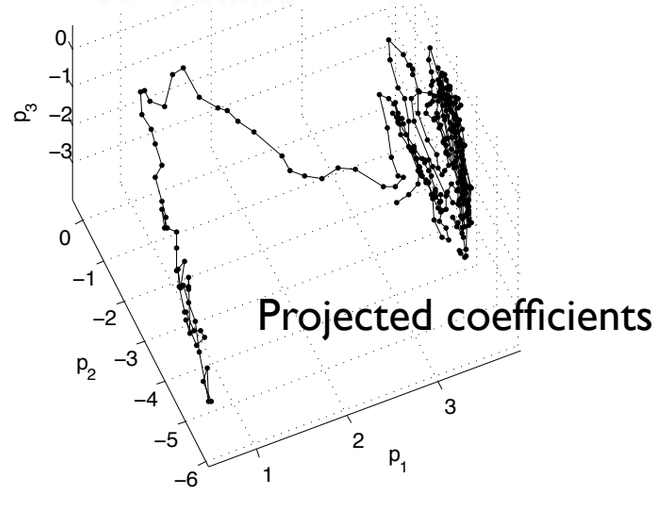
- Application of PCA to spectral envelopes
  - Example: decomposition of a single violin note, with vibrato



$$\mathbf{Y}_\rho = \mathbf{\Lambda}_\rho^{-1/2}\mathbf{P}_\rho^T(\mathbf{X} - E\{\mathbf{X}\})$$

$$\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_D)$$

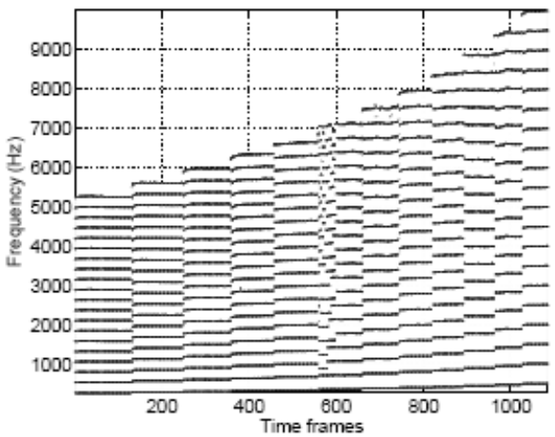
The λ_d are the D largest eigenvalues of the covariance matrix $\mathbf{\Sigma_x}$, whose corresponding eigenvectors are the columns of $\mathbf{P}_\rho$.



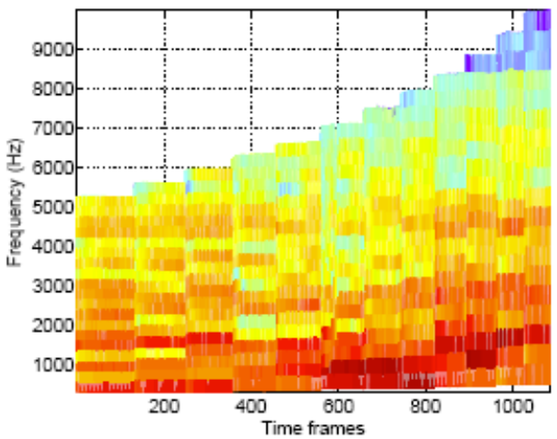
Projected coefficients

Representation stage (2)

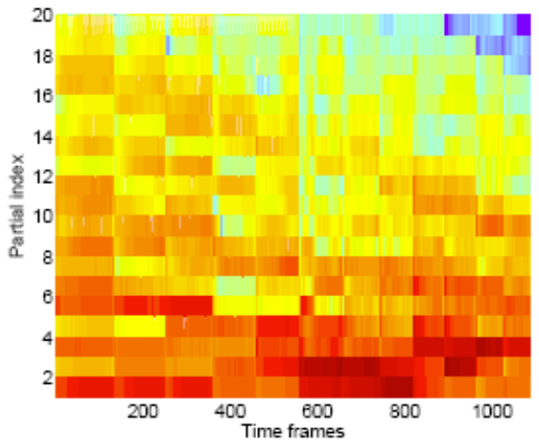
- Arrangement of the data matrix
  - Partial Indexing



Frequency support

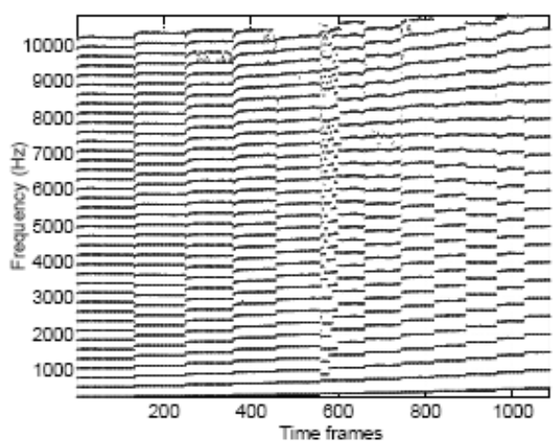


Original partial data

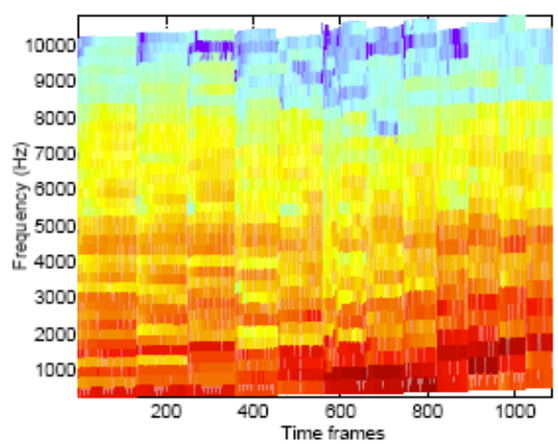


PCA data matrix

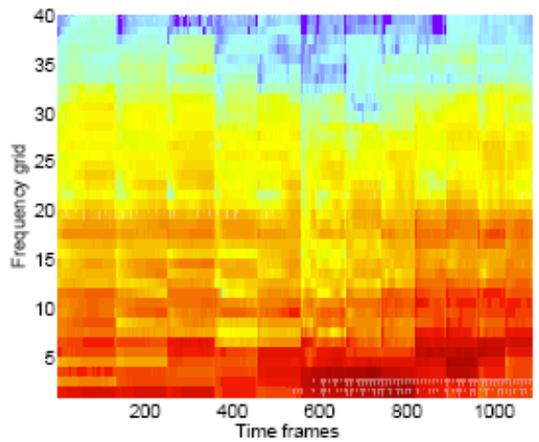
  - Envelope Interpolation (preserves formants)



Frequency support



Original partial data



PCA data matrix

  - Envelope Interpolation performs better according to all criteria (compactness, accuracy, generality) and in classification tasks.

Prototyping stage (I)

- For each instrument, each coefficient trajectory is interpolated to the same relative time positions.
- Each cloud of "synchronous" coefficients is modeled as a D-dimensional Gaussian distribution.
- This originates a prototype curve $\mathcal{C}_i$ that can be modeled as a D-dimensional, non-stationary Gaussian Process with time-varying means and covariances.

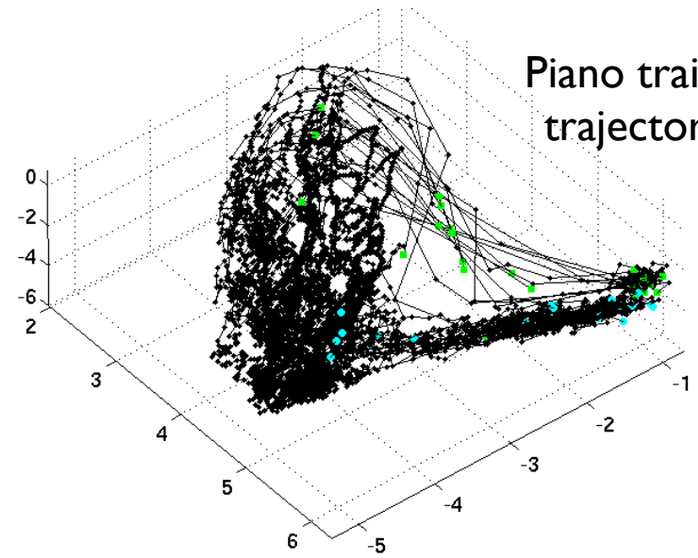
$$\mathcal{C}_i \sim GP\left(\boldsymbol{\mu}_i(r), \boldsymbol{\Sigma}_i(r)\right)$$

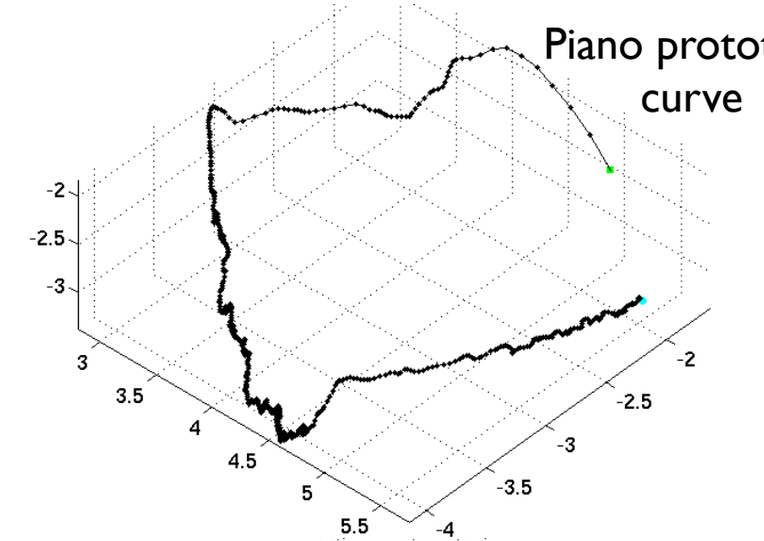
- Projected back to time-frequency, the equivalent is a prototype envelope $\mathcal{E}_i$: a unidimensional GP with time- and frequency-variant mean and variance surfaces.

