



Transcription and separation of drum signals from polyphonic music



Seminar at Enterface'07 workshop

Gaël RICHARD

GET/Telecom Paris (ENST)

Signal and Image processing department





Drum Track Transcription and Separation

O. Gillet and G. Richard » Transcription and Separation of Drum Signals from Polyphonic Music », accepted for publication in the IEEE Trans. on ASLP

Content

- **Drum Track Transcription and Separation Drum track processing**
 - ⇒ **Drum track transcription using separation**
 - ⇒ **Drum track separation using transcription**
- **Short Demo**
 - ⇒ **Drum track remixing**
- **Conclusion**

Introduction

- What is a drum track ?

- ⇒ A polyphonic music signal $s(t)$: 🗣️

- ⇒ The corresponding drum track $d(t)$: 🗣️

- The aim of this work:

- to automatically obtain $d(t)$ from a monophonic or stereophonic signal $s(t)$
- and to automatically annotate $d(t)$ (e.g. obtain metadata describing which instrument of the drum is played when)



Introduction

- A growing effort of the Scientific community
 - ⇒ On the extraction of melodic and tonal information (multipitch estimation, melody transcription, chords and tonality estimation)
 - ⇒ on the estimation of the main rhythmic structure.
- However, too little effort to obtain detailed information about the rhythmic accompaniment played by the drum kit in polyphonic music,
- Wide range of interesting applications
 - ⇒ Musical genre identification (many genre are characterized by their stereotypical drum patterns).
 - ⇒ Query by rhythm information (query by tapping, beatboxing,..)
 - ⇒ Potential new and interesting ways of playing and enjoying music (drum track remixing or automatic DJing).

Drum Track Transcription

- **Monophonic cases**

- ⇒ Represent most work of Drum transcription (see [FitzGerald and J. Paulus, 2006] for a review)

- **Polyphonic cases : 3 classes of approaches**

- ⇒ **Segment & classify**

- Segment the signal in discrete events
- classify the discrete events (bass drum, snare drum,...)
- Some refs [Tanghe &al.2005], [Sandvold &al.2004], [Gillet & Richard2005], [Paulus2006]

- ⇒ **Match & Adapt**

- searching for occurrences of a reference temporal [Zils&al.2002] or time-frequency template [Yoshii&al.2004] within the music signal.

- ⇒ **Isolate & Identify**

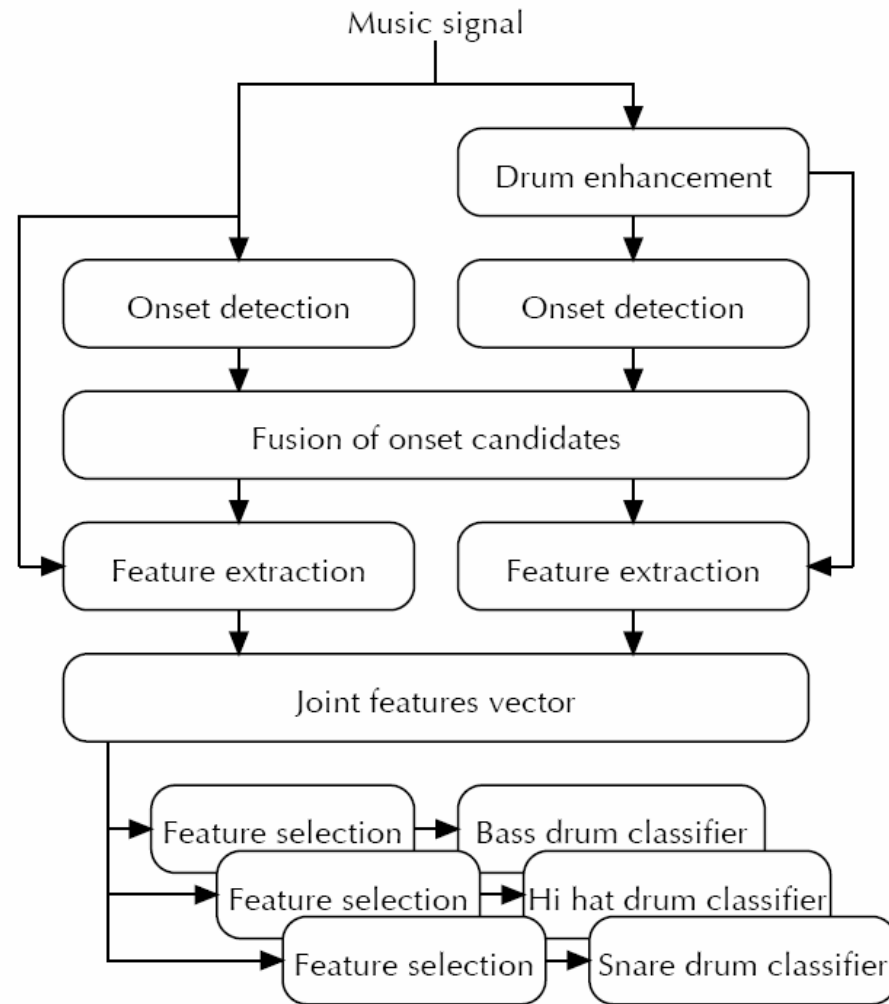
Drum Track Transcription

- **Isolate & Identify**

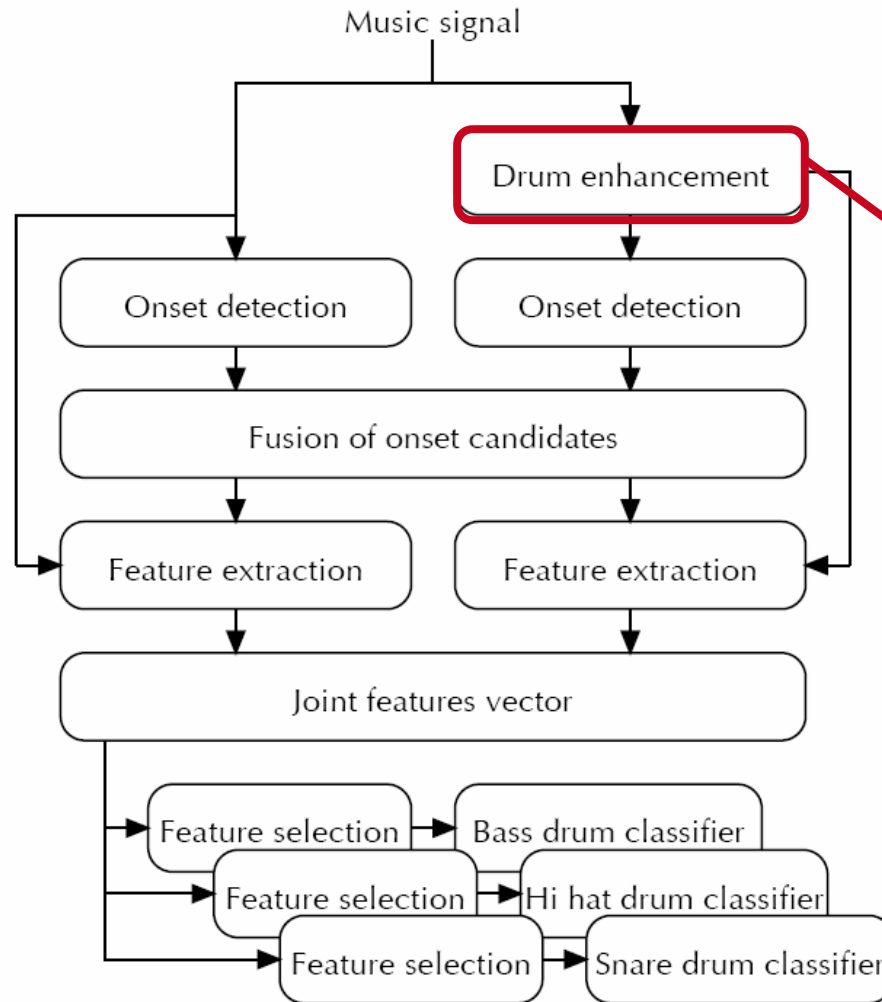
- ⇒ **Intuition:** the drum transcription process should simultaneously gain knowledge on the times at which drum instruments are played, and on their timbre.
- ⇒ **A possible approach:** Decomposition of time-frequency information into a set of independent components
 - Use of Independent Component Analysis
 - Use of Non-Negative Matrix factorization
 - Can be preceded by subspace analysis
 - Some ref: [Fitzgerald & al.2003], [Uhle & Dittmar2004], [Helén and Virtanen2005]
- ⇒ Such approaches highlight the links between music transcription and source separation

