ICML 2015 - Machine Learning for Music Discovery Workshop

Geoffroy Peeters

Frederic Cornu, David Doukhan, Enrico Marchetto, Remi Mignot, Kevin Perros, Lise Regnier

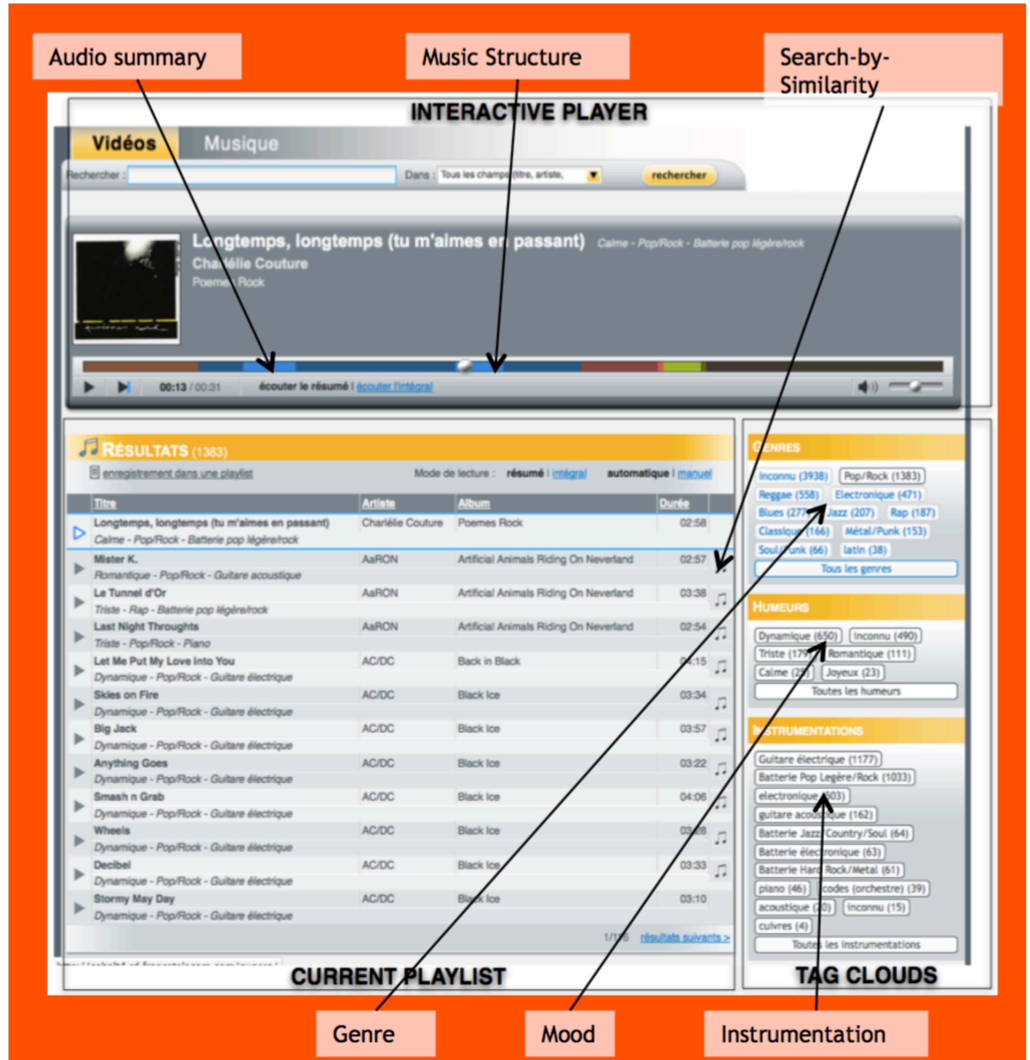
STMS IRCAM CNRS UPMC

This work was partly founded by the French government Programme Investissements d'Avenir (PIA) through the Bee Music Project, and by the European Commission through the SKAT-VG (618067) Project.

WHEN AUDIO FEATURES REACH MACHINE LEARNING

1. Machine Learning for Music Recommendation

Music recommandation



G. Peeters, F. Cornu, D. Tardieu, C. Charbuillet, J. J. Burred, M. Ramona, M. Vian, V. Botherel, J.-B. Rault, and J.-P. Cabanal. A multimedia search and navigation prototype, including music and video-clips. In Proc. of ISMIR (International Society for Music Information Retrieval), Porto, Portugal, October 2012.

2. MIR M.-L. systems over time

Cuidado project (2001-2003)

CUIDADO

Audio Features → Feature Modeling → Feature Selection → Feature Space Transform → Class Model

Audio Features	Feature Modeling	Feature Selection	Feature Space Transform	Class Model
Low-level 1D features Pitch, Centroid, Spread, Roughness	Median, IQR, Vibrato, Tremolo	CFS, Relief-F, IRMFSP	PCA, LDA, Box-Cox	Single-label Bayes, GMM, Decision Tree

- First generic classification system

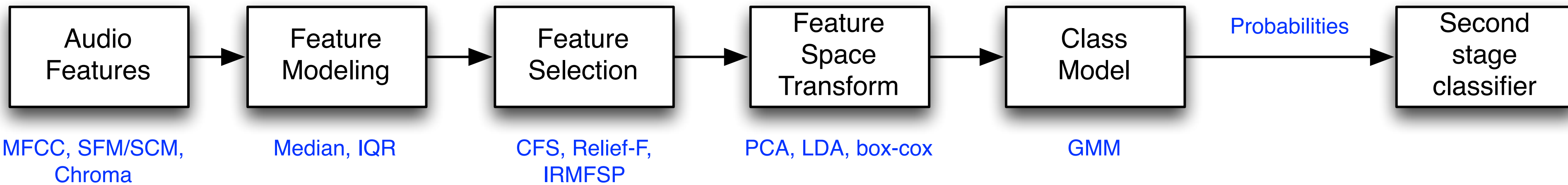
- Target: musical instrument name
- Size: 6.000 audio samples, cross-validation(leave one-dataset out)

- Audio features: 1-D, semantic, coming from perceptual experiment
- Target: clearly defined by the sound source

G. Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. Cuidado project report, Ircam, 2004.