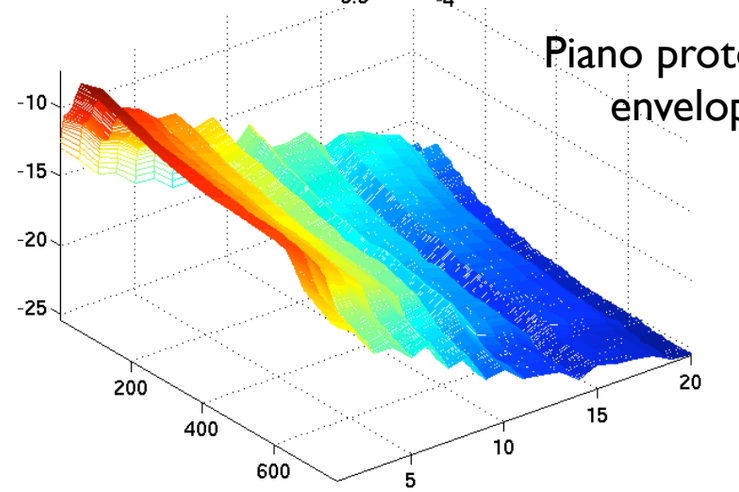
$$\hat{\boldsymbol{\mu}}_{ir} = \mathbf{P}_\rho \boldsymbol{\Lambda}_\rho^{1/2} \boldsymbol{\mu}_{ir} + E\{\mathbf{X}\}$$

$$\hat{\boldsymbol{\sigma}}_{ir}^2 = \mathrm{diag}\left(\mathbf{P}_\rho \boldsymbol{\Lambda}_\rho^{1/2} \boldsymbol{\Sigma}_{ir} (\mathbf{P}_\rho \boldsymbol{\Lambda}_\rho^{1/2})^T\right)$$

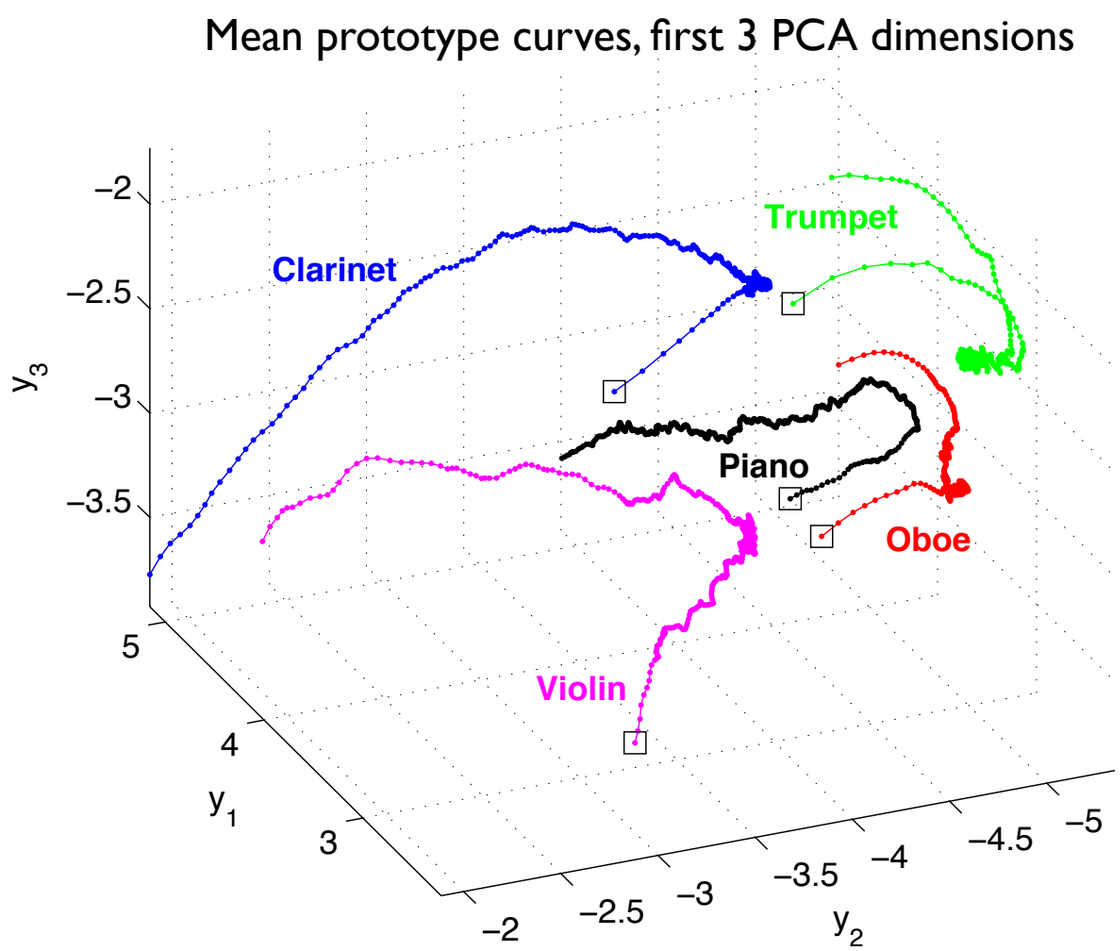
$$\mathcal{E}_i \sim GP\left(\mu_i(t,f), \sigma_i^2(t,f)\right)$$







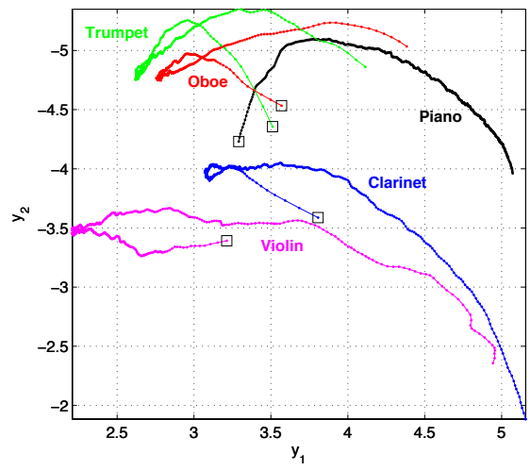
Prototyping stage (2)



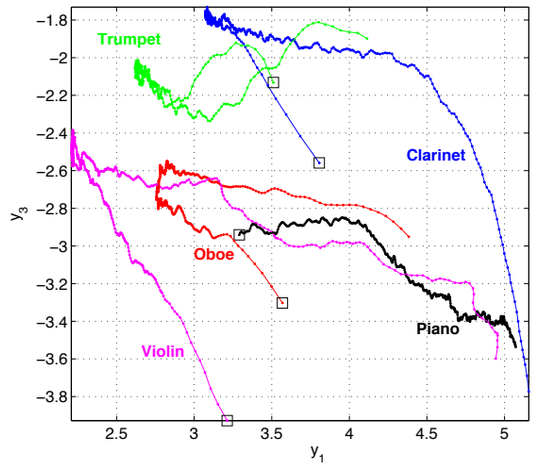
- Practical example
  - 5 instruments: piano, clarinet, trumpet, oboe, violin
  - 423 sound samples, 2 octaves
  - All dynamic levels (forte, mezzoforte, piano)
  - RWC database
  - Common PCA bases
  - Only mean curves represented

