Institut Mines-Télécom

"*MACHINE LISTENING*: AI FOR SOUNDS AND MUSIC"

COLLOQUE IMT - L'INTELLIGENCE ARTIFICIELLE AU COEUR DES MUTATIONS INDUSTRIELLES.
APRIL 4TH, 2019

Gaël RICHARD
Professor, Head of the Image, Data, Signal department

A well established domain for speech …

Speech Recognition

« What a great workshop !»

Speaker Recognition

« Gaël Richard is speaking»

Speech Emotion Recognition

« The person who's speaking is happy but anxious»

Speech Translation

« Quel excellent colloque ! »

« What a great workshop ! »

Goal: to extract « **information** » from the audio signal

Audio Sound capture: *Localisation, Dereverberation, Denoising, …*

Audio Scene recognition: *sound recorded in streets, in subway station, in office,….)*

Audio Event recognition: *Speech, Music, Car noises, birds, …*

Bioacoustics: *Wildlife monitoring, biodiversity, species id…*

Audio Source separation
Demixing music, Singing voice extraction, source localisation,

Music recognition (or Audio ID): *identifying the music recording*

Music Information retrieval
Music transcription, Music similarity, Music genre recognition,
Musical instrument recognition, Music recommendation, Autotagging,

Music Emotion recognition: *Sad vs happy vs dance music, ….*

Music streaming,
music recommendation

Vocal separation, music separation

Music education

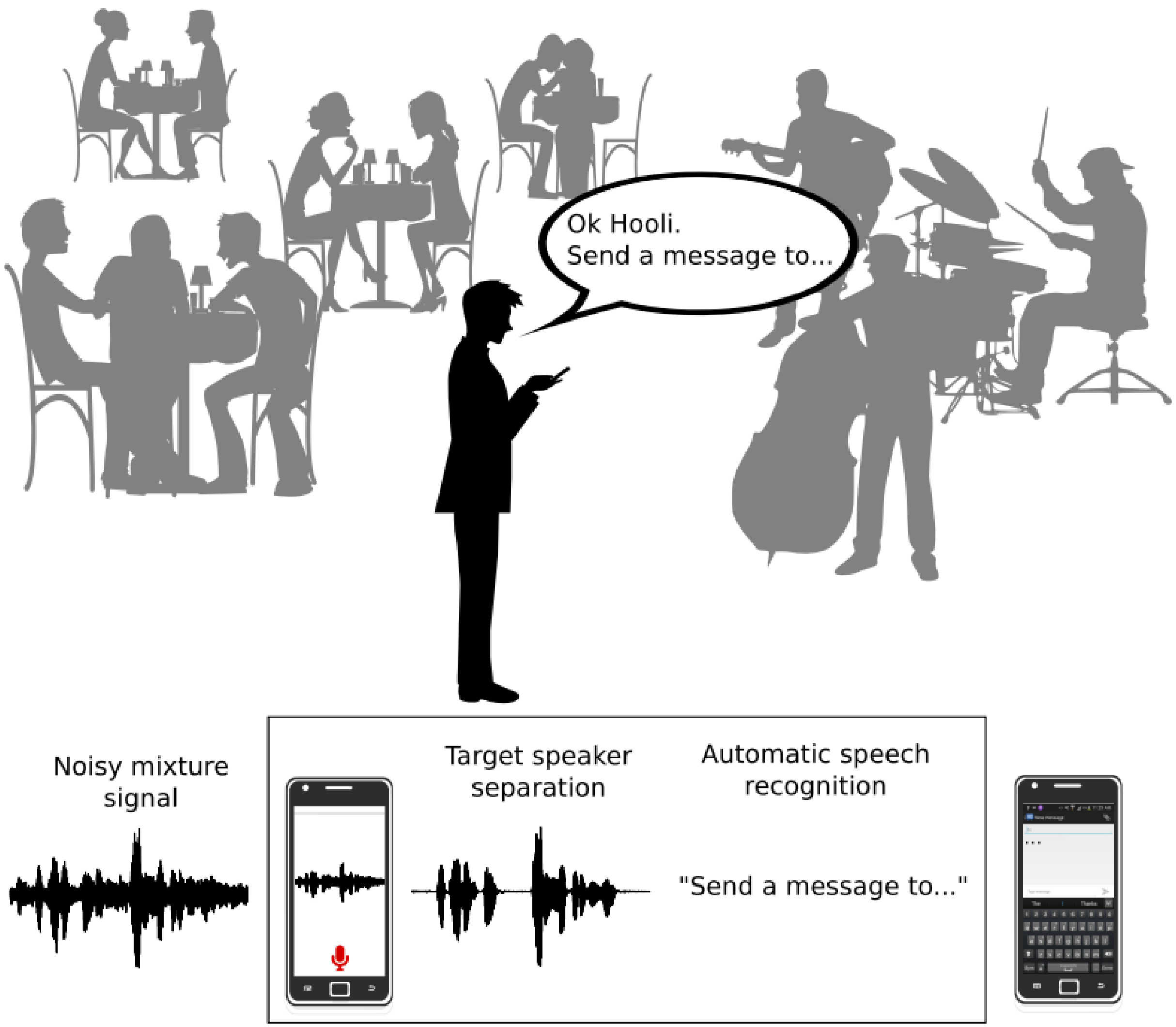
Bioacoustics

Music Identification, Audio Fringerprint

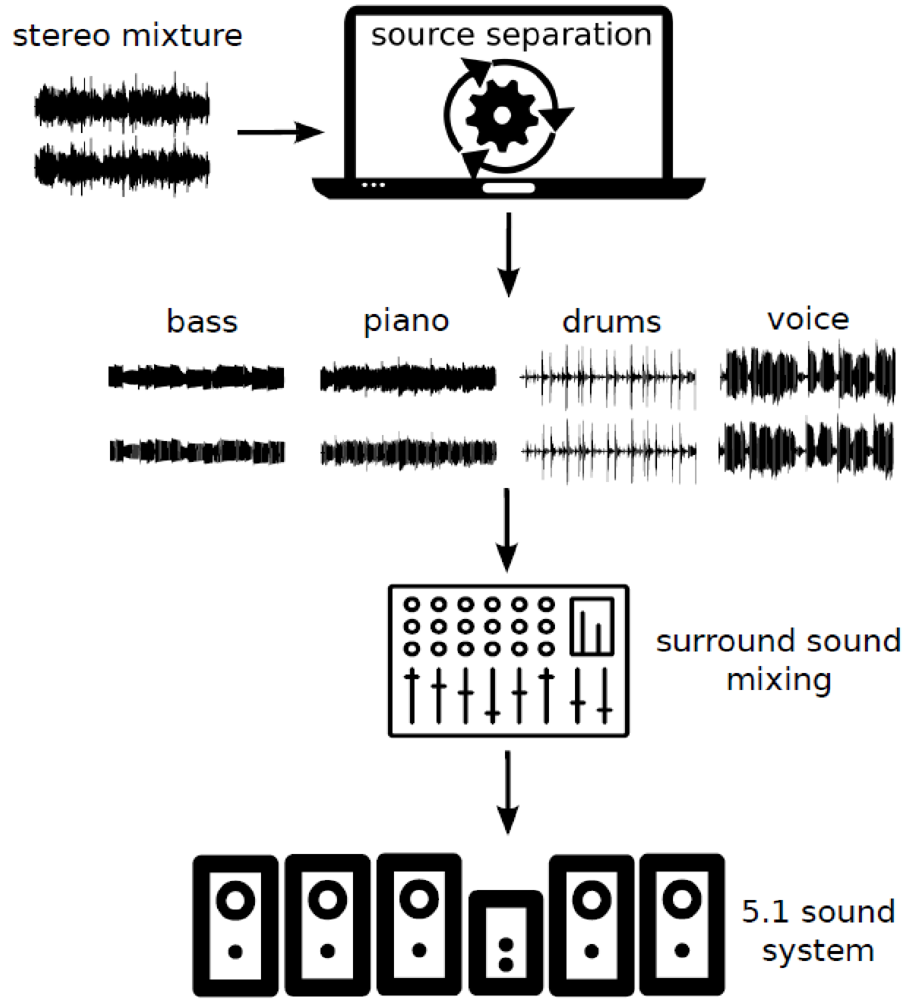
Karaoke, speech to rap conversion

Sound recognition,
smarthomes, smart hearables

Stratégie de marque
musicale, Supervision
musicale (pub.; films)


Ok Hooli.
Send a message to...
Noisy mixture signal
Target speaker separation
Automatic speech recognition
"Send a message to..."