G. Peeters. Automatic classification of large musical instrument databases using hierarchical classifiers with inertia ratio maximization. In Proc. of AES 115th Convention, New York, NY, USA, 2003.

Ecoute project (2006-2008)



- Target: genre, mood classification (single-label)
- Size: 5.000 music extracts (MPO Online/ WMI music catalogue)

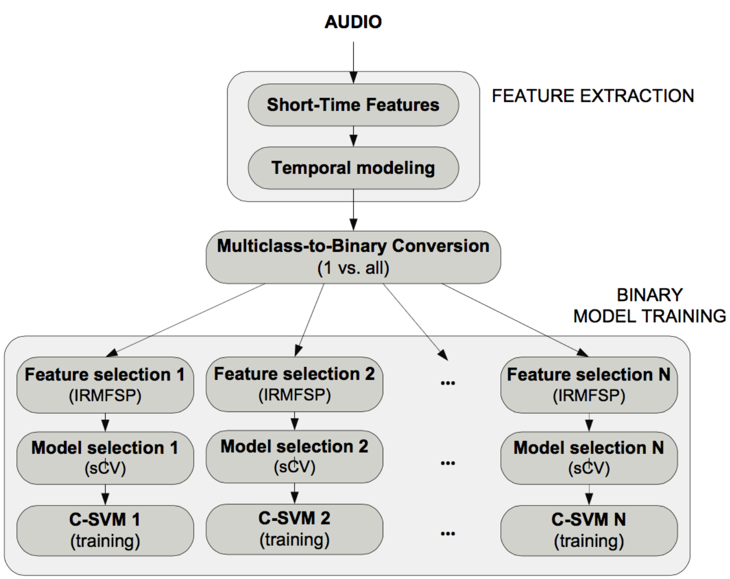
- Audio features: moved to D-dimensional, generic audio-features (no assumptions can be made on audio/music)
- Classifier: second stage classifiers to model probability over time
- Target: somehow hill-defined, needed several attempts

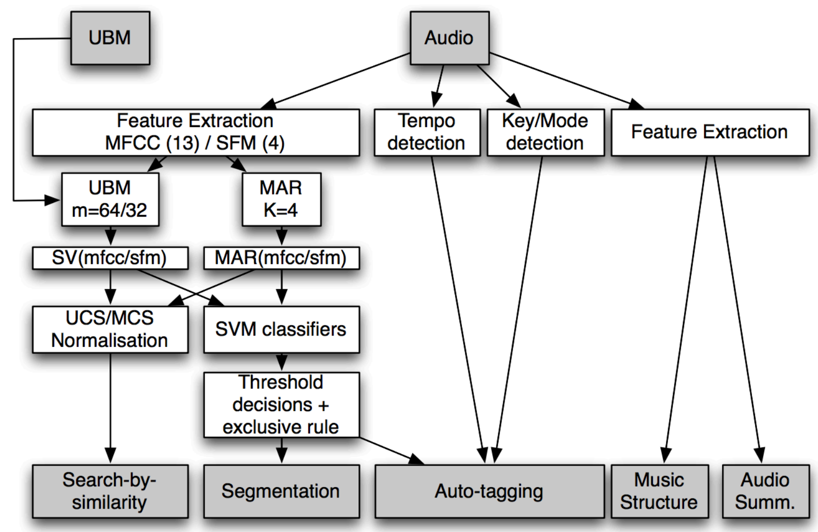
G. Peeters. A generic system for audio indexing: application to speech/ music segmentation and music genre. In Proc. of DAFx (International Conference on Digital Audio Effects), Bordeaux, France, 2007.

Quaero project (2009-2013)

- Target: genre, mood, instru., live/studio, singing, structure
- Size: 30.000 full audio tracks (Orange, INA)

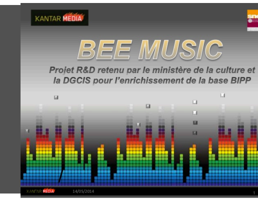
- Audio features: large extent of modeling (UBM Super-Vector, ARM)
- Classifier: Binarization, Multi-Label, Discriminant Classifier (SVM), Threshold Learning



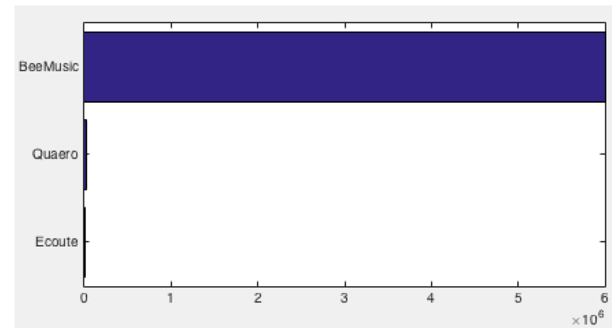


J.-J. Burred and G. Peeters. An adaptive system for music classification and tagging. In Proc. of LSAS (International Workshop on Learning the Semantics of Audio Signals), Graz, Austria, 2009.
C. Charbuillet, D. Tardieu, and G. Peeters. Gmm supervector for content based music similarity. In Proc. of DAFx (International Conference on Digital Audio Effects), pages 425–428, Paris, France, September 2011.

BeeMusic project

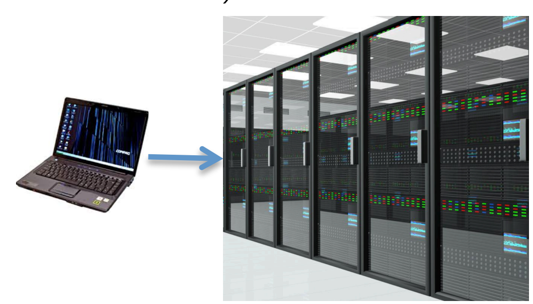


- Target:
  - Labels: Genre, Mood
  - Value:s Valence/Arousal
  - Audio Identification
- Size: 4.000.000 Tracks !!!