- Automatically generated timbre space

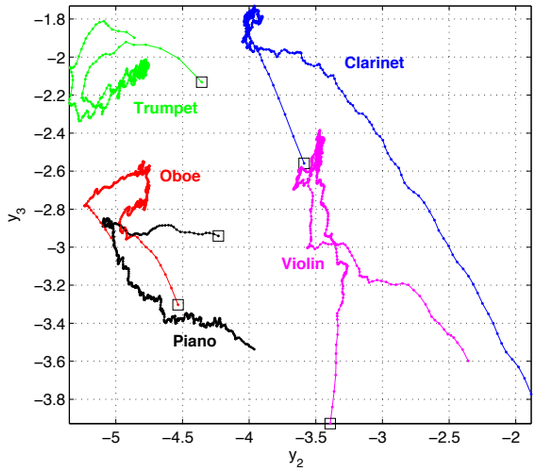
y1,y2 projection



y1,y3 projection

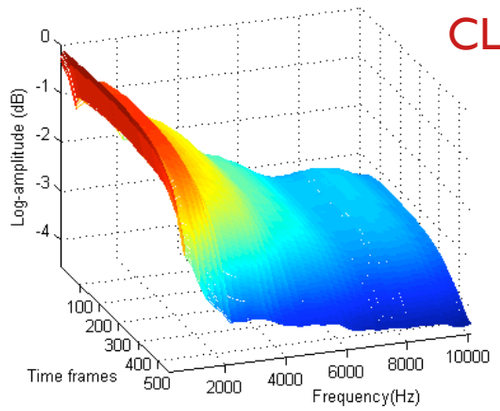


y2,y3 projection

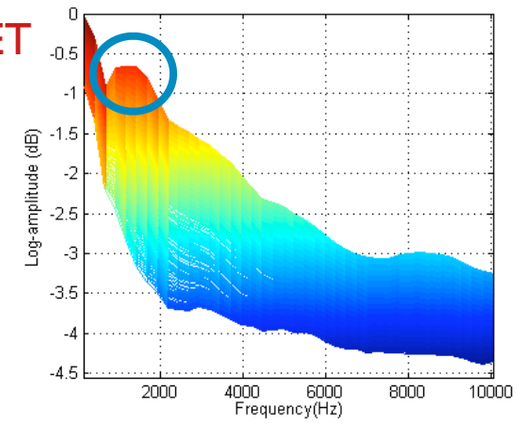


Prototyping stage (3)

CLARINET

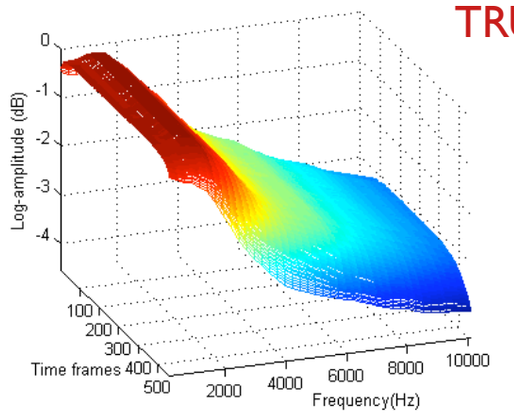


Prototype envelope

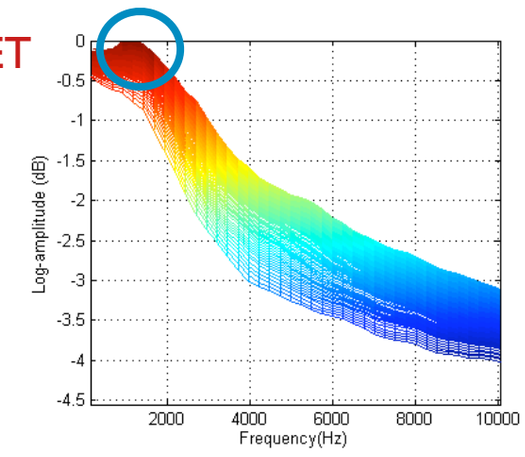


Frequency profile

TRUMPET

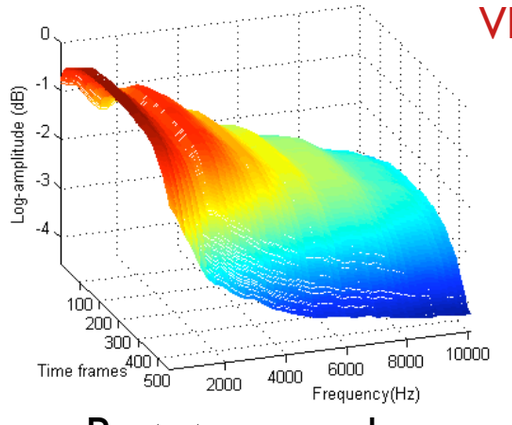


Prototype envelope

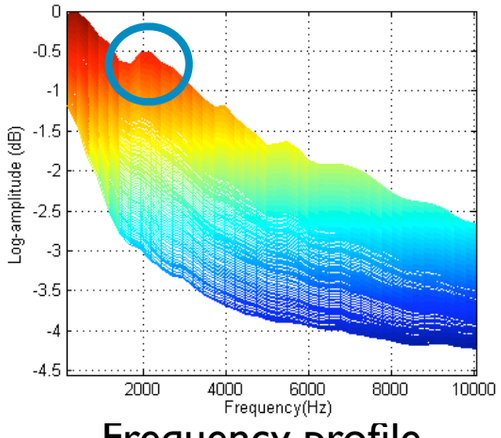


Frequency profile

VIOLIN



Prototype envelope



Frequency profile

- Practical example (cont'd)
  - Projection back into time-frequency domain.
  - The prototype envelopes will serve as templates for the grouping and separation of partials.
- Examples of observed formants:
  - Clarinet:
    first formant, between 1500 Hz and 1700 Hz. [Backus77]
  - Trumpet:
    first formant, between 1200 Hz and 1400 Hz. [Backus77]
  - Violin:
    "bridge hill" around 2000 Hz. [Fletcher98]

[Backus77] J. Backus. *The Acoustical Foundations of Music*. W.W. Norton, 1977.

[Fletcher98] N. H. Fletcher and T. D. Rossing. *The Physics of Musical Instruments*. Springer, 1998.

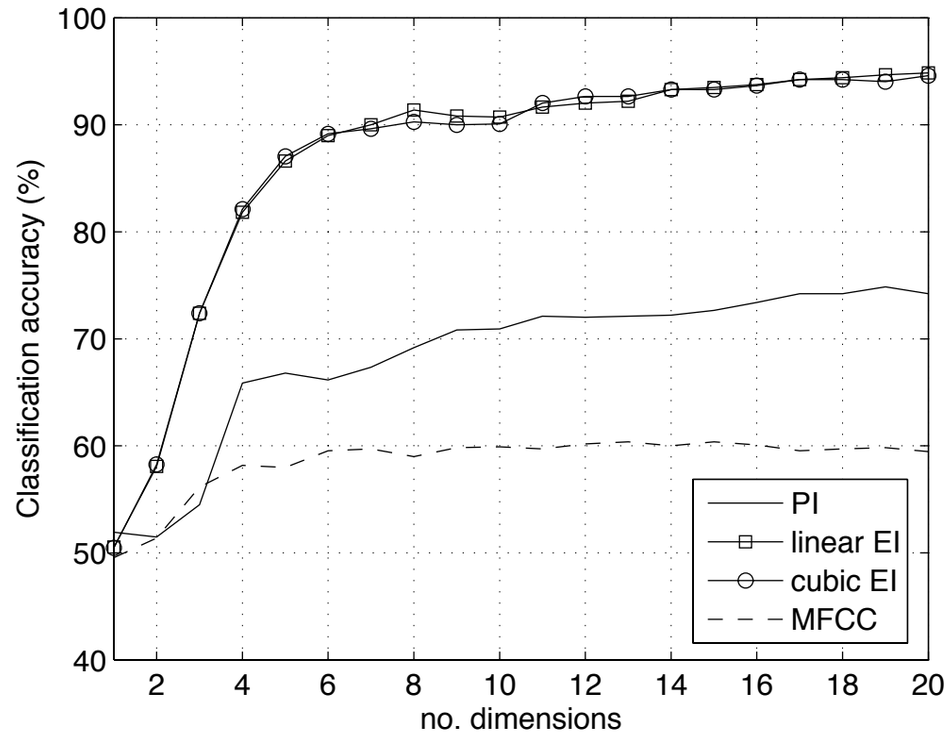
Application to instrument classification

- Classification of isolated-note samples from musical instruments
  - By projecting each input sample as an unknown coefficient trajectory in PCA space and
  - Measuring a global distance between the interpolated, unknown trajectory $\breve{\mathcal{U}}$ and all prototype curves $\mathcal{C}_i$, defined as the average Euclidean distance between their mean points:

$$\frac{1}{R_{max}} \sum_{r=1}^{R_{max}} \sqrt{\sum_{k=1}^{D} (\breve{u}_{rk} - \mu_{irk})^2}.$$

  - Experiment: 5 classes, 1098 files, 10-fold cross-validation, 2 octaves (C4 to B5)
  - Comparison of Partial Indexing (PI) and Envelope Interpolation (EI): 20% improvement with EI
  - Comparison with MFCCs: 34% better with proposed representation method

Averaged classification accuracy
(10-fold cross-validated)