SOURCE SEPARATION FOR REMIXING

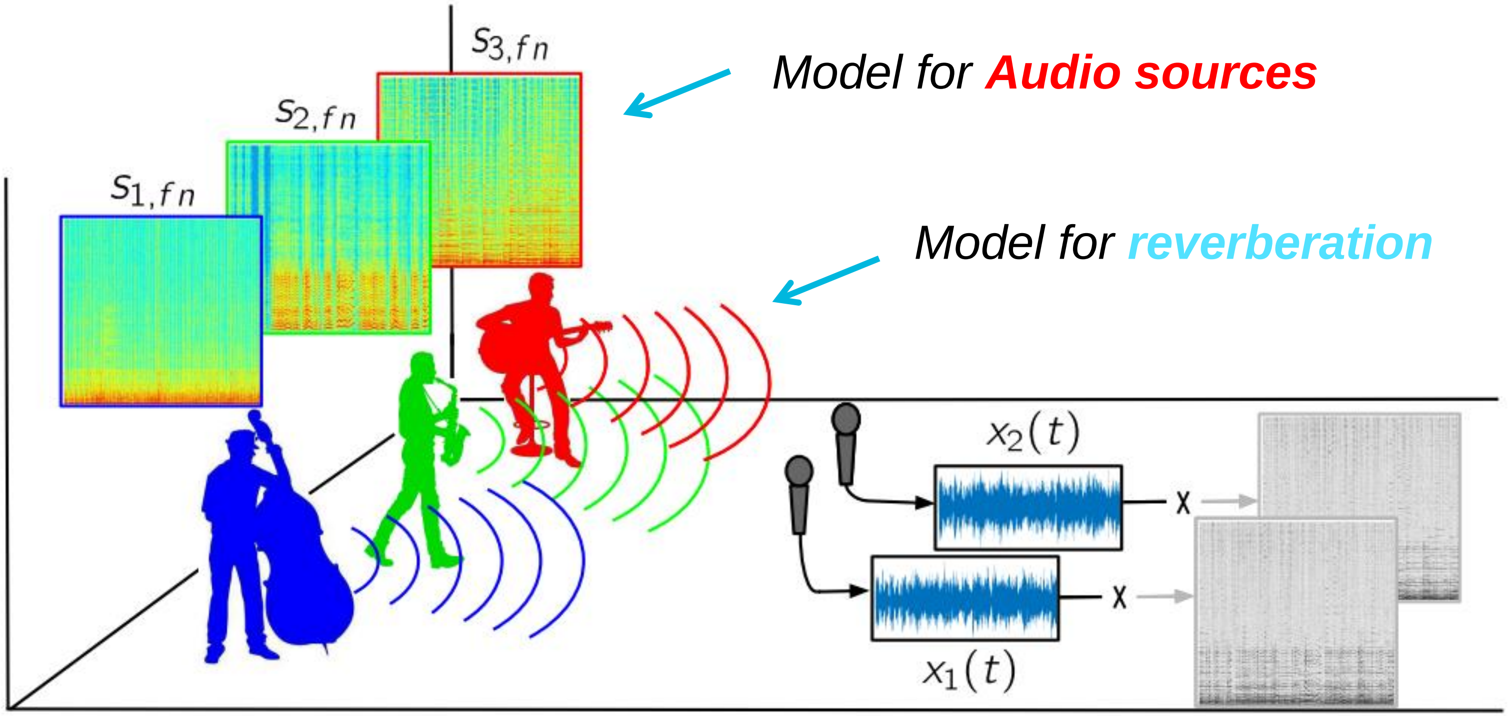


Use case of the ANR project

3DISON

EDISON 3D

- Time-domain mixture representation: $x_i(t) = \sum_{j=1}^{J} [a_{ij} \star s_j](t)$
- Time-frequency source representation: $s_j(t) = \mathcal{T}^{-1}(\{s_{j,fn}\}_{f,n})$



Model for ***Audio sources***

Model for ***reverberation***

Non Negative Matrix factorization

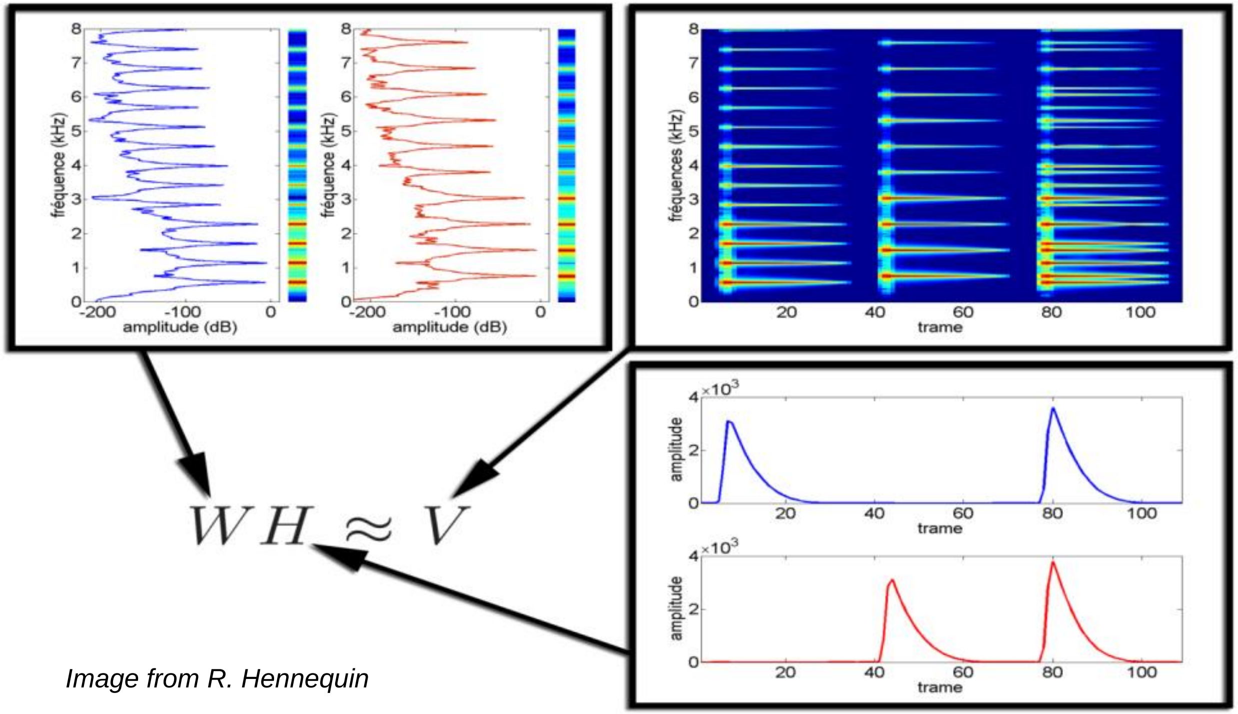
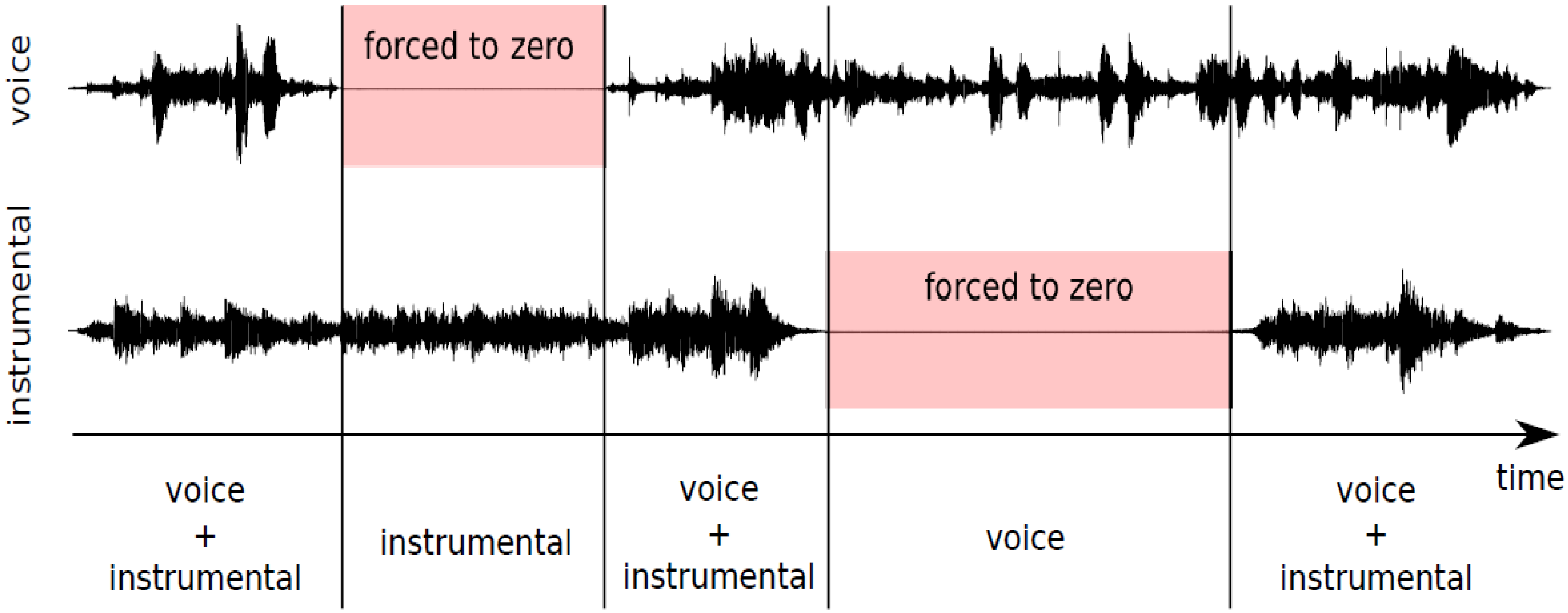