- Labels: real labels (SNEP) = noisy, high unbalancing

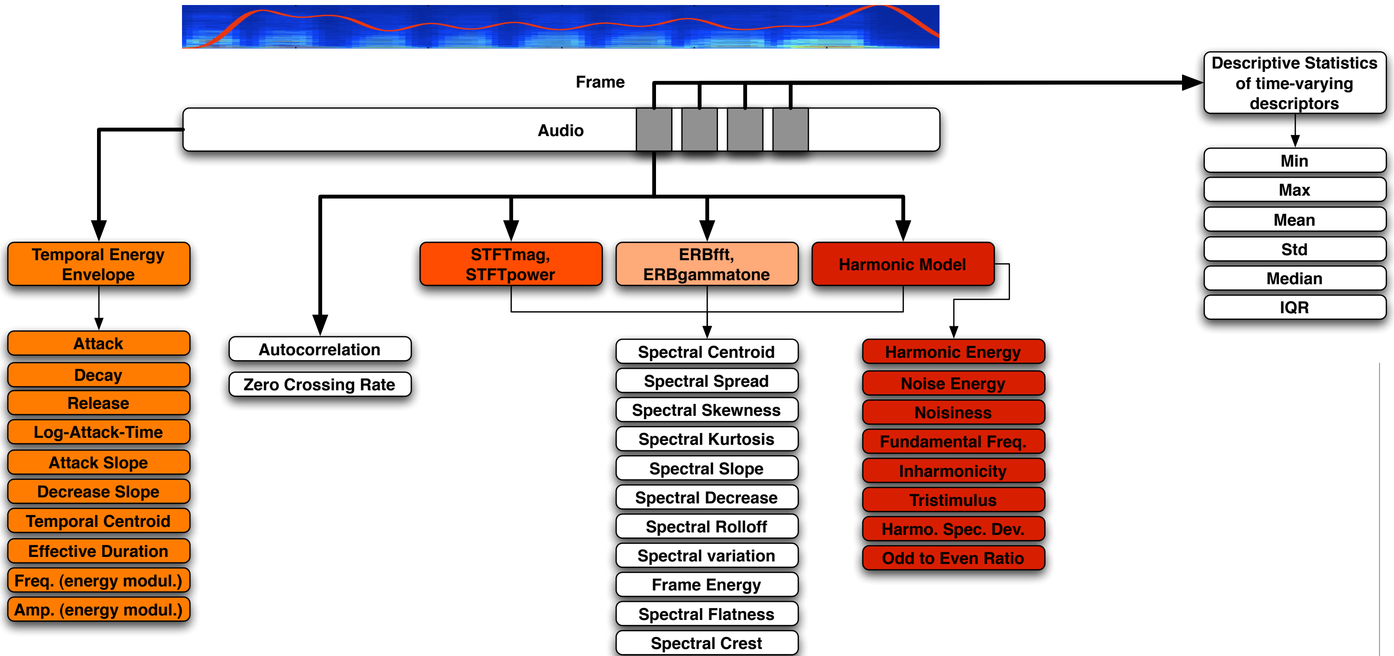
- Developing a system
  - Need to take into account
    - Size of the data: NbDim x NbFiles (4.000.000 * SuperVector>1000)
    - Data transfer
  - A lot of time spent on
    - Data manipulation
    - Cluster configuration
    - Map-Reducing algorithms

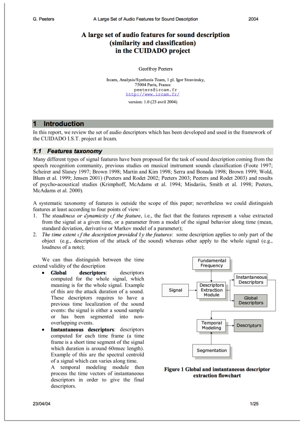


Evolution of Feature Design

Audio Feature Design
Manually Designed Audio Feature

- Inspired by speech processing, perceptual studies, studies on music instruments
- Require specific content
- Generic audio features: MFCC, SFM, Chroma