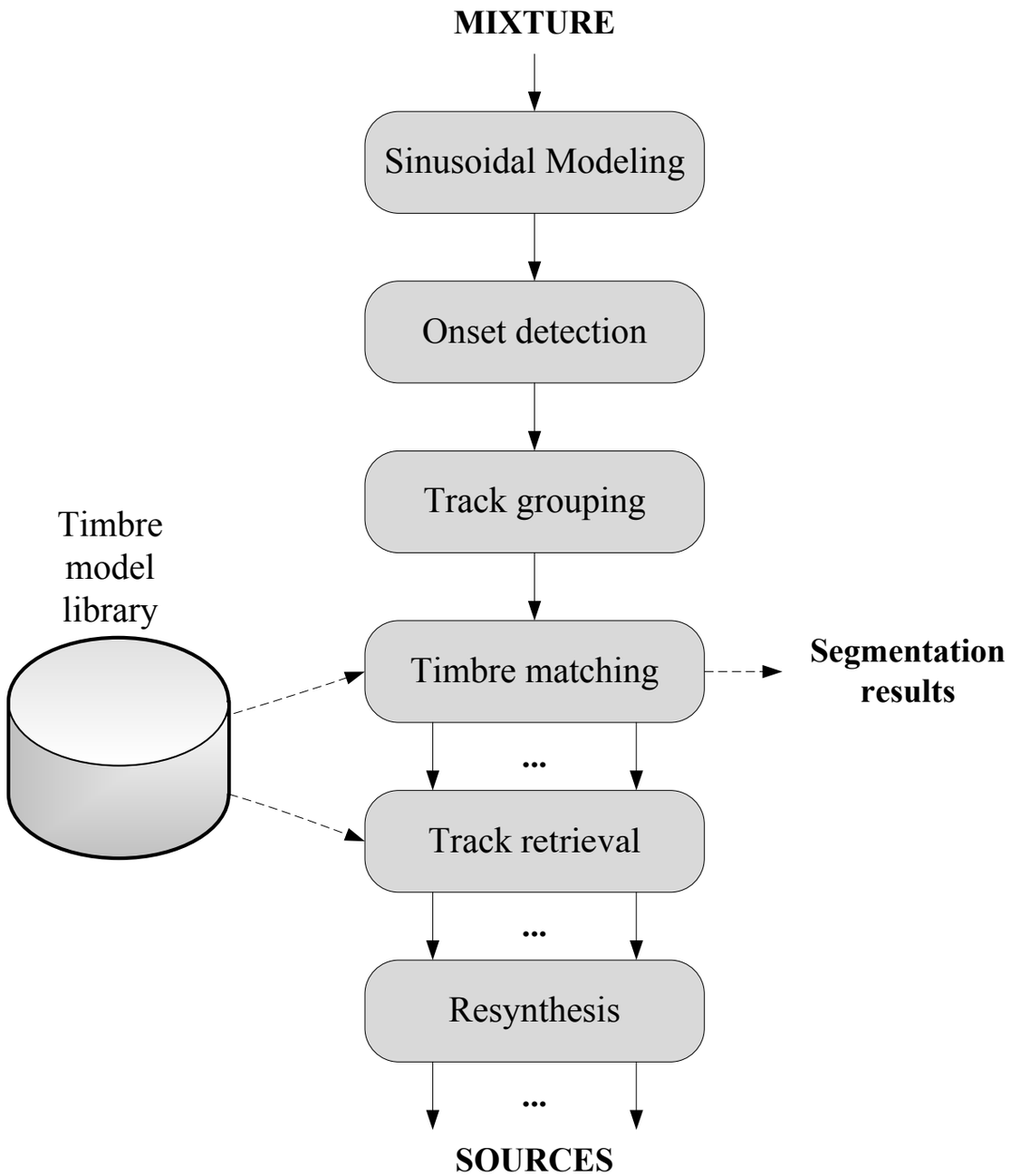


Maximum averaged classification accuracy and standard deviation (STD) (10-fold cross-validated)

Representation	Accuracy	STD
PI	74,86 %	± 2.84%
Linear EI	94,86 %	± 2.13%
Cubic EI	94,59 %	± 2.72%
MFCC	60,37 %	± 4.10%

Monaural separation: overview

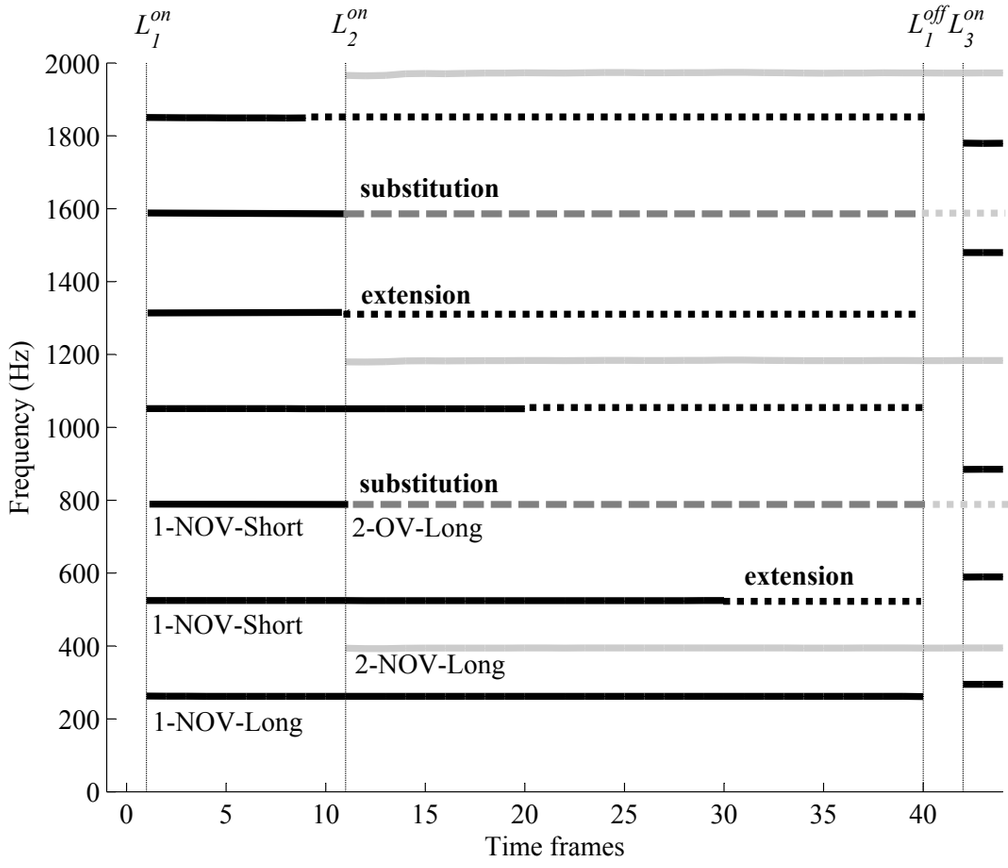
- One channel: the maximally underdetermined situation
  - Underlying idea: to use the obtained prototype envelopes as time-frequency templates to guide the sinusoidal peak selection and grouping for separation.
- Separation is only based on common-fate and good continuation cues of the amplitudes
  - No harmonicity or quasi-harmonicity required
  - No a priori pitch information needed
  - No multipitch estimation stage needed
  - It is possible to separate inharmonic sounds
  - It is possible to separate same-instrument chords as single entities
  - Outputs instrument classification and segmentation data
  - No need for note-to-source clustering
- Trade-off for the above
  - Onset separability constraint



[Burred&Sikora07] J.J. Burred and T. Sikora. Monaural Source Separation from Musical Mixtures based on Time-Frequency Timbre Models. In *Proc. ISMIR*, Vienna, Austria, September 2007.

Track grouping

- Inharmonic sinusoidal analysis on the mixture
- Simple onset detection
  - Based on the number of new sinusoidal tracks at any given frame, weighted by their mean frequency.
- Common-onset grouping of the tracks
  - Within a given frame tolerance from the detected onset.
- Each track on each group can be of the following types:
  1. Nonoverlapping (NOV)
  2. Overlapping with track from previous onset (OV)
  3. Overlapping with synchronous track (from the same onset)
- To distinguish between types 1 and 3:
  - Matching of individual tracks with the models
  - Unsufficient robustness in preliminary tests
  - Origin of onset separability constraint



Timbre matching (I)

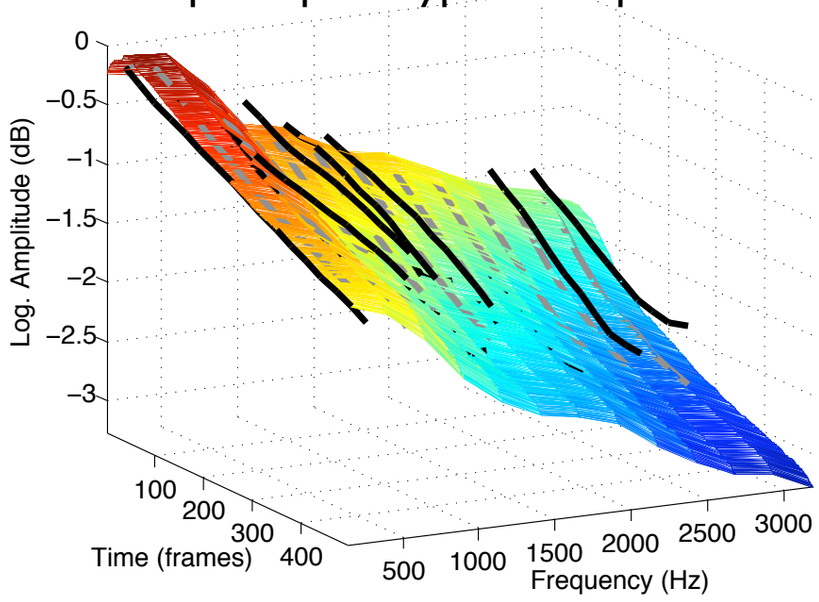
- Each common-onset group of nonoverlapping sinusoidal tracks $\mathcal{T}_o^{NOV}$ is matched against each stored prototype envelope.
- To that end, the following timbre similarity measures have been formulated:
  - Group-wise global Euclidean distance to the mean surface M

$$d(\mathcal{T}_o^{NOV}, \tilde{\mathbf{M}}_{io}) = \sum_{t \in \mathcal{T}_o^{NOV}} \sum_{r=1}^{R_t} |A_{tr} - \mathbf{M}_i(f_{tr})|$$