Drum Track Transcription

- Our approach combines
 - ⇒ Source separation (“Isolate”)
 - ⇒ Segment & classify transcription



Drum Track Transcription



⇒ Cancellation of harmonic sources from stereo signals

⇒ Bandwise harmonic/noise decomposition

Drum Track Enhancement

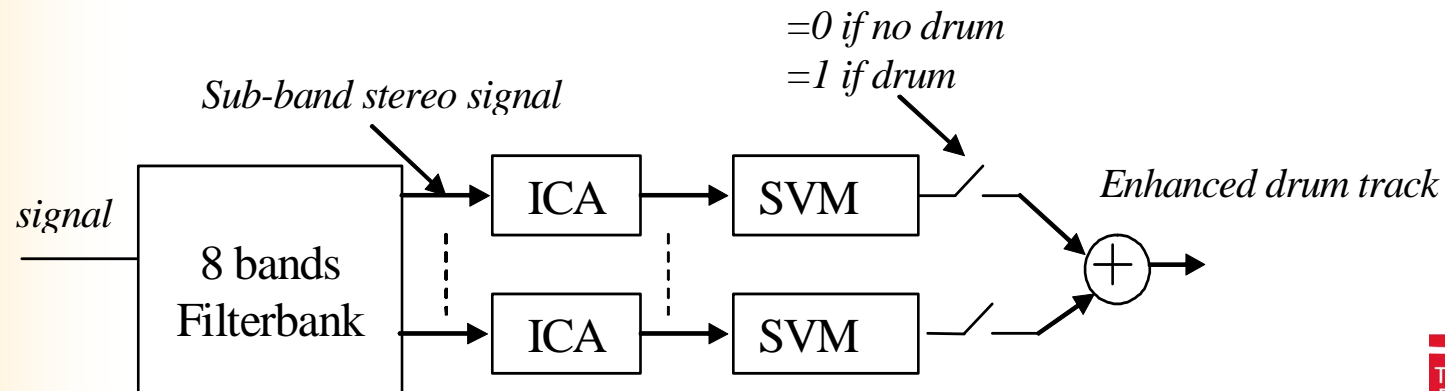
- Cancellation of harmonic sources from stereo signals

- ⇒ Assumptions

- most stereo music is generated by a panoramic mix
- some instruments in the mix are more predominant in some frequency bands than others.

In a narrow frequency band, the signal can be considered as a mixture of a predominant instrument, panned at a given position, and remaining components spread across the stereo field.

- ⇒ Proposed approach (similarly to Barry&al.2004):

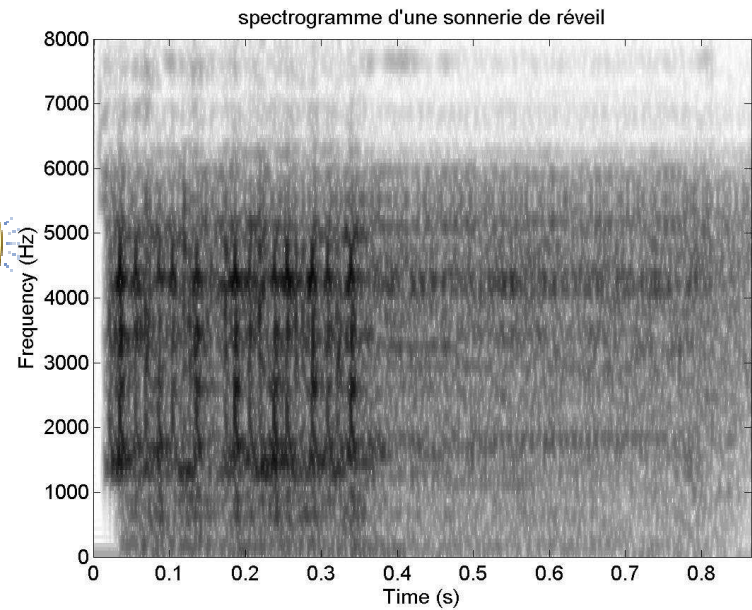
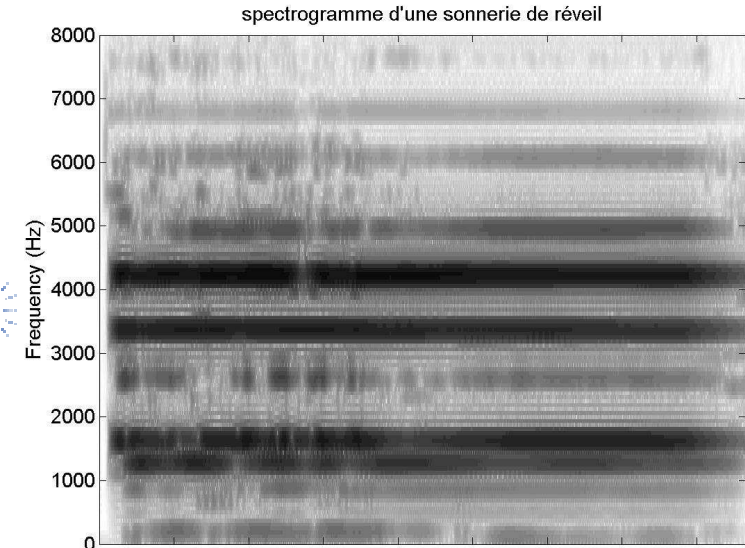
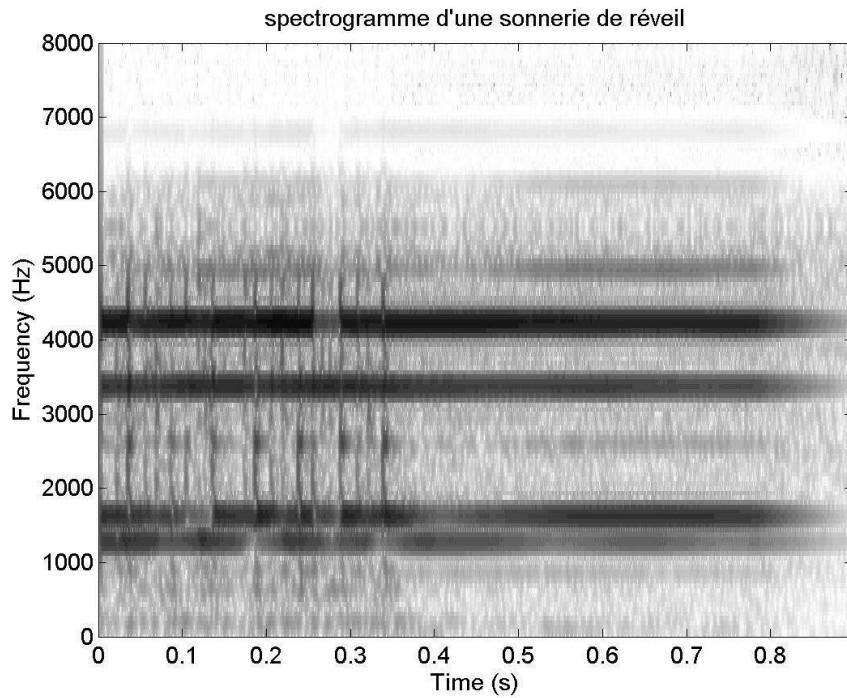