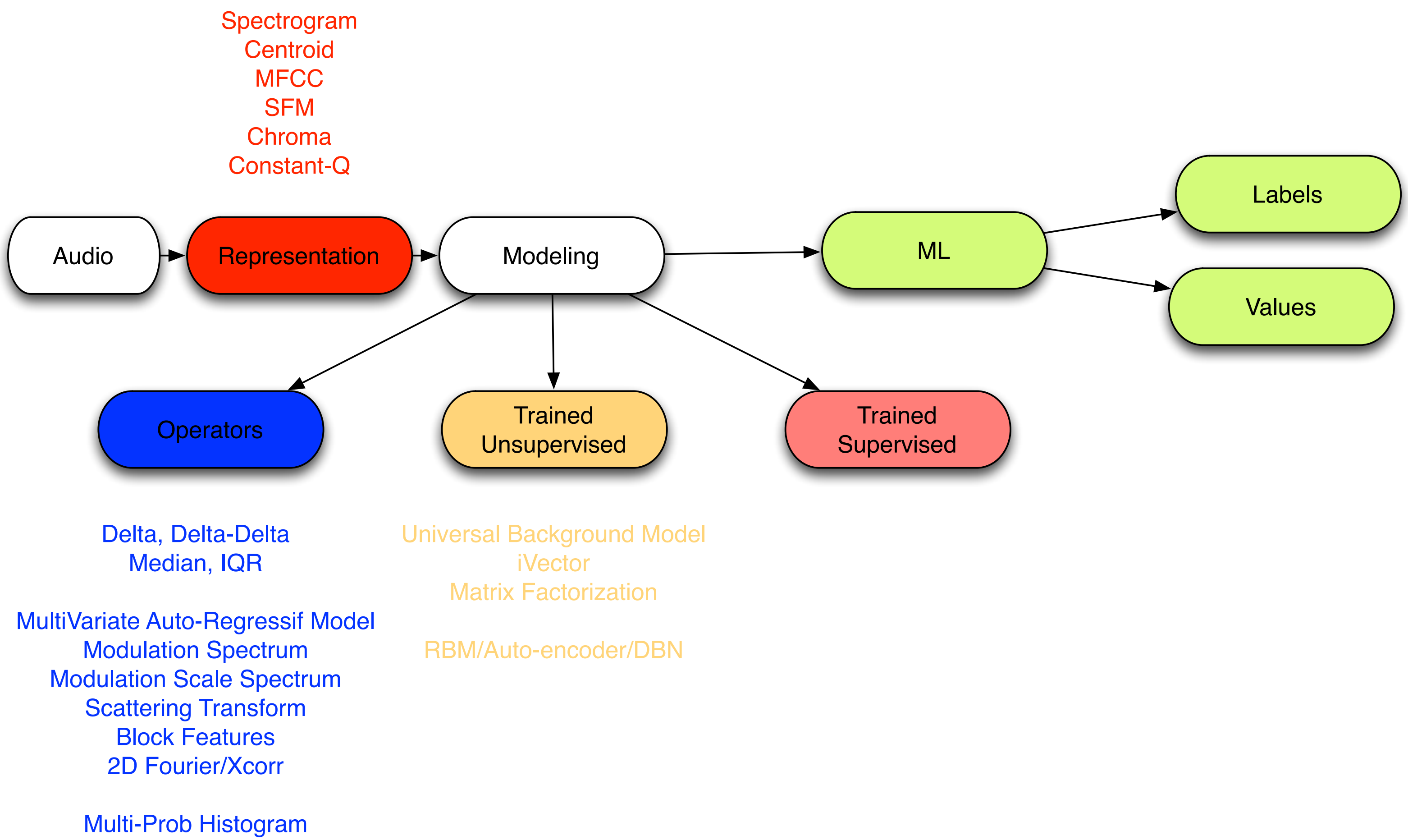




G. Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. Cuidado project report, Ircam, 2004.

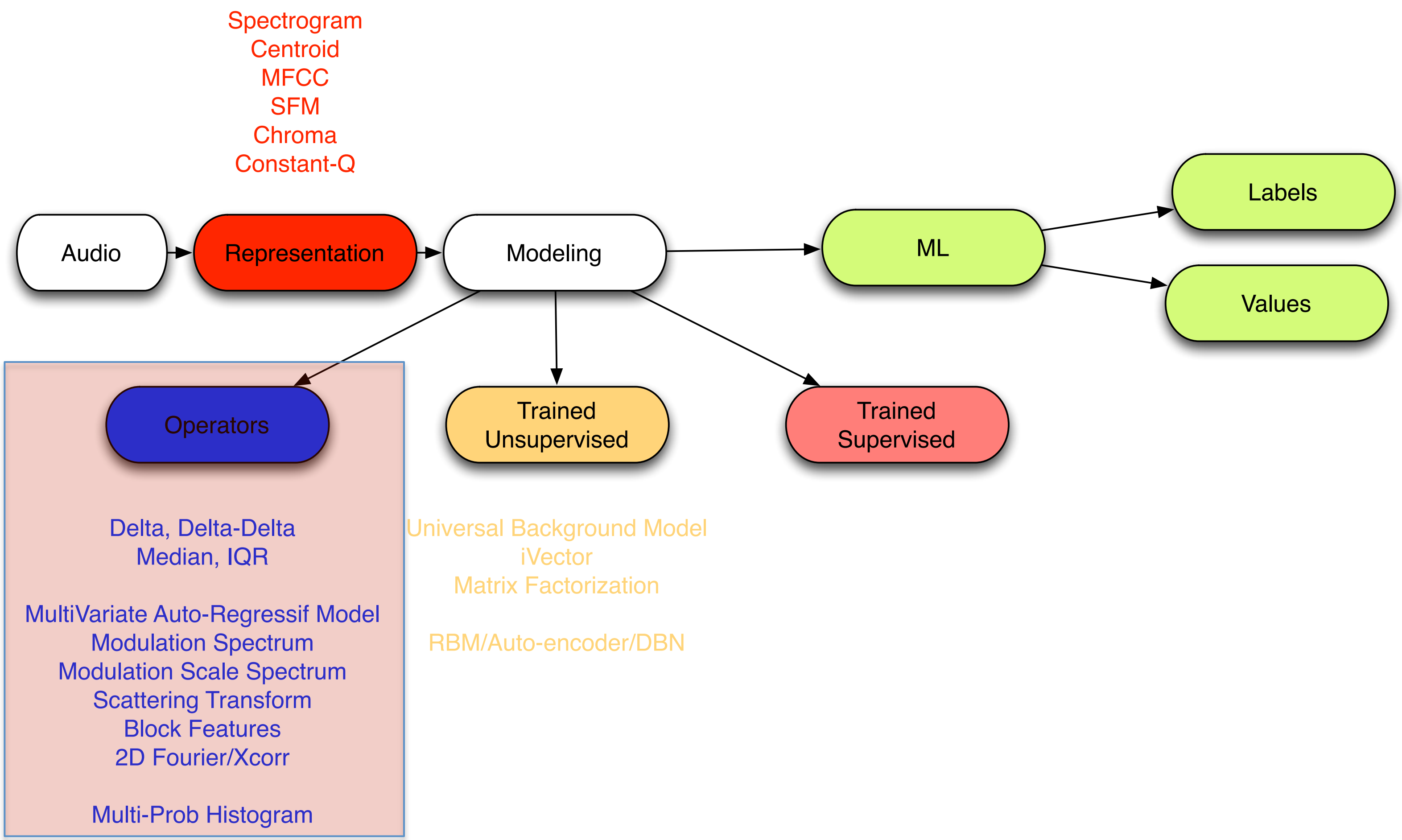
Audio Feature Design
Modelling features (evolution)

Much of the music phenomena is over time



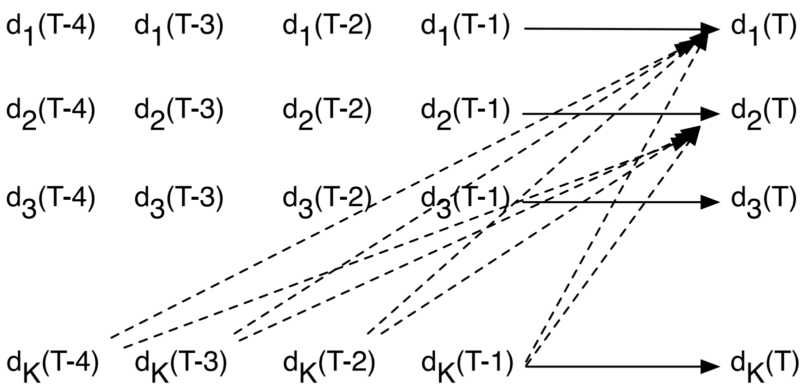
Audio Feature Design
Modelling features (evolution)

Modelling features behaviour over time



Audio Feature Design
Modelling features (evolution)