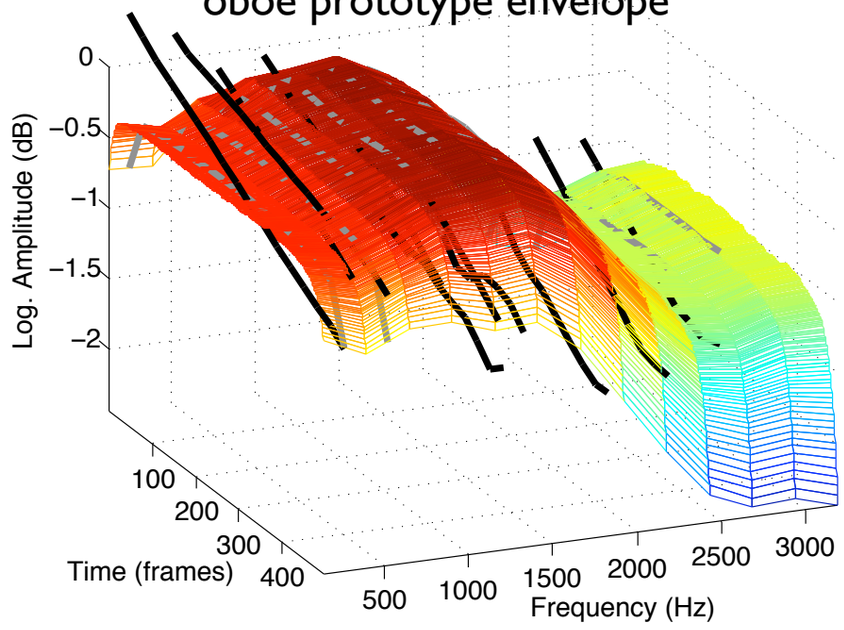
  - Group-wise likelihood to the Gaussian Process with parameter vector $\boldsymbol{\theta}_i = (\mathbf{M}_i, \mathbf{V}_i)$

$$L(\mathcal{T}_o^{NOV} | \boldsymbol{\theta}_i) = \prod_{t \in \mathcal{T}_o^{NOV}} \prod_{r=1}^{R_t} p\left(A_{tr} | \mathbf{M}_i(f_{tr}), \mathbf{V}_i(f_{tr})\right)$$

Good match: piano track group against piano prototype envelope



Bad match: piano track group against oboe prototype envelope



Timbre matching (2)

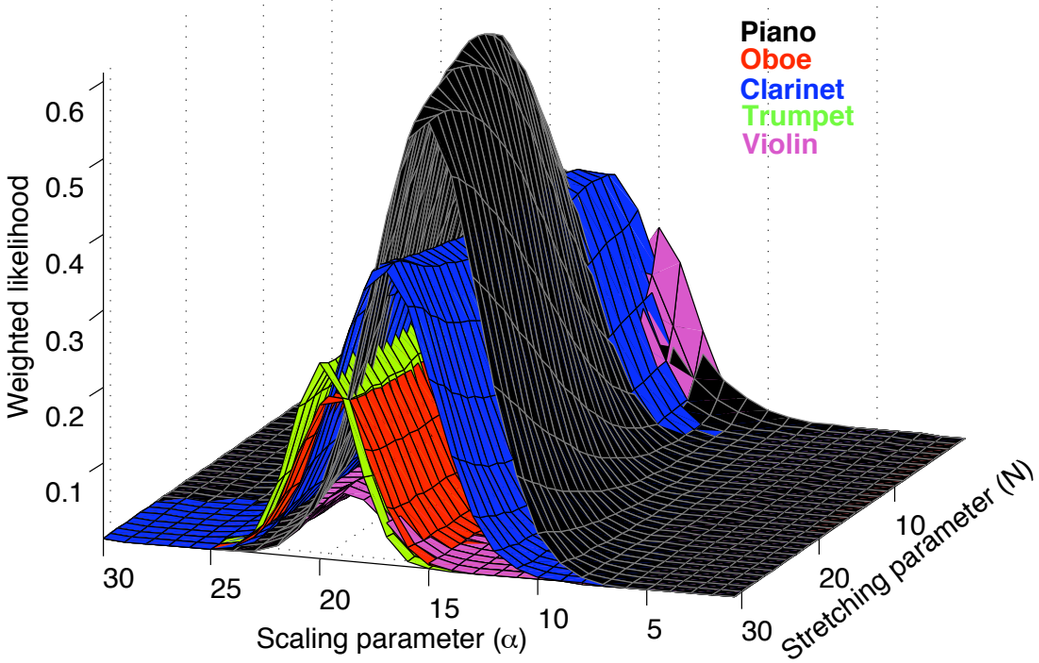
- To allow robustness against amplitude scalings and note lengths, the similarity measures are redefined as optimization problems subject to two parameters:
  - Amplitude scaling parameter α
  - Time stretching parameter N (A_{tr}^N and f_{tr}^N denote the amplitude and frequency values for a track that has been stretched so that its last frame is N.)

$$d(\mathcal{T}_o^{NOV}, \tilde{\mathbf{M}}_{io}) = \min_{\alpha,N} \left\{ \sum_{t \in \mathcal{T}_o^{NOV}} \sum_{r=1}^{R_t} |A_{tr}^N + \alpha - \mathbf{M}_i(f_{tr}^N)| \right\}$$

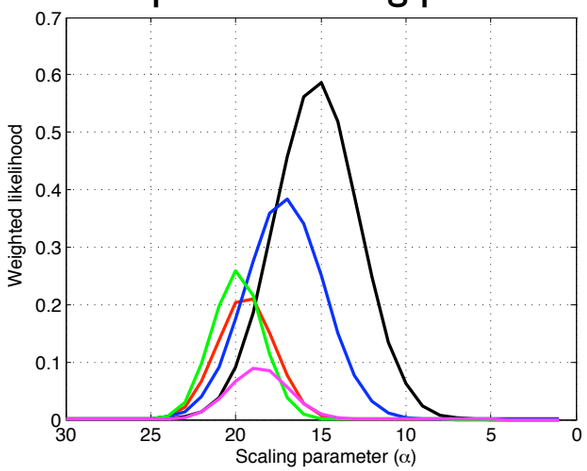
$$L_w(\mathcal{T}_o^{NOV}|\boldsymbol{\theta}_i) = \max_{\alpha,N} \left\{ \prod_{t \in \mathcal{T}_o^{NOV}} w_t \prod_{r=1}^{R_t} p\left(A_{tr}^N + \alpha | \mathbf{M}_i(f_{tr}^N), \mathbf{V}_i(f_{tr}^N)\right) \right\}$$

  - Weighted likelihood: $w_t = e^{R_t/\bar{f}_t}$
    - $\bar{f}_t$ is the track mean frequency
    - R_t is the track length
  - Unweighted likelihood: $w_t = 1$

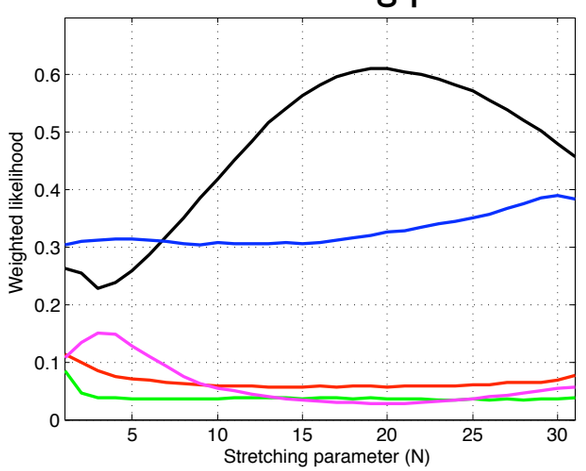
Exhaustive optimization surface (piano note)



Amplitude scaling profile



Time stretching profile



Application to polyphonic instrument recognition

- Same model library:
  - 5 classes (piano, clarinet, oboe, trumpet, violin)
- Each experiment contains 10 mixtures of 2 to 4 instruments
- Comparison of the 3 optimization-based timbre similarity measures
  - Euclidean, Likelihood and Weighted Likelihood
- Comparison between consonant intervals and dissonant intervals
- Note-by-note accuracy, cross-validated