Drum Track Enhancement

- **Bandwise harmonic/noise decomposition**
 - ⇒ Assumptions
 - Drum instruments are mostly non-harmonic (unpitched) and present in well-defined frequency subbands
 - ⇒ Proposed approach:
 - Decomposition using a 8 octave-bands filterbank
 - Use of a Harmonic/noise decomposition

Feature extraction

- « Harmonic » + noise decomposition



Harmonic/noise decomposition

- Use of a signal model :

⇒ EDS model : *Exponentially Damped Sinusoidal model*

$$x(t) = s(t) + w(t)$$

⇒ *with*

$$s(t) = \sum_{m=1}^M a_m e^{-d_m t} \cos(2\pi f_m t + \phi_m)$$

$$s(t) = \sum_{m=1}^M \left(\alpha_m z_m^t + \alpha_m^* z_m^{*t} \right)$$

$$\alpha_m = \frac{1}{2} a_m \exp(i\phi_m)$$

$$z_m = \exp(-d_m + 2i\pi f_m)$$

« High Resolution » methods

● Principle:

⇒ Data matrix
(*Hankel structure*)

$$H = \begin{bmatrix} x(0) & x(1) & \dots & x(L-1) \\ x(1) & x(2) & \dots & x(L) \\ \vdots & \vdots & \ddots & \vdots \\ x(L-1) & x(L) & \dots & x(N-1) \end{bmatrix}$$

⇒ That can be written:

$$A = \text{Diag}(\alpha_1, \dots, \alpha_M, \alpha_1^*, \dots, \alpha_M^*)$$

$$H = E A E^T$$

$$E = \begin{bmatrix} 1 & \dots & 1 & 1 & \dots & 1 \\ z_1 & \dots & z_M & z_1^* & \dots & z_M^* \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ z_1^{L-1} & \dots & z_M^{L-1} & z_1^{*L-1} & \dots & z_M^{*L-1} \end{bmatrix}$$

⇒ Or by using a Eigen Value
Decomposition (EVD)

$$H = U \Lambda U^t$$

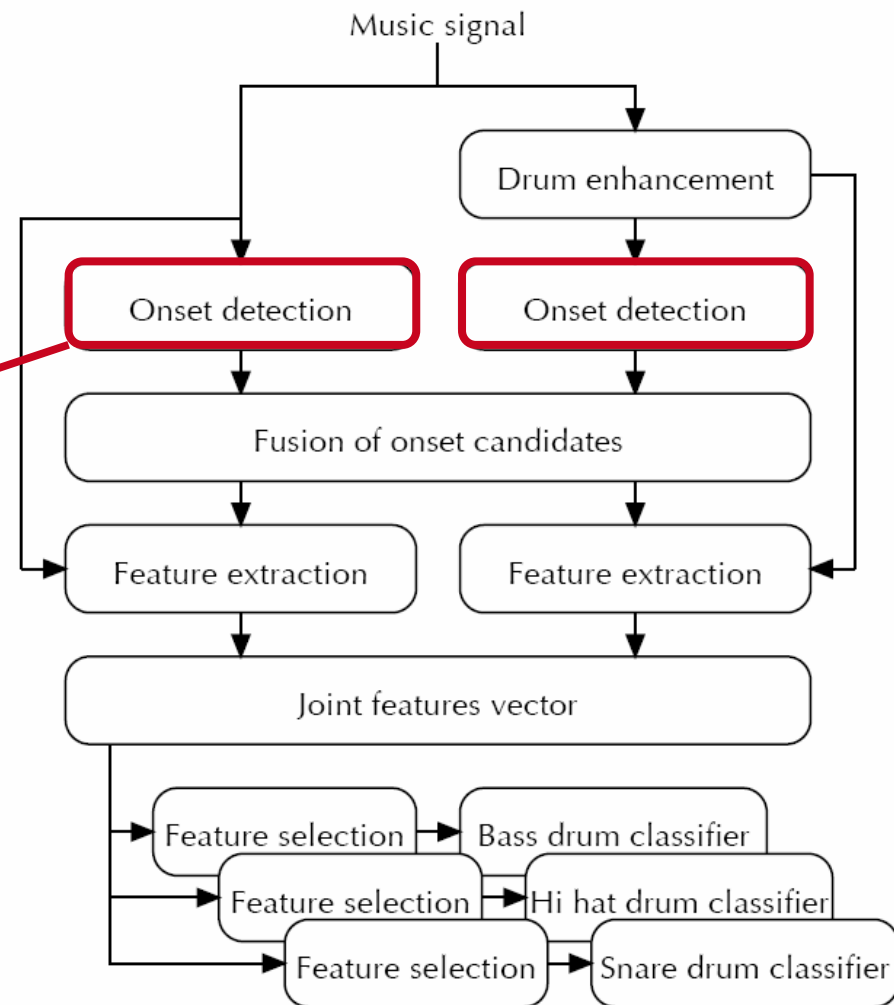
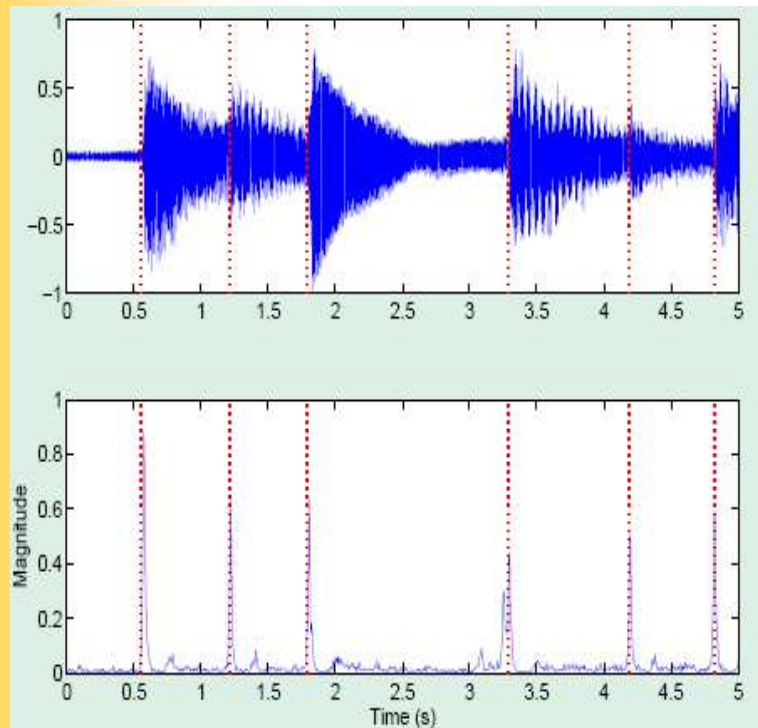
« High Resolution » methods

- ⇒ U^S is estimated from U (keeping the columns associated to the 2K dominant eigen vectors of the decomposition). It is a basis of the signal subspace
- ⇒ The signal and noise signals can be obtained by projection:

$$s = U_S U_S^H x$$

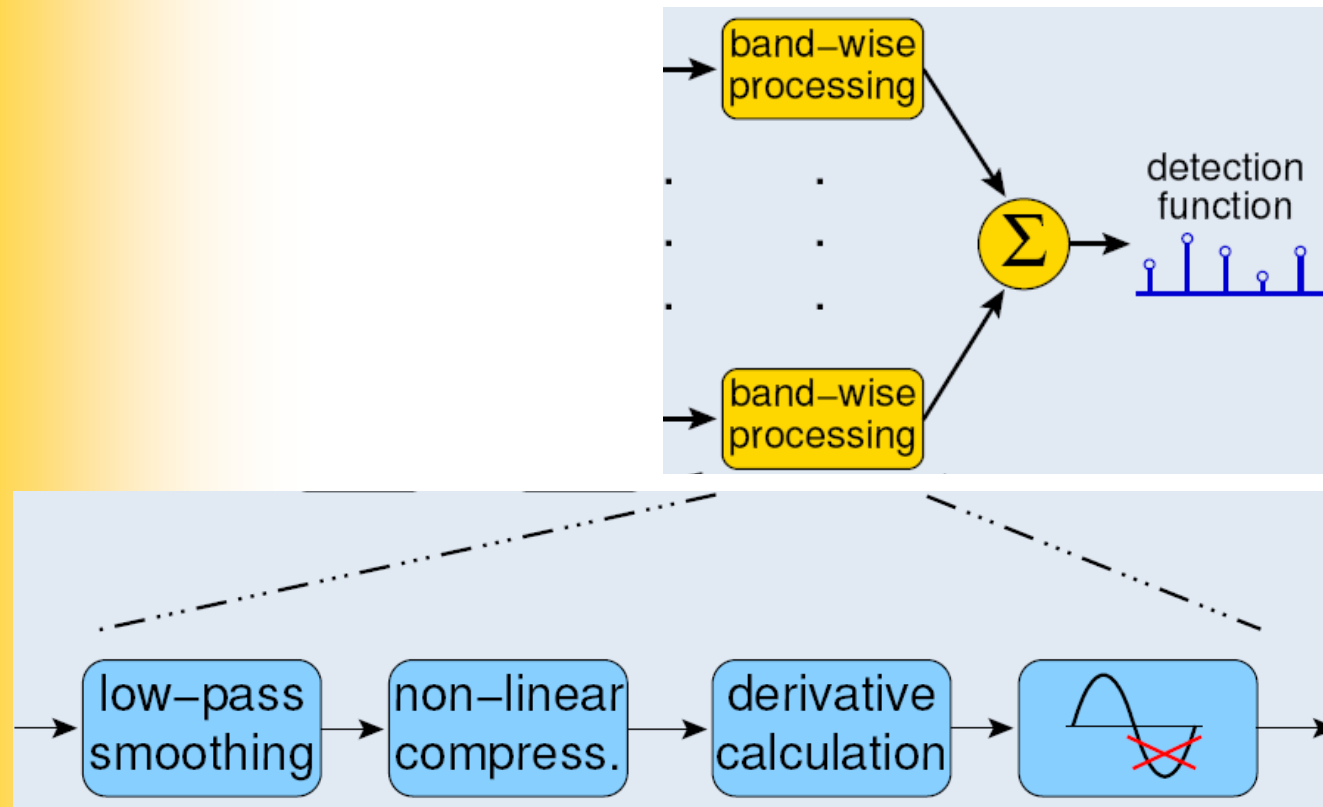
$$w = x - s = (I - U_S U_S^H) x$$

Drum Track Transcription



Drum track transcription

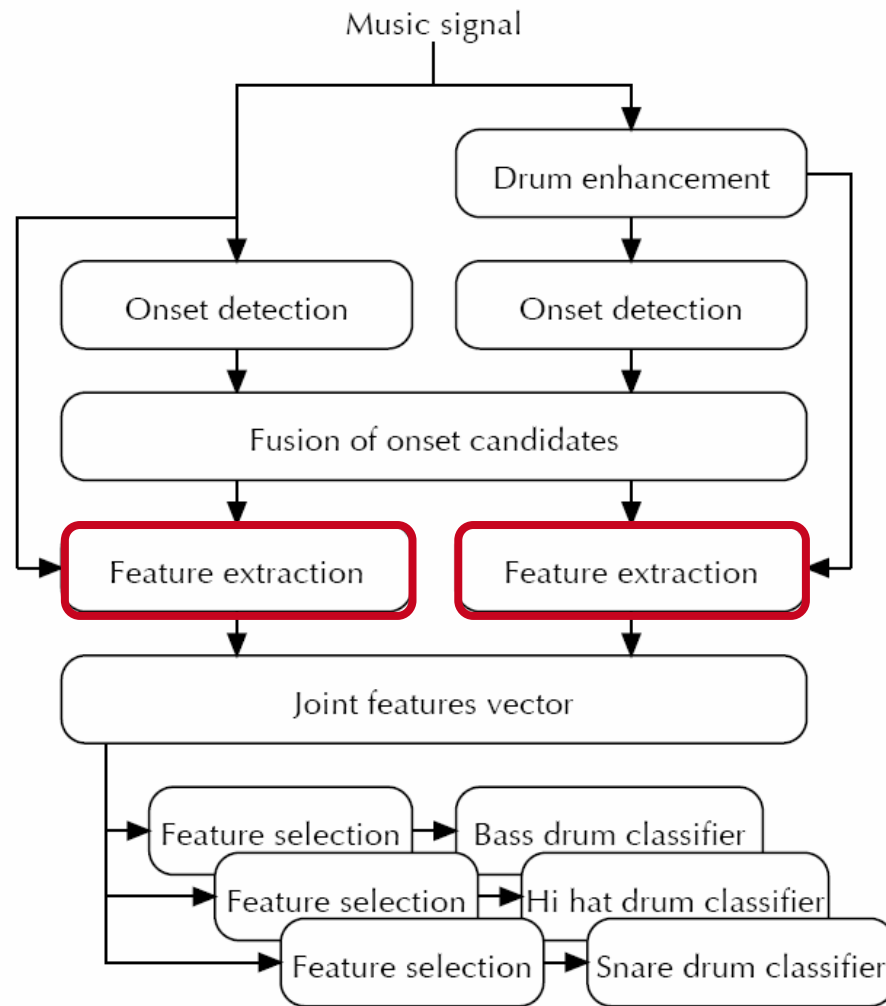
- Onset detection (Alonso&al.2007)



Drum track transcription

- Fusion of onsets from the two parts (original signal and drum enhanced signal):
 - ⇒ A simple sum
- Parameters of the onset detector are adjusted to favor a high recall rate (at the cost of a lower precision rate).
- Onsets of other instruments will be discarded by the classification stage

Drum Track Transcription



Drum Track Transcription

- Feature extraction

- ⇒ No consensus in the community on the ideal set of features for drum transcription
- ⇒ No consensus in the community on the ideal window duration/integration

- Our approach:

- ⇒ Use a large set of features and automatically select the most appropriate by feature selection algorithms
- ⇒ Compute the feature on the duration of a stroke (between two onsets previously detected).

Drum Track Transcription

- Our feature set: 147 features

(Most of them described in Peeters'04, IRCAM tech. report)

- ⇒ Temporal features (6)
- ⇒ Energy distribution features (23)
- ⇒ Cepstral features (78)
- ⇒ Perceptual features (26)

- ⇒ To obtain centered and unit variance features, a linear transformation is applied to each computed feature.

Drum Track Transcription

- 2 feature selection algorithms used:
 - ⇒ **IRMFSP based on Fisher determinant** (*Peeters et al.*)

Principle : « to select uncorrelated features one by one that enable good separation between classes with respect to the within-class spreads ».

- ⇒ **Recursive feature elimination with support vector machines (RFE-SVM)** [Guyon&al.2002]

- ⇒ *Principle*: “The RFE-SVM algorithm iteratively removes from the entire feature set the features whose contribution to the decision function of a linear support vector machine (SVM) classifier is minimal »

Drum Track Transcription

- Feature selection: results
 - ⇒ The final number of features retained was selected by a grid search from the set $D(d) = \{4, 8, 16, 32, 64, 96\}$.
 - ⇒ RFE-SVM performed better than IRMFSP except for small feature sets (less than 8 features).
 - ⇒ In the following IRMFSP is used for feature selection when $d \in \{4, 8\}$, and RFE-SVM is used in the other cases.