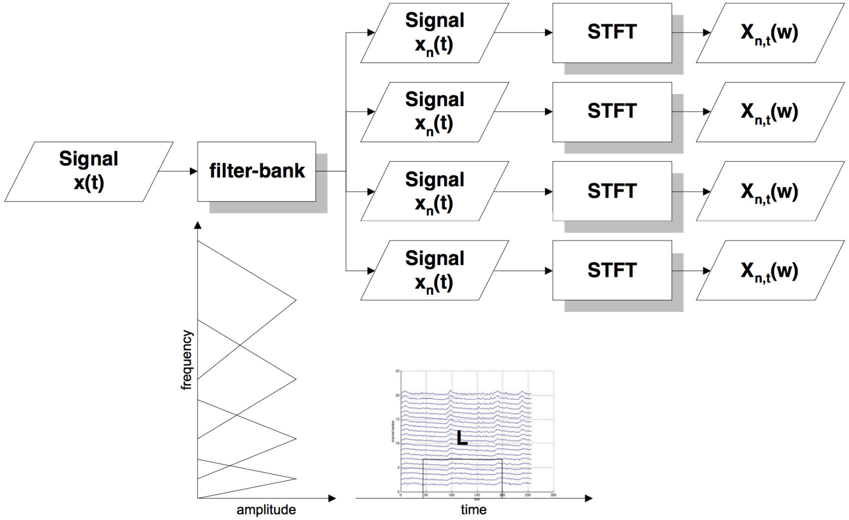
- Multi-Variate Auto-Regressive Model
  - Modelling the time evolution of the audio-features using an AR model, and their joint dependence



F. Bimbot, L. Mathan, A. De Lima, and G. Chollet. Standard and target driven ar-vector models for speech analysis and speaker recognition. In Proc. of IEEE ICASSP (International Conference on Acoustics, Speech, and Signal Processing), volume 2, pages 5–8, San Francisco, California, USA, 1992.

Audio Feature Design
Modelling features (evolution)

- Modulation Spectrum
  - Modelling jointly the time and frequency evolution using Fourier Transform
  - Shift-Invariant
  - Audio Identification -> Music Similarity -> Music Structure



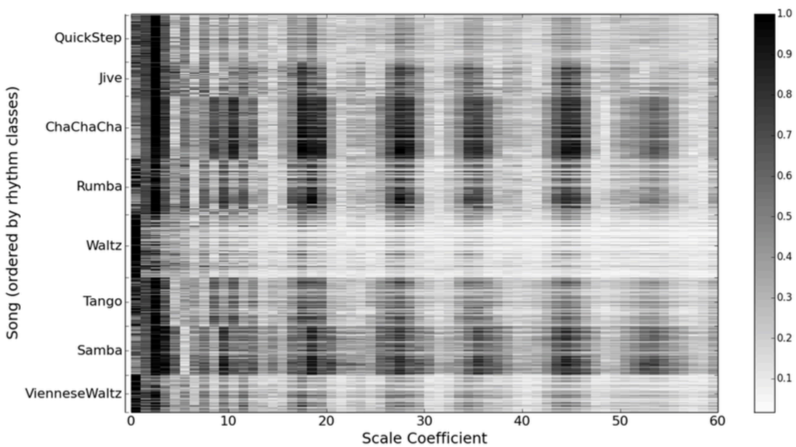
$$x(\omega,\tau)=\frac{1}{\sqrt{2\pi}}\int_t x(t)h(\tau-t)e^{-j\omega t}dt$$

$$X(\omega,\Omega)=\frac{1}{\sqrt{2\pi}}\int_\tau |x(\omega,\tau)|e^{-j\Omega\tau}d\tau$$

X. Rodet, L. Worms, and G. Peeters. Method for characterizing a sound signal. US 2005/0163325 A1 / EP 1459214 A1 / JP 2005-513576 A / WO 2003/056455, 2003.

Audio Feature Design
Modelling features (evolution)

- Modulation Scale Spectrum
  - Modelling jointly the time and frequency evolution using Scale Transform
  - Shift-Invariant and Tempo-Invariant



Method	Exp. 1			Exp. 2		
	Accuracy	C	K	Accuracy.	C	K
Jensen	-	-	-	48.4 %		1
Holzapfel	86.9 %	40	5	-	-	-
Holzapfel (re-implemented)	87.82 %	40	11	66.48 %	20	5
Peeters	87.96 %	-	-	-	-	-
Modulation Scale Transform	**93.12 %**	**60**	**5**	**75.52 %**	**20**	**5**

$$x(\omega, \tau) = \frac{1}{\sqrt{2\pi}} \int_t x(t) h(\tau - t) e^{-j\omega t} dt$$

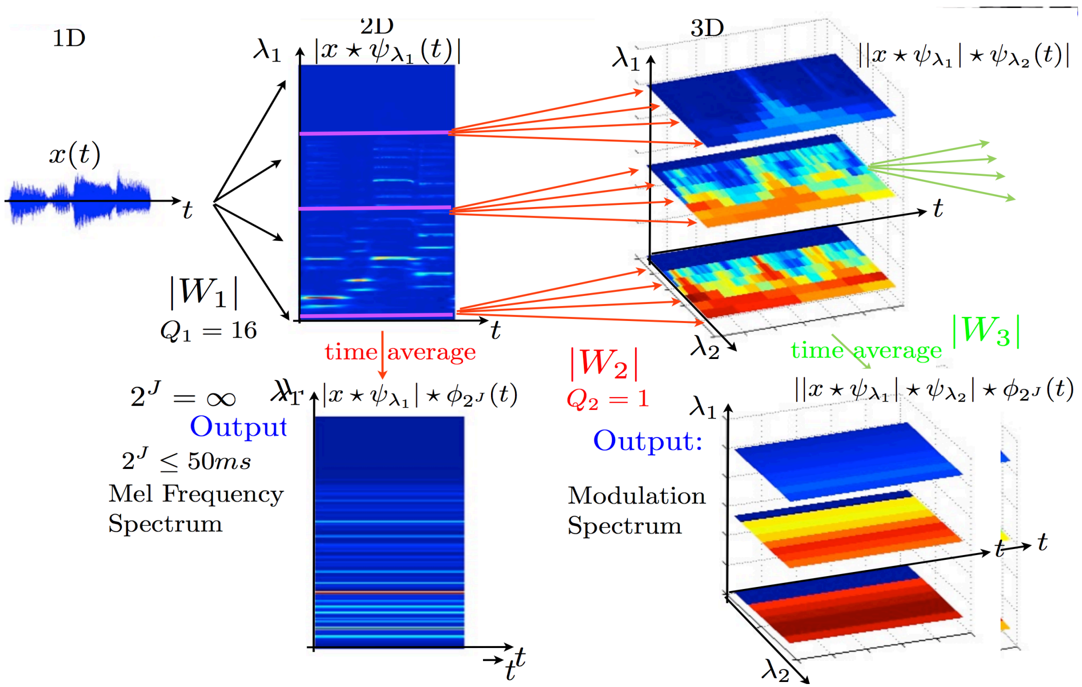
$$D(\omega, c) = \frac{1}{\sqrt{2\pi}} \int |x(\omega, e^{\tau})| e^{\frac{1}{2}t} e^{-jc\tau} d\tau$$