Detection accuracy (%) for simple mixtures of one note per instrument

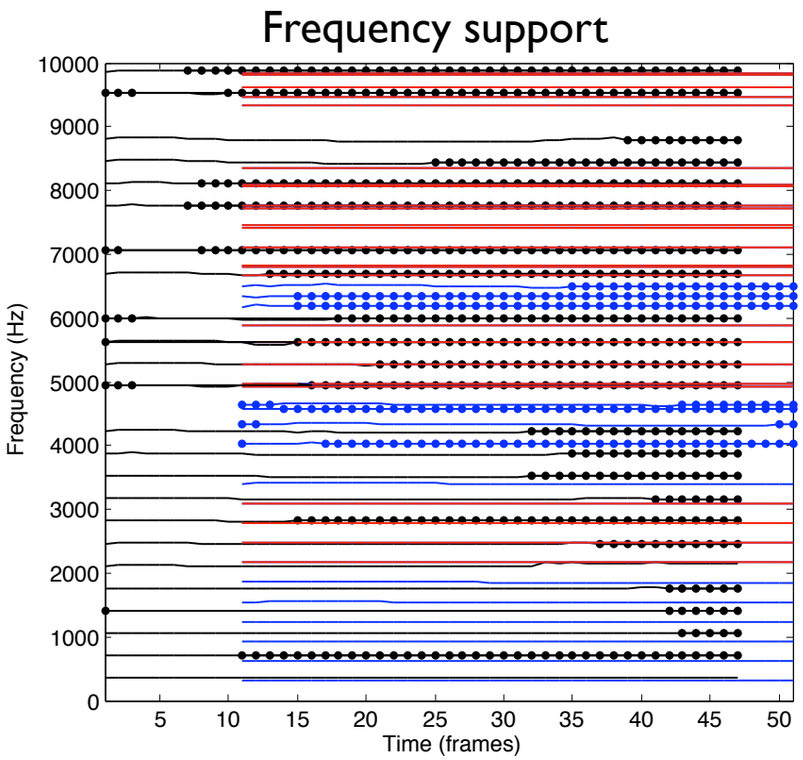
	Consonant (EXP 1)				Dissonant (EXP 2)			
Polyphony	2	3	4	**Av.**	2	3	4	**Av.**
Euclidean distance	63.14	34.71	40.23	46.03	73.81	69.79	42.33	61.98
Likelihood	66.48	**53.57**	**51.95**	**57.33**	**79.81**	57.55	56.40	64.59
Weighted likelihood	**76.95**	43.21	40.50	53.55	**79.81**	**77.79**	**61.40**	**73.00**

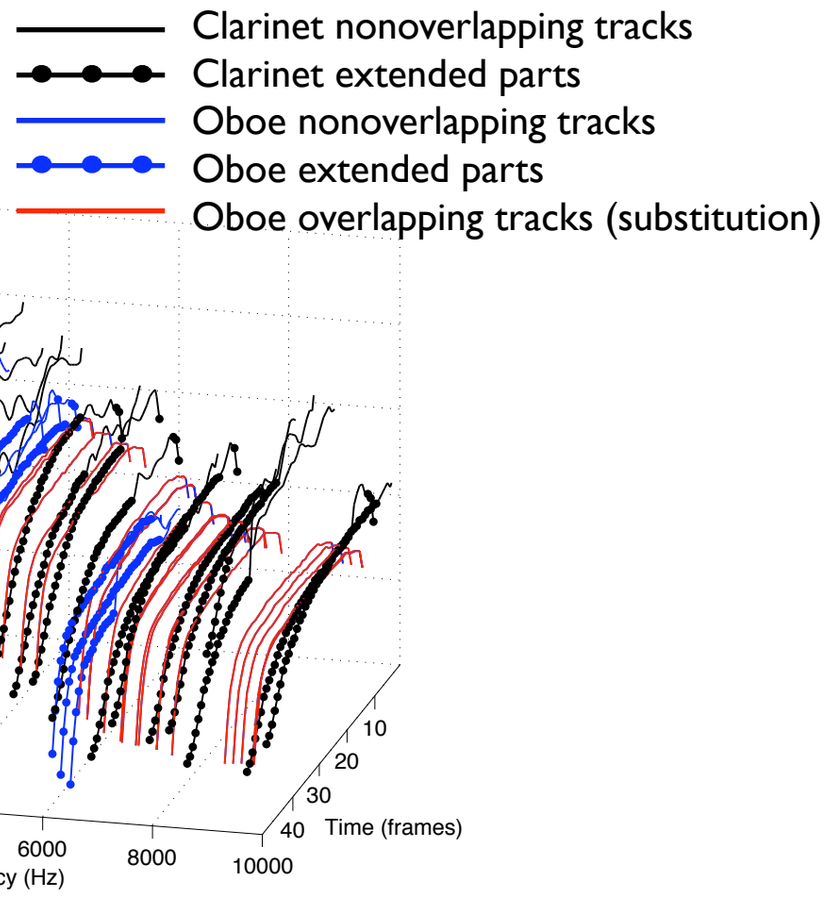
Detection accuracy (%) for mixtures of sequences containing several notes

	Sequences (EXP 3)		
Polyphony	2	3	**Av.**
Euclidean distance	64.66	50.64	57.65
Likelihood	63.68	**56.40**	**60.04**
Weighted likelihood	**65.16**	54.35	59.76

Track retrieval

- Goal: to retrieve the missing and overlapping parts of the sinusoidal tracks by interpolating the selected prototype envelope
- 2 operations:
  - Extension: tracks (of types 1 and 3) shorter than the current note are extended towards the onset (pre-extension) or towards the offset (post-extension), ensuring amplitude smoothness.
  - Substitution: overlapping tracks (type 2) are retrieved from the model in their entirety by linearly interpolationg the prototype envelope at the track's frequency support.
- Finally, the tracks are resynthesized by additive synthesis.





Evaluation of Mono Separation

- Experimental setups: (170 mixtures in total)

Type	Name	Source content	Harmony	Instruments	Polyphony
Basic	EXP 1	Individual notes	Consonant	Unknown	2,3,4
	EXP 2	Individual notes	Dissonant	Unknown	2,3,4
	EXP 3	Sequence of notes	Cons., Diss.	Unknown	2,3
	EXP 3k	Sequence of notes	Cons., Diss.	Known	2,3
Extended	EXP 4	One chord	Consonant	Unknown	2,3
	EXP 5	One cluster	Dissonant	Unknown	2,3
	EXP 6	Sequence with chords	Cons., Diss.	Known	2,3
	EXP 7	Inharmonic notes	-	Known	2

- Reference measure: Spectral Signal-to-Error Ratio (SSER)

$$\mathrm{SSER} = 10\log_{10}\frac{\sum_{r,k}|S(r,k)|^2}{\sum_{r,k}(|S(r,k)| - |\hat{S}(r,k)|)^2}$$

- Basic experiments:

Source type	Polyphony 2	3	4
Individual notes, consonant (EXP 1)	6.93 dB	5.82 dB	5.35 dB
Individual notes, dissonant (EXP 2)	9.38 dB	8.36 dB	5.95 dB
Sequences of notes (EXP 3k)	6.97 dB	7.34 dB	-

- Extended experiments:

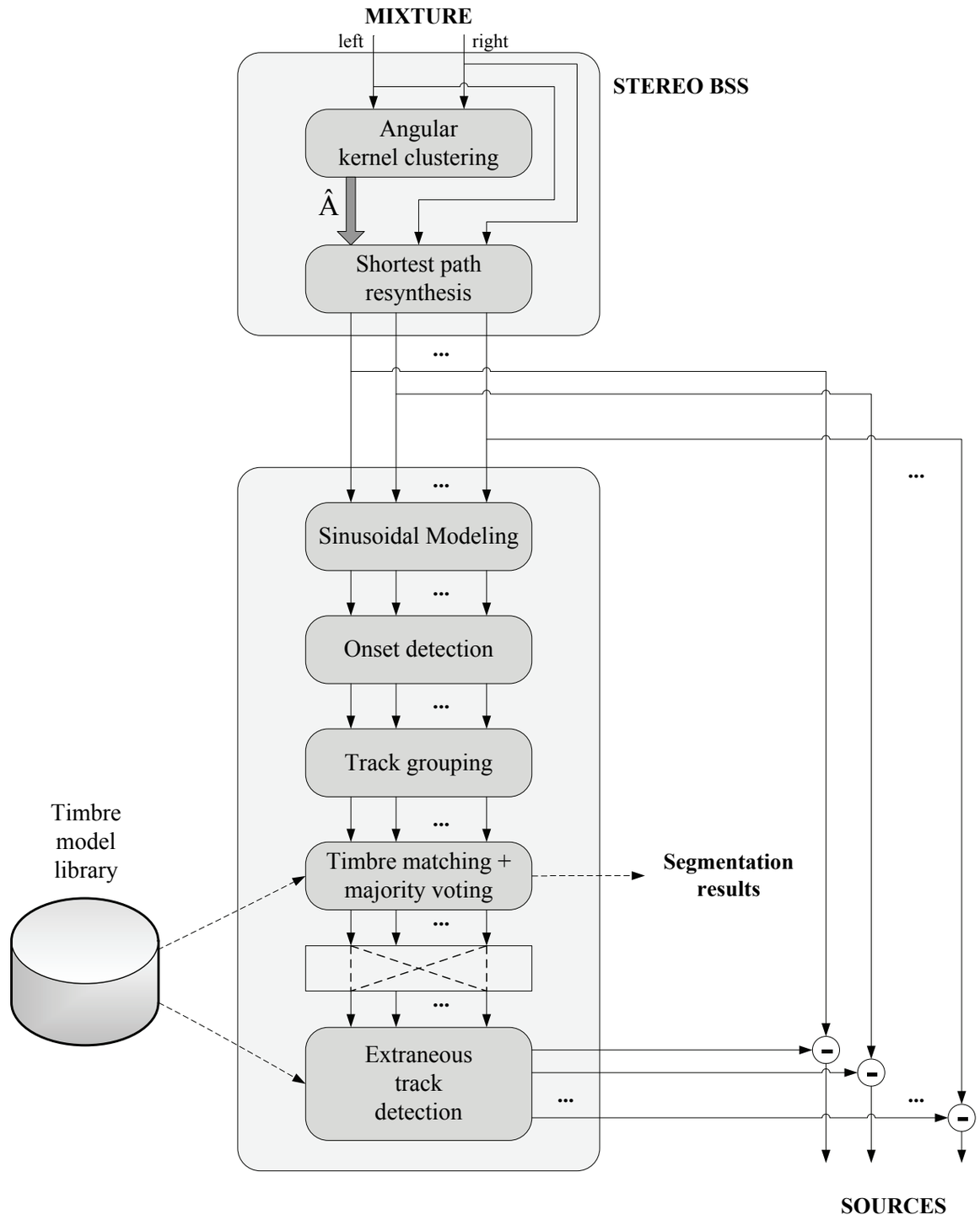
Source type	No. Instruments 2	3
One chord (EXP 4)	7.12 dB	6.74 dB
One cluster (EXP 5)	4.81 dB	4.77 dB
Sequences with chords and clusters (EXP 6)	4.99 dB	6.29 dB
Inharmonic notes (EXP 7)	7.84 dB	-