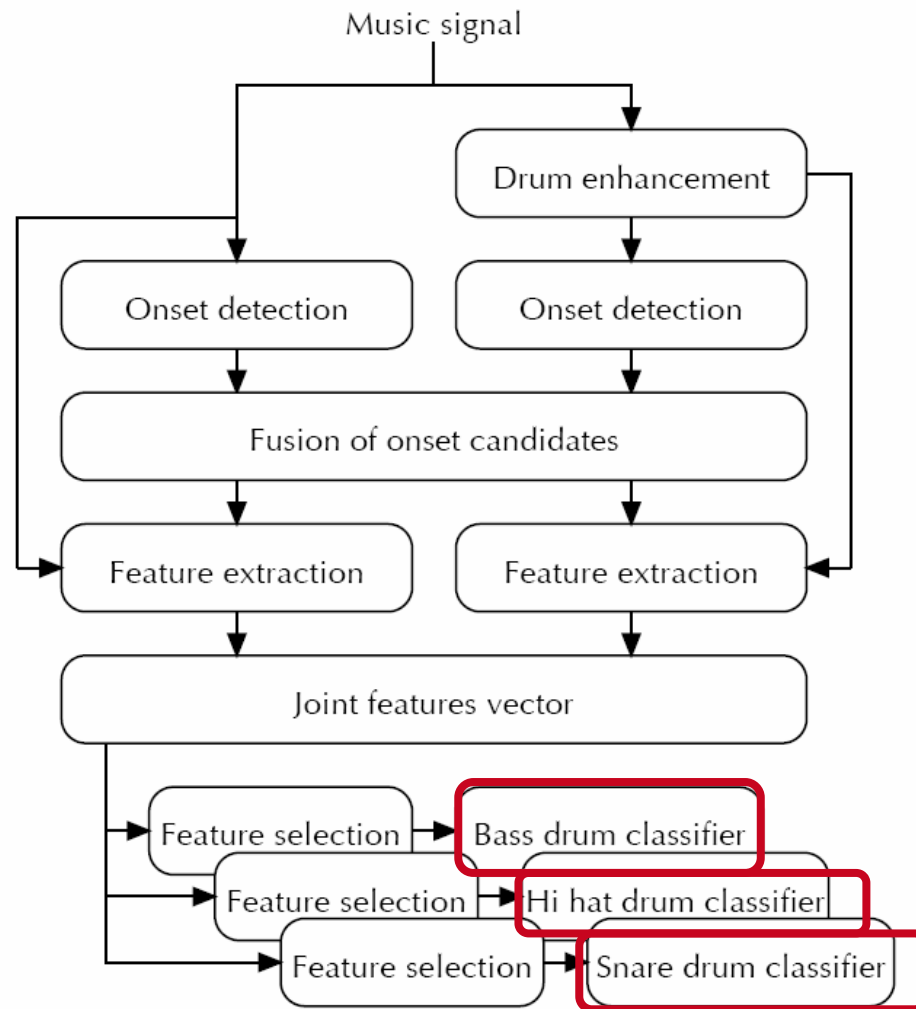
Drum track transcription: results

- Which features are selected (10 per instrument) ?

Temporal (T), Energy in frequency bands (E), Spectral (S), Cepstral (C) and Perceptual (P) subsets

	Original signal						Drum-enhanced signal					
	T	E	S	C	P	Total	T	E	S	C	P	Total
Accompaniment $-\infty$ dB												
BD	1	5	0	1	1	8	2	0	0	0	0	2
SD	1	1	1	1	1	5	0	2	1	1	1	5
HH	0	2	0	0	1	3	1	1	3	1	1	7
Accompaniment -6 dB												
BD	1	3	0	1	1	6	1	1	0	2	0	4
SD	2	1	0	1	0	4	0	3	0	3	0	6
HH	2	0	0	0	2	4	1	0	3	1	1	6
Accompaniment $+0$ dB												
BD	0	2	0	0	0	2	1	4	0	3	0	8
SD	2	2	0	0	0	4	2	1	0	3	0	6
HH	1	0	0	0	0	1	1	1	5	1	1	9
Accompaniment $+6$ dB												
BD	0	4	0	0	0	4	1	4	0	1	0	6
SD	2	1	0	0	0	3	2	3	0	2	0	7
HH	2	0	0	0	0	2	1	0	4	0	3	8

Drum Track Transcription



Drum Track Transcription: Classification

- Classification of K instruments ($K=3$: {bass drum, snare drum, hi-hat}): 2 possible approaches:

- ⇒ one $2^{|K|}$ -class classifier
- ⇒ $|K|$ binary classifiers

each of them detecting the presence or absence of a target instrument.

- one $2^{|K|}$ -class classifier: homogenous classes in unbalanced proportions.
- $|K|$ binary classifiers :
 - ⇒ less homogenous classes
 - ⇒ but the number of positive and negative training examples is more balanced for each classifier.

Drum Track Transcription: Classification

- Classifier used: C- Support Vector Machines (C-SVM)
 - ⇒ With a normalized Gaussian kernel

$$K(x, y) = \exp\left(-\frac{\|x - y\|^2}{2d\sigma^2}\right)$$

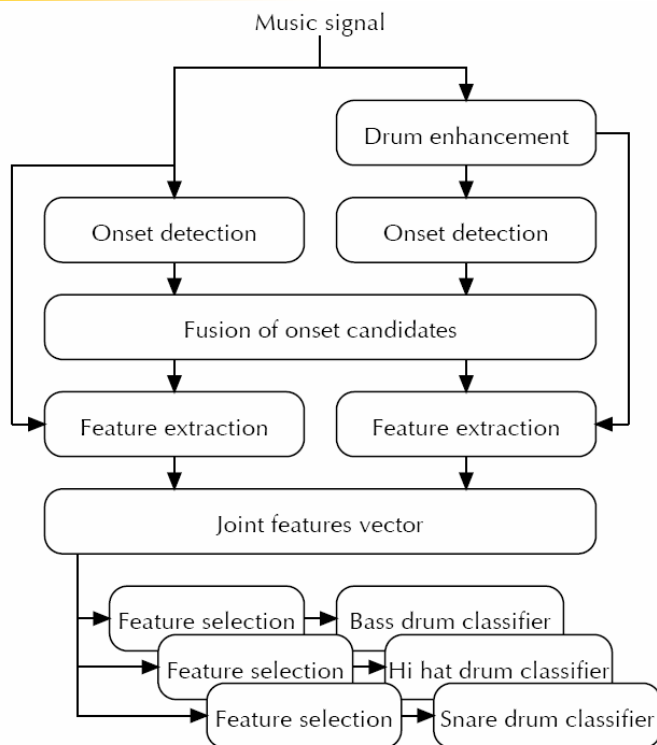
where d is the number of features

- Posterior probabilities of class membership is obtained according to the method described by Platt [Platt2000] (also used for further information fusion)

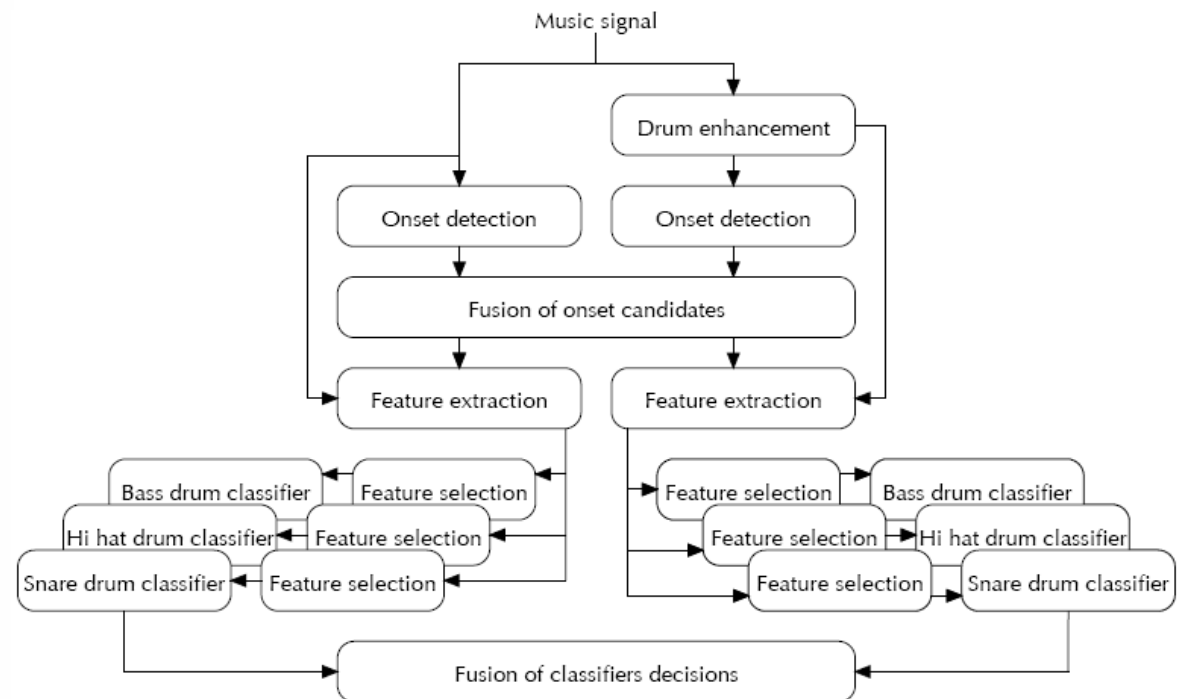
Drum Track Transcription: Fusion

- 2 different fusion schemes

Early fusion



Late fusion



Drum Track Transcription: Fusion

- Several fusion approaches were tested, but only two are retained:
 - ⇒ **Sum**
 - ⇒ **Maximum**