Holzapfel and Y. Stylianou. A scale tranform based method for rhythmic similarity of music. In Proc. of IEEE ICASSP (International Conference on Acoustics, Speech, and Signal Processing), Taipei, Taiwan, 2009.

U. Marchand and G. Peeters. The modulation scale-spectrum and its application to rhythm-content description. In Proc. of DAFx (International Conference on Digital Audio Effects), Erlangen, Germany, 2014.

Audio Feature Design
Modelling features (evolution)

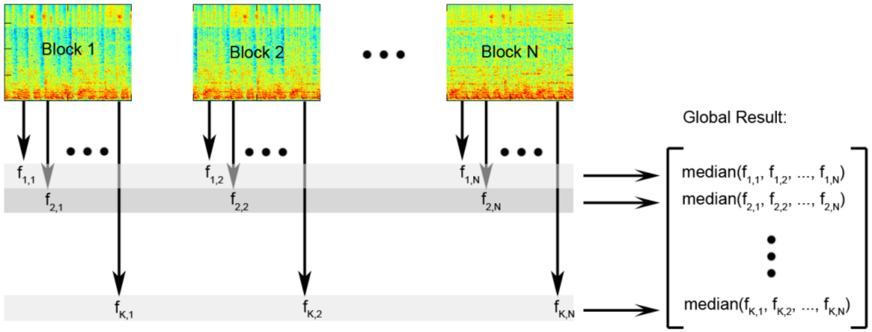
- Scattering network
  - model successively the time and frequency evolution using Wavelet transform, mutliple-layers, marginal

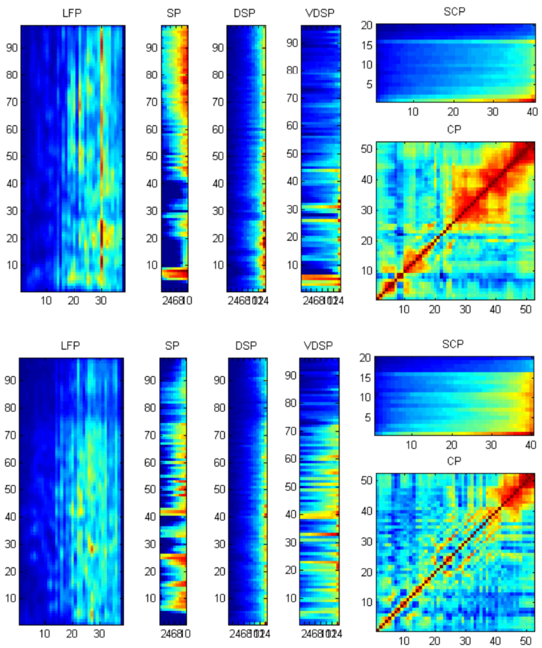


J. Andén and S. Mallat. Multiscale scattering for audio classification. In Proc. of ISMIR (International Society for Music Information Retrieval), pages 657–662, Miami, Florida, USA, 2011.

Audio Feature Design
Modelling features (evolution)

- Block Features
  - Model the variation insight a block using various histogram/ranking statistics
  - Also use some sort of modulation spectrum (onset coefficients)

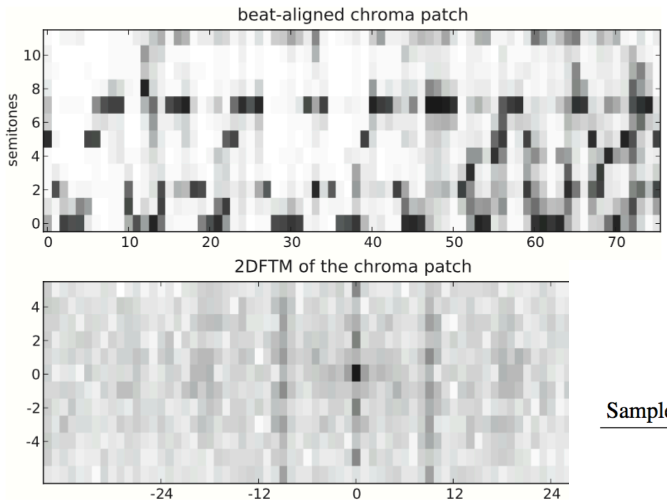


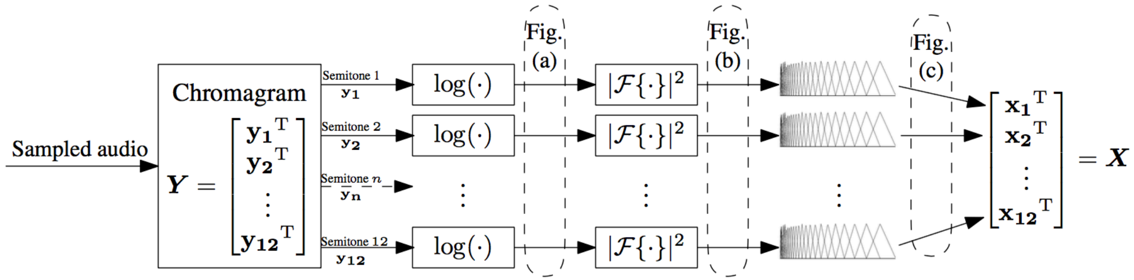


K. Seyerlehner. Content-Based Music Recommender Systems: Beyond simple Frame-Level Audio Simi- larity. PhD thesis, Johannes Kepler Universit ̈at, Linz, Austria, December 2010.