Stereo separation

- Extension of the previous mono system to take into account spatial diversity in linear stereo mixtures (M = 2)

$$x_m(t) = \sum_{n=1}^{N} a_{mn} s_n(t), \qquad m = 1, \ldots, M.$$

- Principle:
  - A first Blind Source Separation (BSS) stage exploiting spatial diversity for a preliminary separation, solely assuming sparsity (Laplacian sources). After [Bofill&Zibulevsky01].
  - Refine the partially-separated BSS channels applying a modified version of the previous sinusoidal and model-based methods.
- No onset separation required!



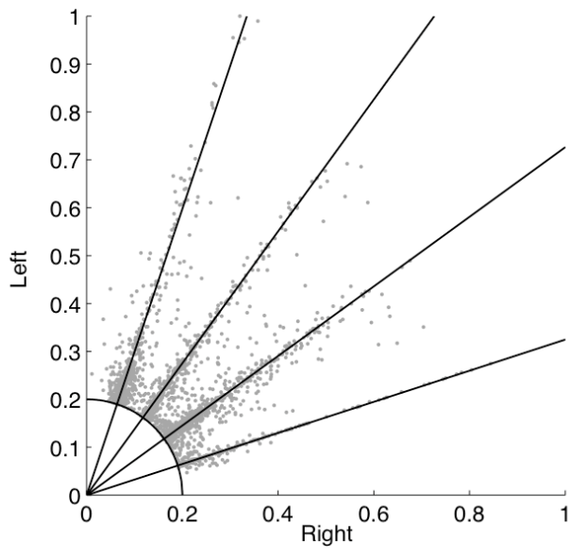
BSS stage: mixing matrix estimation

- To increase sparsity, both BSS stages are performed in the STFT domain.
- If the sources are enough sparse, the mixture bins (with radii $\rho_{rk} = \sqrt{x_{1,rk}^2 + x_{2,rk}^2}$ and angles $\theta_{rk} = \arctan(x_{2,rk}/x_{1,rk})$) concentrate around the mixing directions.
- The mixing matrix can be thus recovered by angular clustering.
- To smooth the obtained polar histogram, kernel-based density estimation is used, with a triangular polar kernel.

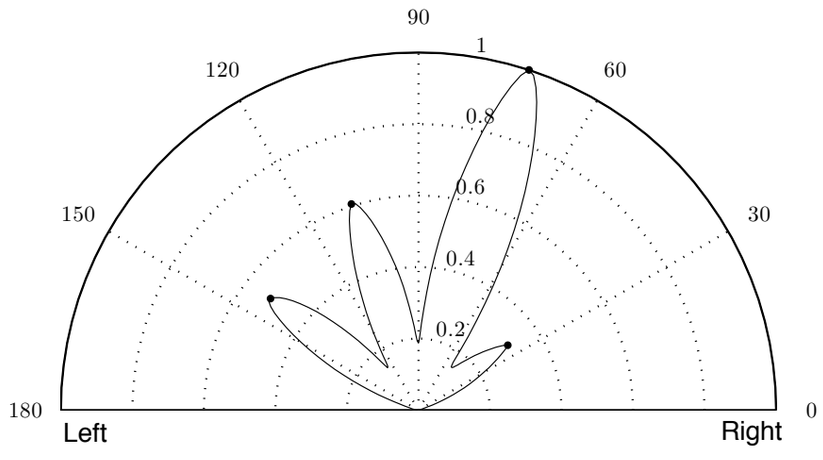
Estimated density: $$\hat{p}(\theta) = \sum_{r,k} \rho_{rk} K(\lambda(\theta - \theta_{rk}))$$

Triangular kernel: $$K(\theta) = \begin{cases} 1 - \frac{\theta}{\pi/4} & \text{if } |\theta| < \pi/4 \\ 0 & \text{otherwise} \end{cases}$$

Mixture scatter and found directions



Estimated density (polar)



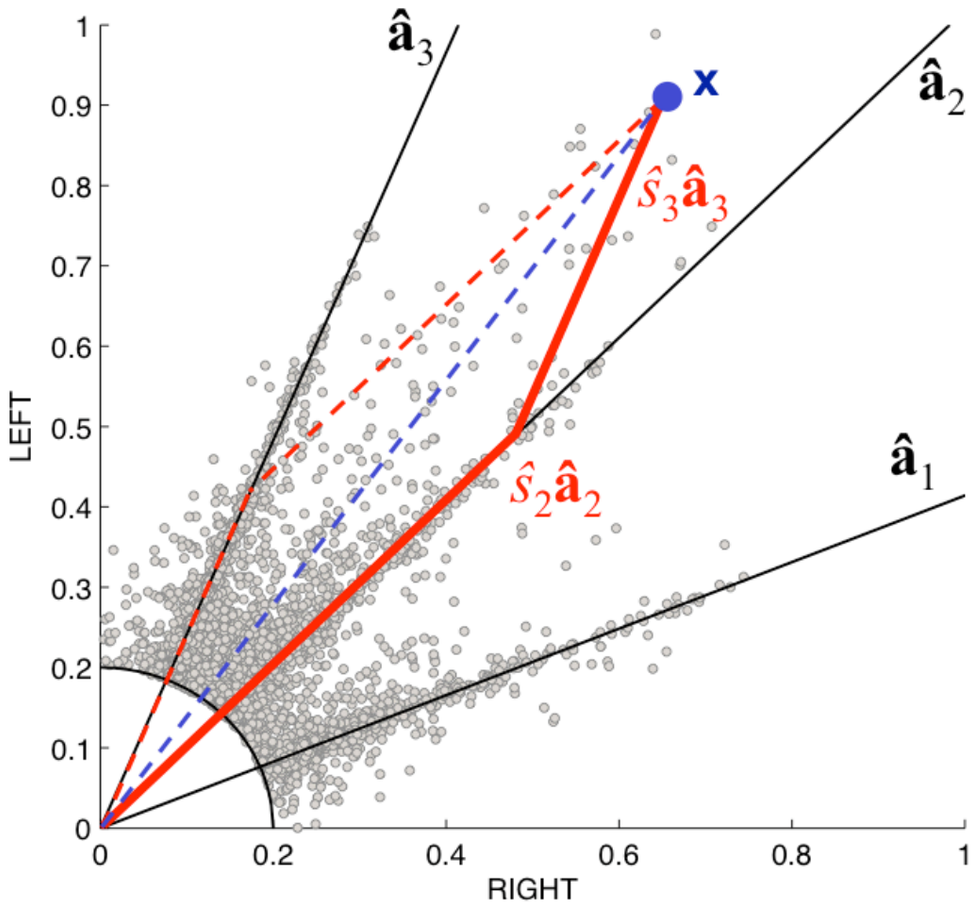
[Bofill&Zibulevsky01] P. Bofill and M. Zibulevsky. Underdetermined Blind Source Separation Using Sparse Representations. *Signal Processing*, Vol. 81, 2001.