Evaluation: use of ENST-Drums (Ismir'06)

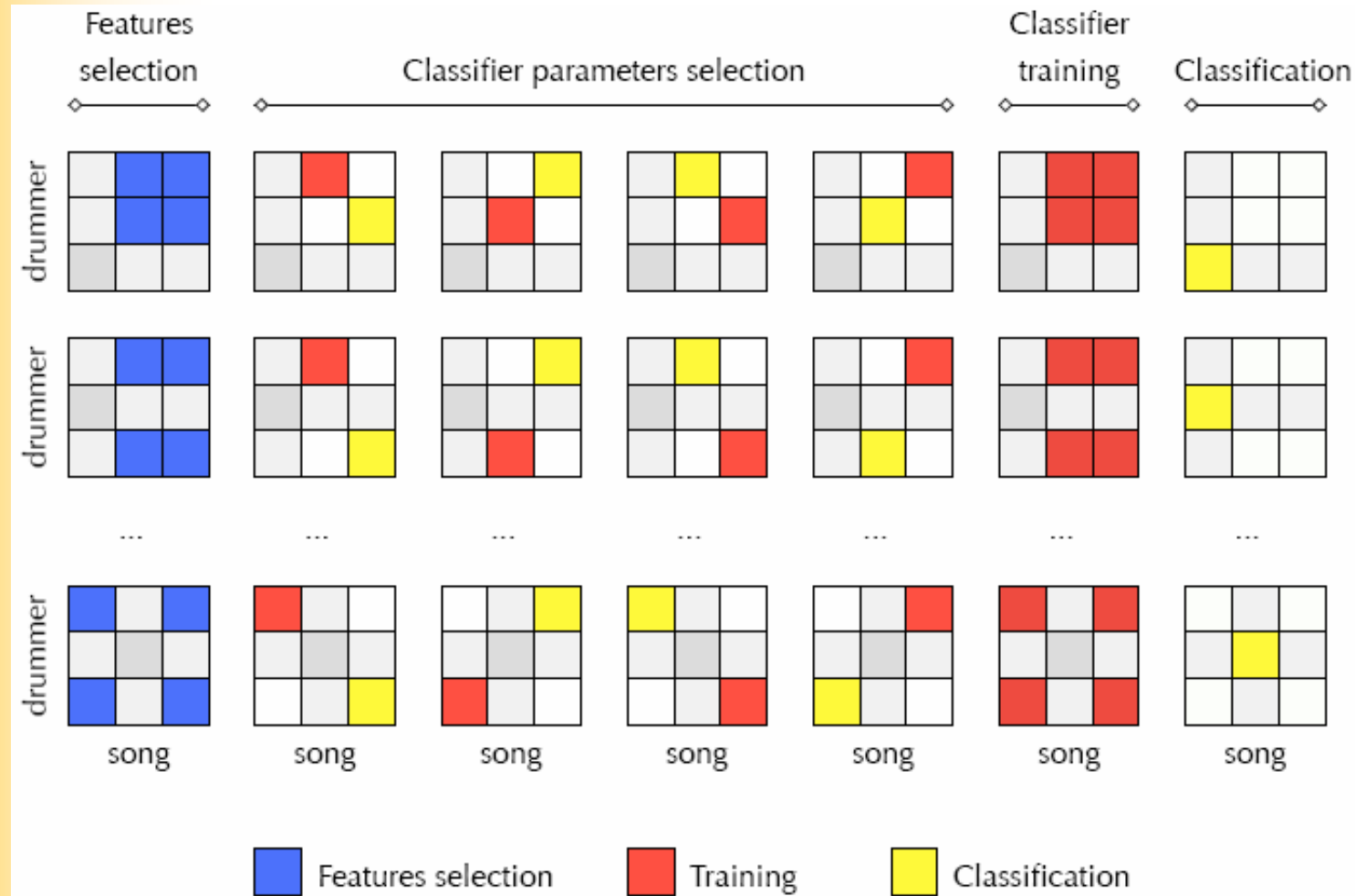
- A fully annotated audiovisual database for
 - ⇒ Automatic drum transcription
 - ⇒ Drum track extraction / separation
 - ⇒ Multimodal drum performance analysis
 - ⇒ Evaluation (a large part is publicly available at production cost)

Item	Drummer 1		Drummer 2		Drummer 3	
	Sequences	Events	Sequences	Events	Sequences	Events
Hits	29	139	31	180	48	283
Phrases	66	5339	74	9305	68	10467
Soli	7	1420	5	1613	5	1983
Accompaniment (Minus one CD)	17	8856	17	8788	17	9382
Accompaniment (MIDI file)	24	8224	24	6274	24	7357
Total	143	23978	151	26160	162	29472

ENST-Drums (Ismir'06)



The nested cross-validation protocol



Drum track transcription: evaluation

- Evaluation metrics: Precision & Recall

$$P = \frac{N_c}{N_d} \quad , \quad R = \frac{N_c}{N}$$

- ⇒ N_d : the number of strokes detected by the system
- ⇒ N_c : the number of correct strokes detected
- ⇒ N : the total number of true strokes for instrument k

- F-measure is:

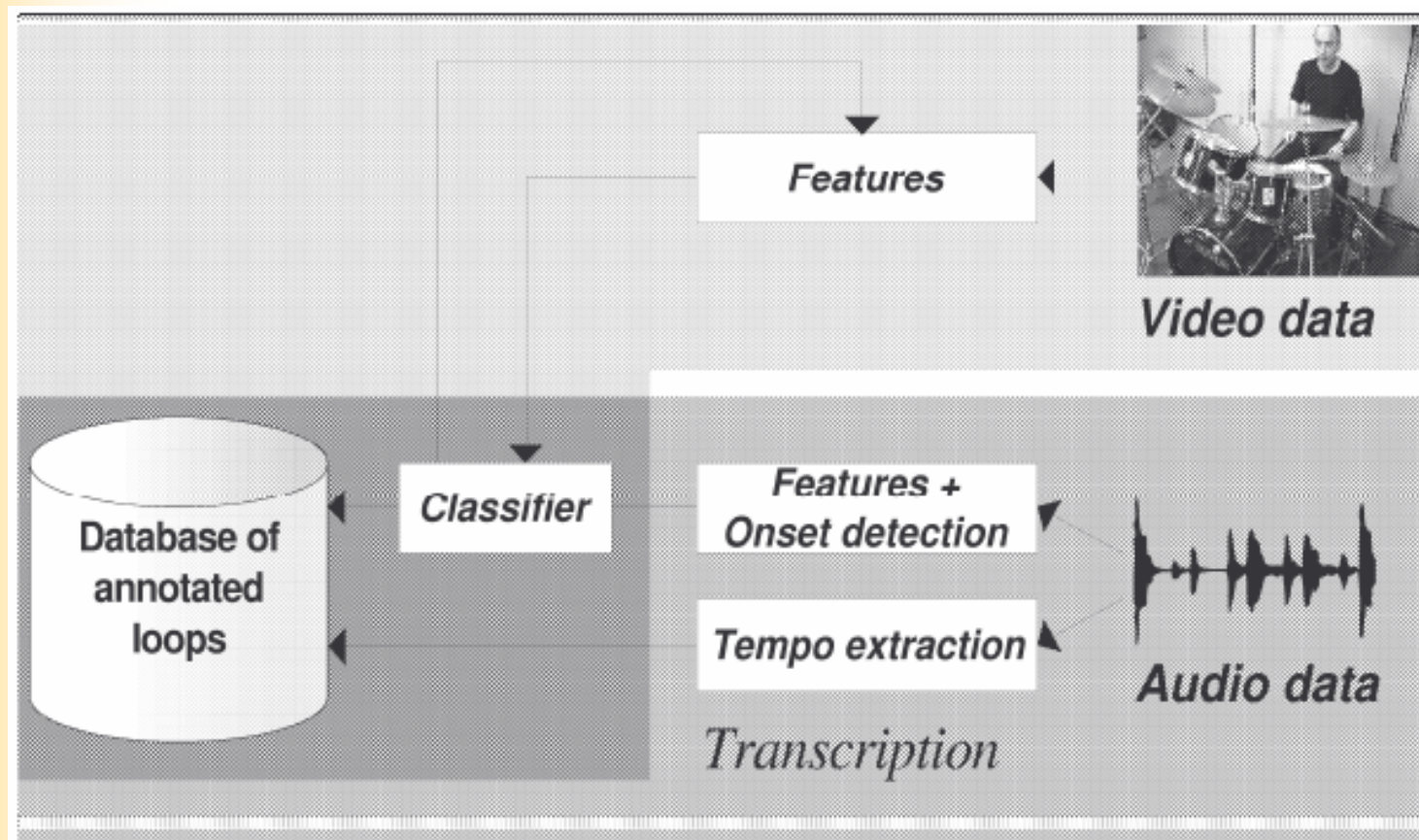
$$\text{F-measure} = \frac{2PR}{P + R}$$

Drum track transcription: results

Instrument	Original signal only			Drum-enhanced signal			Early fusion			Late fusion		
	Prec.	Recall	F meas.	Prec.	Recall	F meas.	Prec.	Recall	F meas.	Prec.	Recall	F meas.
Accompaniment $-\infty$ dB												
BD	66.4%	67.8%	67.1%	60.4%	75.2%	67.0%	62.8%	62.7%	62.8%	65.6%	80.5%	72.3%
SD	52.4%	80.1%	63.3%	57.0%	70.1%	62.9%	51.1%	78.3%	61.8%	58.5%	75.7%	66.0%
HH	81.3%	76.8%	79.0%	82.5%	78.6%	80.5%	86.5%	76.6%	81.3%	85.2%	79.2%	82.1%
Accompaniment -6 dB												
BD	65.7%	72.1%	68.7%	54.3%	69.3%	60.9%	63.7%	61.5%	62.6%	64.6%	79.2%	71.1%
SD	54.7%	72.4%	62.3%	57.3%	69.0%	62.6%	56.6%	75.1%	64.5%	57.7%	73.2%	64.5%
HH	81.2%	75.8%	78.4%	79.5%	78.4%	79.0%	80.5%	77.3%	78.9%	82.4%	78.2%	80.3%
Accompaniment $+0$ dB												
BD	61.7%	58.4%	60.0%	54.1%	65.8%	59.4%	61.1%	61.0%	61.1%	62.0%	70.2%	65.8%
SD	46.4%	66.7%	54.7%	50.6%	66.1%	57.4%	52.0%	69.5%	59.5%	50.6%	70.7%	59.0%
HH	80.8%	70.6%	75.4%	79.5%	73.3%	76.3%	78.9%	74.9%	76.8%	83.1%	73.0%	77.7%
Accompaniment $+6$ dB												
BD	60.0%	54.3%	57.0%	55.1%	58.5%	56.8%	55.5%	54.9%	55.2%	60.9%	62.6%	61.7%
SD	37.6%	54.7%	44.6%	41.3%	56.5%	47.7%	48.0%	58.7%	52.8%	42.8%	60.4%	50.1%
HH	76.7%	65.6%	70.6%	74.7%	68.4%	71.4%	74.7%	67.7%	71.1%	78.0%	68.0%	72.6%

Use the video to transcribe drum sequences

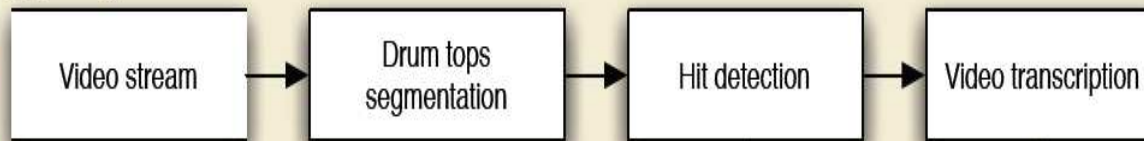
- Use of audiovisual features



Multimodal drum transcription

(from K. Mc Guinness & al., Eusipco07)

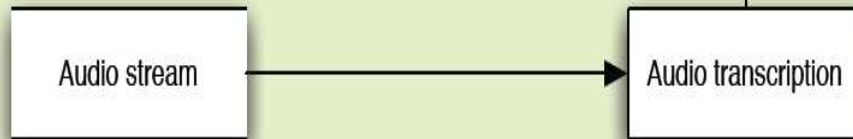
Video processing



Multimodal processing



Audio processing



Drum track separation

- Towards improved separation and remixing
 - ⇒ Time/Frequency/Subspace Masking
 - Score informed separation
 - ⇒ Separation with Wiener filtering
 - Improvement for drum track separation

Drum Track separation

- Time/Frequency/Subspace Masking

⇒ As previously seen, a « drum signal » can be reconstructed from the 8 subbands of the noise signal:

$$\sum_{k=1}^8 \hat{x}_{\tau k}$$

⇒ Possible improvements:

$$s = \sum_k \alpha_{hk} \hat{x}_{hk} + \alpha_{\tau k} \hat{x}_{\tau k}$$

- With, for each drum instrument frequency/subspace temporal envelopes, which model the distribution of energy in the harmonic and stochastic components of each sub-band

Drum Track separation

- Time/Frequency/Subspace Masking

- ⇒ **1. Extraction of the frequency/subspace temporal envelopes:**

- Amplitude envelope for the harmonic and noise component for each instrument and subbands are computed
- Fit these envelope signals with exponentially decaying envelopes resulting in envelope signals

$$e_{hk}^i \quad e_{rk}^i$$

- Average on several solo hits.

Drum Track separation

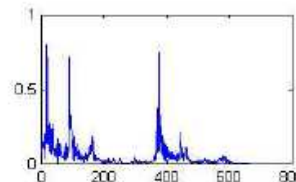
- Time/Frequency/Subspace Masking

⇒ **2. Detection of Drum events:**

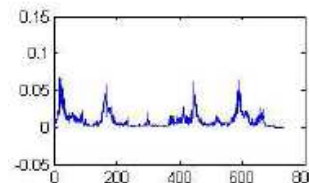
- Any transcription algorithm could be used
- Here a simple drum detection is used:

$$D^i(n) = \sum_{k=1}^8 \sum_{n=0}^{N-1} \left[e_{hk}^i(n) \hat{x}_{hk}(n_0 + n) + e_{rk}^i(n) \hat{x}_{rk}(n_0 + n) \right]^2$$

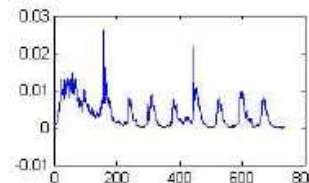
Bass drum



snare drum



cymbal



Drum Track separation

⇒ 3. Remasking:

$$\alpha_{hk}(n) = \max_i (\mathbb{I}^i \star e_{hk}^i)(n)$$

$$\alpha_{rk}(n) = \max_i (\mathbb{I}^i \star e_{rk}^i)(n)$$

- Where $\mathbb{I}^i(n)$ is equal to 1 if n is the onset of a note played by the drum instrument i , 0 otherwise.