Audio Feature Design
Modelling features (evolution)

- Modelling dependencies between frequency/scale bands
  - 2D-Fourier
  - 2D-Auto-Correlation

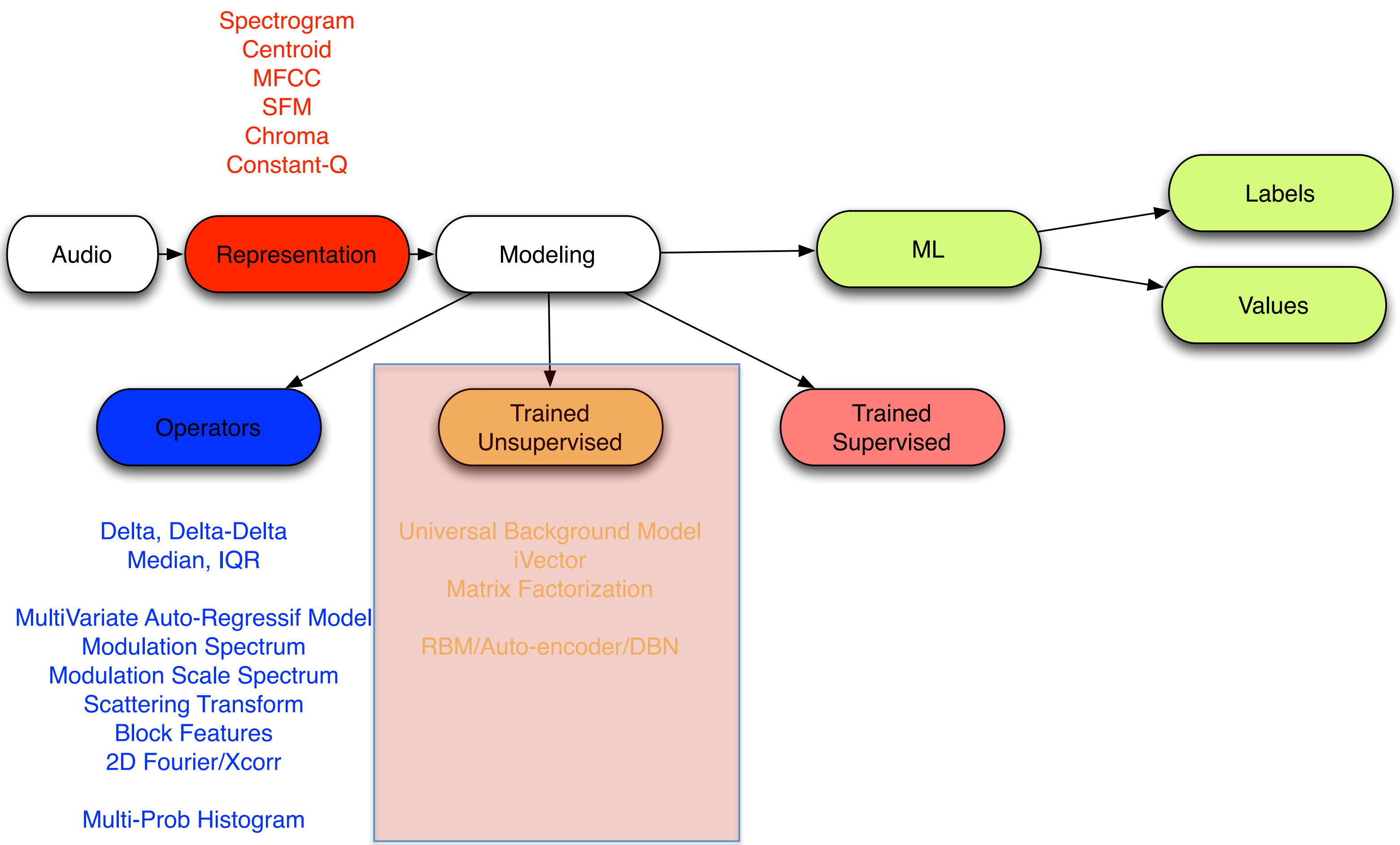




T. Bertin-Mahieux and D. P. Ellis. Large-scale cover song recognition using the 2d fourier transform magnitude. In Proc. of ISMIR (International Society for Music Information Retrieval), Porto, Portugal, 2012.

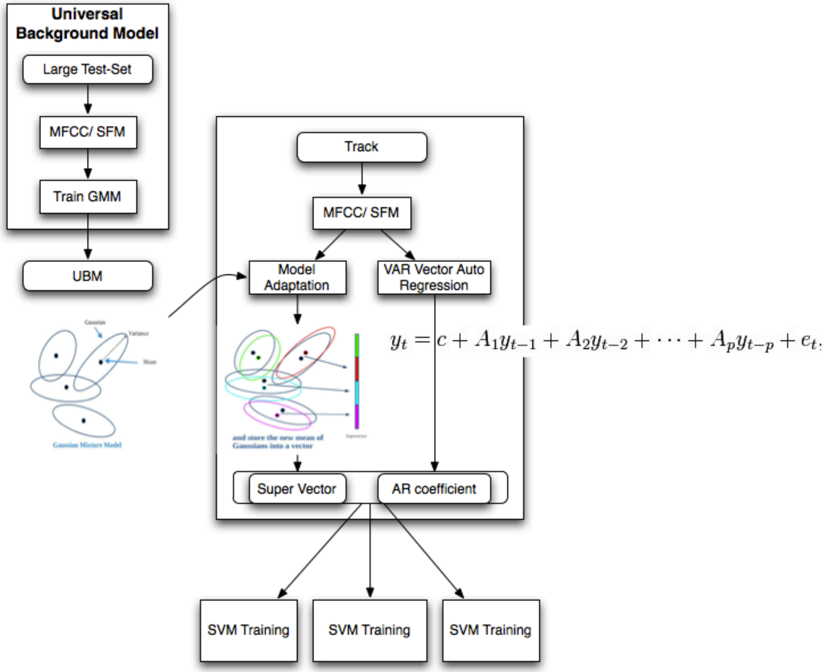
Audio Feature Design
Modelling features (evolution)

Modelling features behaviour over time



Audio Feature Design
Modelling features (evolution)