BSS stage: source estimation

- Sparsity assumption: sources are Laplacian: $p(c) = \frac{\lambda}{2} e^{-\lambda|c-\mu|}$
- Given an estimated mixing matrix $\hat{\mathbf{A}}$ and assuming the sources are Laplacian, source estimation is the $L1$-norm minimization problem:

$$\hat{\mathbf{s}}_{rk} = \underset{\mathbf{x}_{rk}=\hat{\mathbf{A}}\mathbf{s}_{rk}}{\operatorname{argmin}} \left\{ \sum_{n=1}^{N} |s_{n,rk}| \right\}$$

Example of shortest-path resynthesis



- This minimization problem can be interpreted geometrically as the shortest-path algorithm:

• For each bin $\mathbf{x}$, a reduced 2 x 2 mixing matrix $\hat{\mathbf{A}}_\rho = [\hat{\mathbf{a}}_a, \hat{\mathbf{a}}_b]$ is defined, whose columns are the mixing directions enclosing it.

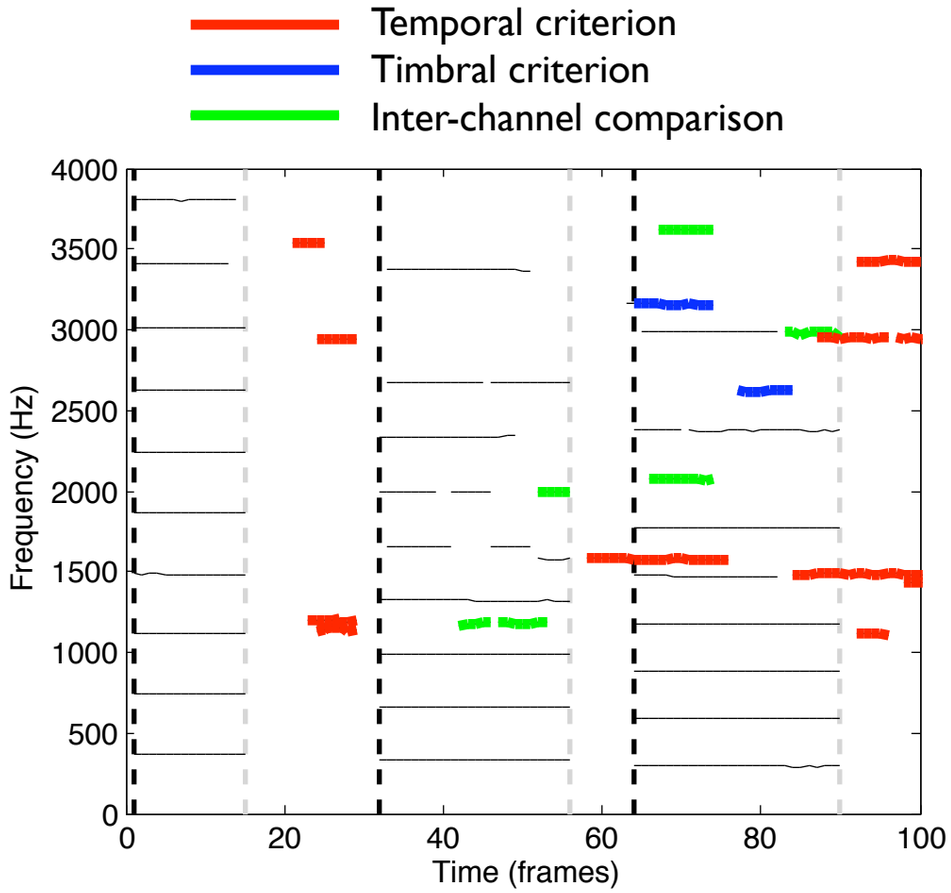
• Source estimation is performed by inverting the determined 2 x 2 subproblem and by setting all other N-M sources to zero:

$$\begin{cases} \hat{\mathbf{s}}_{\rho,rk} = \mathbf{A}_\rho^{-1}\mathbf{x}_{rk} \\ \hat{s}_n = 0, \quad \forall n \neq a, b. \end{cases}$$

Extraneous track detection

- After BSS, the same sinusoidal modeling, onset detection, track grouping and timbre matching stages are applied to the partially-separated channels.
  - All of these stages are now far more robust because the interfering sinusoidal tracks have already been partially suppressed.

Example: three piano notes, separated from a 3-voice mixture with an oboe and a trumpet.



- New module: extraneous track detection
  - Detects interfering tracks most probably introduced by the other channels, according to three criteria:

1. Temporal criterion. Deviation from onset/offset.
2. Timbral criterion. Matching of individual tracks, with the best timbre matching parameters. Length dependency must be cancelled:

$$L(\mathbf{t}_t|\boldsymbol{\theta}_i) = \left[\prod_{r=1}^{R_t} p\left(A_{tr}|\mathbf{M}_i(f_{tr}), \mathbf{V}_i(f_{tr})\right)\right]^{\frac{1}{R_t}}$$

3. Inter-channel comparison. Search tracks in the other channels with similar frequency support and decide according to average amplitudes.

- Finally, extraneous sinusoidal tracks are subtracted from the BSS channels.

Evaluation of Stereo Separation

- Same instrument model database (5 classes)
- 10 mixtures per experimental setup, 110 mixtures in total, cross-validated
- Polyphonic instrument detection accuracy (%):

Polyphony	Consonant (EXP 1s) 2	3	4	**Av.**	Dissonant (EXP 2s) 2	3	4	**Av.**
Euclidean distance	63.33	77.14	76.57	72.35	60.95	**86.43**	78.00	75.13
Likelihood	**86.67**	**84.29**	**82.38**	**84.45**	**81.90**	81.95	**81.33**	**81.73**
Weighted likelihood	70.00	70.95	66.38	69.11	78.10	78.62	74.67	77.13

Polyphony	Sequences (EXP 3s) 2	3	**Av.**
Euclidean distance	64.71	59.31	62.01
Likelihood	67.71	**74.44**	**71.08**
Weighted likelihood	**69.34**	58.34	63.84

- Separation quality
  - Apart from SSER, Source-to-Distortion (SDR), Source-to-Interferences (SIR) and Source-to-Artifacts Ratios (SAR) can be now computed (locked phases)
  - Comparison with applying only track retrieval to the BSS channels

Source type	Polyph.	Track retrieval SSER	Sinusoidal subtraction SSER	SDR	SIR	SAR
Individual notes, cons. (EXP 8s)	3	13.36	18.26	17.35	40.48	17.39
	4	14.88	15.31	14.96	36.25	15.06
Individual notes, diss. (EXP 9s)	3	11.88	21.72	20.91	44.56	21.03
	4	15.10	18.93	18.24	40.36	18.30
Sequences with chords (EXP 10s)	3	11.21	17.95	17.17	32.30	17.44
	4	10.57	12.16	11.18	26.26	11.51

Source type	Polyph.	Track retrieval SSER	Sinusoidal subtraction SSER	SDR	SIR	SAR
Individual notes, cons. (EXP 1s)	3	13.92	21.13	20.70	43.77	20.77
	4	12.10	17.13	16.78	40.83	16.83
Individual notes, diss. (EXP 2s)	3	14.37	24.20	23.63	47.01	23.72
	4	12.06	21.33	20.76	43.74	20.81
Sequences of notes (EXP 3s)	3	12.52	22.00	21.48	44.79	21.53

Overall improvements:

- Compared to mono separation:
  5-7 dB SSER
- Compared to stereo track retrieval:
  5-10 dB SSER
- Compared to using only BSS:
  2-4 dB SDR and SAR
  3-6 dB SIR

Conclusions

- Timbre models
  - Representation of prototype spectral envelopes as either curves in PCA space or templates in time-frequency
  - Use for musical instrument classification: 94.86% accuracy with 5 classes.
- Monaural separation (based on sinusoidal modeling and timbre models)
  - No harmonicity assumption: can separate inharmonic sounds and chords
  - No multipitch estimation
  - No note-to-source clustering
  - Drawback: onset separation required
  - Use for polyphonic instrument recognition: 79.81% accuracy for 2 voices, 77.79% for 3 voices and 61% for 4 voices.
- Stereo separation (based on sparsity-BSS, sinusoidal mod. and timbre models)
  - All the above features, plus:
  - Keeps (partially separated) noise part
  - Far more robust
  - No onset separation required
  - Better than only BSS and than stereo track retrieval
  - Use for polyphonic instrument recognition: 86.67% accuracy for 2 voices, 86.43% for 3 voices and 82.38% for 4 voices.

Outlook

- Separation-for-understanding applications
  - Use of the separation systems in music analysis or transcription applications
- Improvement of the timbre models
  - Test other transformations, e.g. Linear Discriminant Analysis (LDA)
  - Other methods for extracting prototype curves, e.g. Principal Curves
  - Separation of envelopes into Attack-Decay-Sustain-Release phases
  - Morphological description of timbre as connected objects (clusters, tails)
- Other applications of the timbre models
  - Further investigation into the perceptual plausibility of the generated spaces
  - Synthesis by navigation in timbre space
  - Morphological (object-based) synthesis in timbre space
- Improvement of timbre matching for classification and separation
  - Other timbre similarity measures
  - More efficient parameter optimization, e.g. with Dynamic Time Warping (DTW)
  - Avoiding the onset separation constrained in the monaural case.
- Extension to more complex mixtures
  - Delayed and convolutive (reverberant) mixtures
  - Higher polyphonies