Drum Track separation

- Separation with Wiener Filtering (extension and adaptation of the work by Beneroya&al2006)
- Principle:
 - ⇒ Considering 2 Gaussians processes s_1 and s_2 of PSD $\sigma_1^2(f)$ and $\sigma_2^2(f)$
 - ⇒ The optimal estimate of s_i can be obtained by filtering the mixture with filter of gain:

$$\frac{\sigma_i^2(f)}{\sigma_1^2(f) + \sigma_2^2(f)}$$

- ⇒ Since audio sources are not stationary and Gaussian, the sources are assumed to be mixtures of stationary Gaussian processes, with slowly time-varying coefficients

$$s_i(n) = \sum_{k \in K_i} a_i(n) b_k(n)$$

Drum Track separation

Separation with Wiener Filtering (2)

- **Estimation consists in:**

- ⇒ Obtaining a time-frequency representation $S_x(l, m)$ of x by means of the STFT - where l is the frequency bin index, and m a frame index.
- ⇒ Decomposing for every time frame t the observed power spectra as a sum of the spectral templates

$$S_x(l, m) \approx \sum_{k \in K_1 \cup K_2} a_k(m) \sigma_k^2(l)$$

- ⇒ Estimating the time-frequency representation of the source s_i

$$S_{s_i}(l, m) = \frac{\sum_{k \in K_i} a_k(m) \sigma_k^2(l)}{\sum_{k \in K_1 \cup K_2} a_k(l) \sigma_k^2(m)}$$

Drum Track separation

Separation with Wiener Filtering

- **Optimisation for Drum track separation**

- ⇒ A) Use of Non-Negative Matrix factorisation (NMF) to obtain 16 elementary PSD from solo drums (training database)
- ⇒ B) Simple adaptation: It consists in extending, during the decomposition step, the set of drum spectral templates with the PSD of the stochastic component of x observed for frame m .
- ⇒ C) Window size switching scheme :
 - small windows efficient for segments with drum onsets
 - Long windows efficient for sustained parts
 - 2 window size ($L=1024$ or 256) as a result of the onset detection module

Drum separation evaluation

- Use of performance metrics (Gribonval&al2003)

- ⇒ s_d is the original drum signal
- ⇒ s_a is the accompaniment signal
- ⇒ ϵ_{artif} is the residual of the projection

$$\hat{s}_d = \langle \hat{s}_d, s_d \rangle s_d + \langle \hat{s}_d, s_a \rangle s_a + \epsilon_{artif}$$

- ⇒ We define Signal to Distorsion Ratio (SDR), Signal to Interference Ratio (SIR) and Signal to Artefact Ratio (SAR)

$$SDR = 10 \log_{10} \frac{\|\langle \hat{s}_d, s_d \rangle s_d\|^2}{\|\langle \hat{s}_d, s_a \rangle s_a + \epsilon_{artif}\|^2}$$

$$SIR = 10 \log_{10} \frac{\|\langle \hat{s}_d, s_d \rangle s_d\|^2}{\|\langle \hat{s}_d, s_a \rangle s_a\|^2}$$

$$SAR = 10 \log_{10} \frac{\|\langle \hat{s}_d, s_d \rangle s_d + \langle \hat{s}_d, s_a \rangle s_a\|^2}{\|\epsilon_{artif}\|^2}$$

Separation results

Method	Accompaniment -6 dB			Accompaniment +0 dB			Accompaniment +6 dB		
	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
Variable gain	3.9	11.2	6.1	1.2	5.2	4.9	-3.5	-1.2	3.7
NMF+SVM	5.2	14.4	6.2	2.2	10.7	3.5	-1.4	6.9	0.2
Spectral modulation	0.7	13.8	1.3	-0.8	8.0	0.9	-3.9	2.1	0.0
Sub-band ICA from stereo signal	5.7	10.0	9.7	0.1	4.9	5.9	-6.3	-2.2	2.6
Noise subspace projection	8.3	10.2	14.5	3.0	4.3	11.5	-2.7	-1.6	8.9
TFS masking	7.6	14.0	9.6	3.4	6.8	7.7	-2.4	-0.6	6.3
Score-informed TFS masking	7.5	15.9	8.7	4.6	10.0	7.1	0.4	4.1	4.7
Wiener filter	8.6	10.4	14.8	3.1	9.4	5.1	-0.4	4.8	2.9
Wiener filter, enhanced	10.1	15.7	12.2	5.5	10.7	8.0	0.2	5.1	3.9

- **DEMO (Wiener filter, enhanced)**
- **Other demos at <http://tsi.enst.fr/~gillet/ENST-drums/separation/>**

Conclusion

- We argue that audio transcription and audio source separation can benefit from each other and should be done in parallel
- Impact of source separation on the robustness of « traditional » audio features is however not well known
 - ⇒ Fusion and feature selection however proved to be useful to cope with this lack of knowledge
- Our experiments show that it is easier to separate when the signal is transcribed and vice-versa
 - ⇒ Similarity with estimation problems with hidden variables
 - the set of parameters to estimate (drum transcription) and latent variables (separated signal or model of drum instrument) are difficult to optimize jointly...
 - ... but easy with respect to each other
 - ⇒ Opens the path for future iterative scheme (transcription/separation) until convergence..

A few links at ENST ...

- www.enst.fr/~grichard/
- www.enst.fr/~rbadeau/
- www.enst.fr/~gillet/
- www.enst.fr/~malonso/

And a few papers....

- **Rythm extraction**

- ⇒ M. Alonso, R. Badeau, B. David et G. Richard, *Musical tempo estimation using noise subspace projections*, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '03), New Paltz, New York, 19-22 octobre 2003,
- ⇒ M. Alonso, G. Richard, B. David, "Efficient beat-tracking system based on harmonic+noise decomposition," submitted to "EURASIP Journal on Advances in Signal Processing, Vol. 2007, article 82795, 2007.

- **Noise/signal subspace tracking**

- ⇒ R. Badeau, G. Richard et B. David, *Sliding window adaptive SVD algorithms*, IEEE Transactions on Signal Processing, January 2004.
- ⇒ R. Badeau, B. David and G. Richard, "A new perturbation analysis for signal enumeration in rotational invariance techniques", IEEE Transactions on Signal Processing, 2006
- ⇒ R. Badeau, B. David and G. Richard, "Fast Approximated Power Iteration Subspace Tracking", IEEE Transactions on Signal Processing, vol. 53, n°8, August 2005.

- **Percussive signals transcription**

- ⇒ O. Gillet et G. Richard , « *Extraction and Remixing of Drum tracks from polyphonic music signals* », IEEE-WASPAA'05, New Paltz, NY, 2005
- ⇒ O. Gillet et G. Richard , « Drum loops retrieval from spoken queries », Journal of Intelligent Information Systems, 24:2/3, pp 159-177, Springer Science, 2005

- **Musical instruments recognition / Sparse representation**

- ⇒ S. Essid, G. Richard and B. David, "Musical Instrument Recognition by pairwise classification strategies", IEEE Transactions on Speech and Audio, à paraître en 2006
- ⇒ S. Essid, G. Richard, B. David, « *Instrument Recognition in Polyphonic Music Based on Automatic Taxonomies*, in IEEE Trans. on Audio, Speech and Language Processing, vol 14, N°1, Jan. 2006
- ⇒ P. Leveau, E. Vincent, G. Richard and L. Daudet, « Instrument-Specific Harmonic Atoms for Mid-Level Musical Audio Representation » revised version submitted to IEEE Trans on ASSP.

- **Multiple fundamental frequencies estimation**

- ⇒ J. Rosier, Y. Grenier "Unsupervised Classification Techniques for Multipitch Estimation" in Proc. Of 116th Convention of the Audio Engineering Society, Mai 2004