- Trained (unsupervised) feature representation
  - Universal Background Model/ Super-vector
  - Identity i-Vector $M = m + Tw$



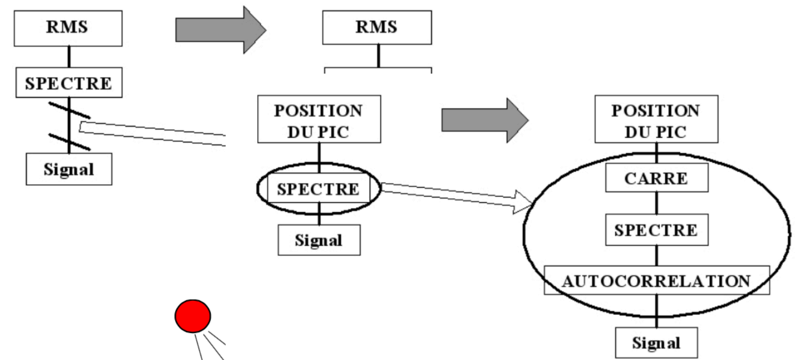
D. Reynolds, T. Quatieri, and R. Dunn. Speaker verification using adapted gaussian mixture models. Digital signal processing, 10(1-3):19–41, 2000.

C. Charbuillet, D. Tardieu, and G. Peeters. Gmm supervector for content based music similarity. In Proc. of DAFx (International Conference on Digital Audio Effects), pages 425–428, Paris, France, September 2011.

N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. Audio, Speech, and Language Processing, IEEE Transactions on, 19(4):788–798, 2011.

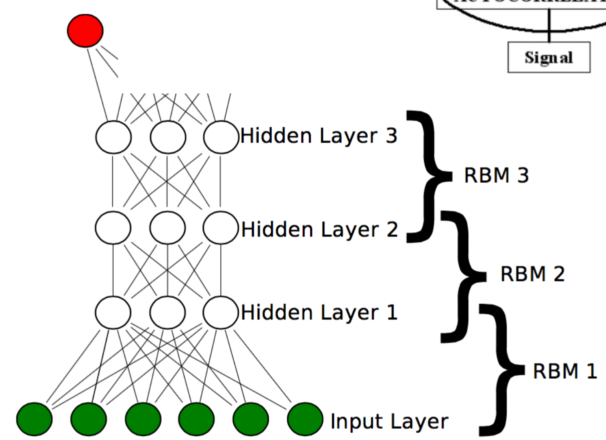
Audio Feature Design
Modelling features (evolution)

- <u>Non-linearly Trained</u> (unsupervised) feature representation
  - EDS system



From Pachet et al. ISMIR 2004

  - RBM/Auto-encoder/DBN



From Hamel et al. ISMIR 2010

P. Hamel and D. Eck. Learning features from music audio with deep belief networks. In Proc. of ISMIR (International Society for Music Information Retrieval), Utrecht, The Netherlands, 2010.

F. Pachet and A. Zils. Automatic extraction of music descriptors from acoustic signals. In Proc. of ISMIR (International Society for Music Information Retrieval), Barcelona (Spain), 2004.

Audio Feature Design
Results

Mirex-2011

Summary Results [top]

Algorithm	Classification Accuracy	Normalised Classification Accuracy
PH2	0.8007	0.8007
WR1	0.7557	0.7557
TCCP4	0.7527	0.7527
SSKS2	0.7496	0.7496

DBN [Hamel, Ismir, 2010]

	Accuracy
MFCCs	0.790
Layer 1	0.800
Layer 2	0.837
Layer 3	0.830
All Layers	**0.843**

SVM + SIM [Marchetto/Peeters, BeeMusic project]

Dataset: GTZAN **Validation: 10-Folds**				
Accuracy, Precision, Recall and F-measure	**Simil.**	**Block**	**MFCC-based**	**MSS**
82.2000 83.0621 82.2000 81.7330	x			
78.3000 79.3717 78.3000 77.9550		x		
80.7000 81.2960 80.7000 80.3091			**x**	
65.3000 65.4959 65.3000 64.1907				x
83.5000 84.3739 83.5000 83.1979	x	x		
84.7000 85.5060 84.7000 84.4027	x		x	
84.2000 84.8188 84.2000 83.7024	x			x
85.3000 86.3197 85.3000 85.1759		x	x	
82.0000 82.9002 82.0000 81.6536		x		x
82.5000 83.4090 82.5000 82.2621			x	x
85.8000 86.6846 85.8000 85.6584	x	x	x	
85.9000 86.8224 85.9000 85.7755		x	x	x
85.9000 86.5272 85.9000 85.6315	x	x		x
85.8000 86.6070 85.8000 85.5263	x		x	x
87.4000 88.0476 87.4000 87.1981	**x**	**x**	**x**	**x**

Scattering Network [Mallat, IEEE, TSP, 2010]

Representations	GTZAN	TIMIT
Δ-MFCC (T = 23 ms)	20.2 ± 5.4	18.5
Δ-MFCC (T = 740 ms)	18.0 ± 4.2	60.5
State of the art (excluding scattering)	9.4 ± 3.1 [8]	16.7 [43]
	$T = 740$ ms	$T = 32$ ms
Time Scat., $l = 1$	19.1 ± 4.5	19.0
Time Scat., $l = 2$	10.7 ± 3.1	17.3
Time Scat., $l = 3$	10.6 ± 2.5	18.1
Time & Freq. Scat., $l = 2$	9.3 ± 2.4	16.6
Adapt Q_1, Time & Freq. Scat., $l = 2$	8.6 ± 2.2	15.9

Discussion

- Continuum between manually and automatically designed audio features
  - EDS or RBM/AE/DBN algorithms rarely start from audio waveforms but rather from higher-level representation inspired by manually designed audio features.
  - Manually designed audio features are rarely used directly, but rather as input to higher-level modelling (such as UBM) which are based on ML algorithms.
- Manual feature design ?
  - un-tractable with the size of the data-set
  - limited to the inspiration of the researcher.
- Automatic feature design ?
  - can help going beyond this and thanks to the increasing availability of computational resources can now be applied to large scale database.
  - It is therefore very welcome.
- Question:
  - apart from its performances,
  - how to get knowledge out of automatically designed features ?

Questions ?