Image from R. Hennequin



S. Leglaive, R. Badeau, G. Richard, "Multichannel Audio Source Separation with Probabilistic Reverberation Priors", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 24, no. 12, December 2016

Simon Leglaive, Roland Badeau, Gaël Richard, Separating Time-Frequency Sources from Time-Domain Convolutive Mixtures Using Non-negative Matrix Factorization. WASPAA , Oct. 2017 New Paltz, US.

Audio ID = find high-level metadata from a music recording

Audio identification

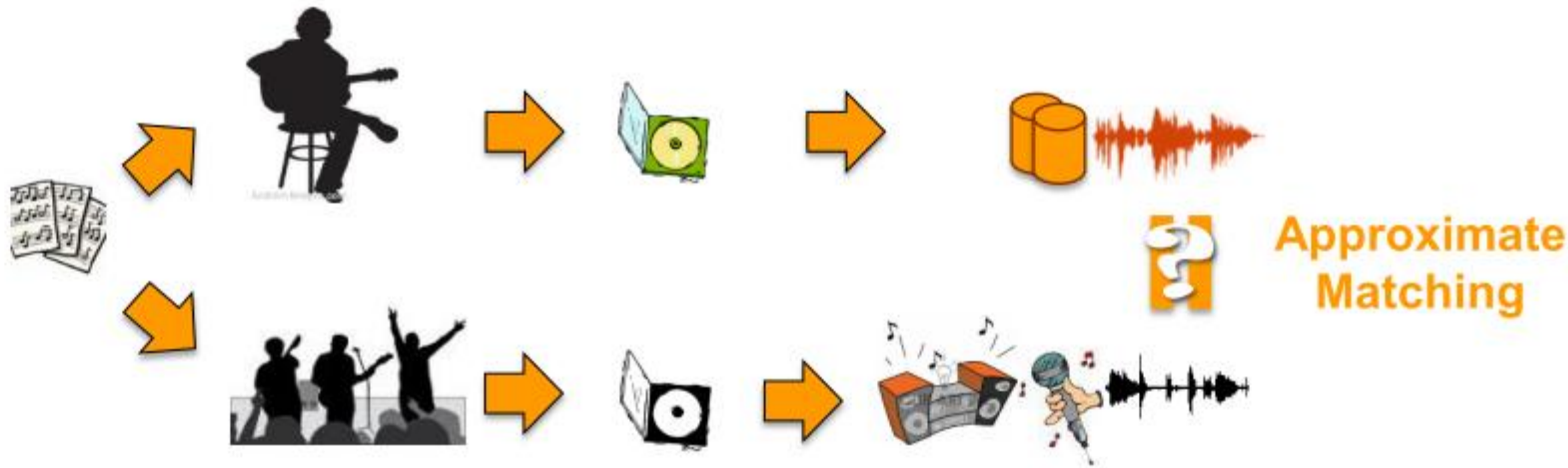
Information of the recording (title, artist, etc.., …)

Challenges:

Efficiency in adverse conditions (distorsion, noises,..)
Scale to “Big data” (bases > millions of titles)
Rapidity / Real time

Product example :

SHAZAM



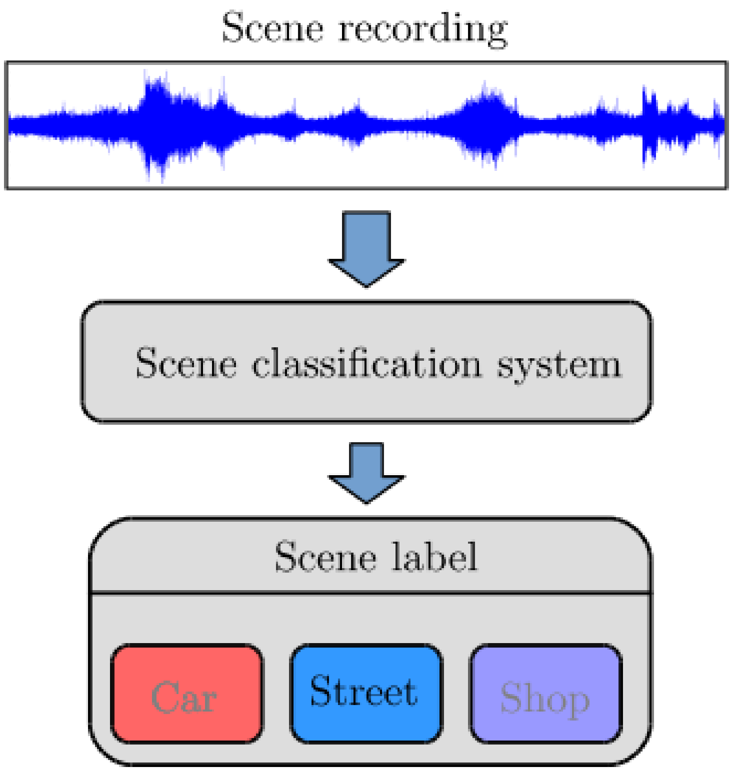
Audio recordings recognition

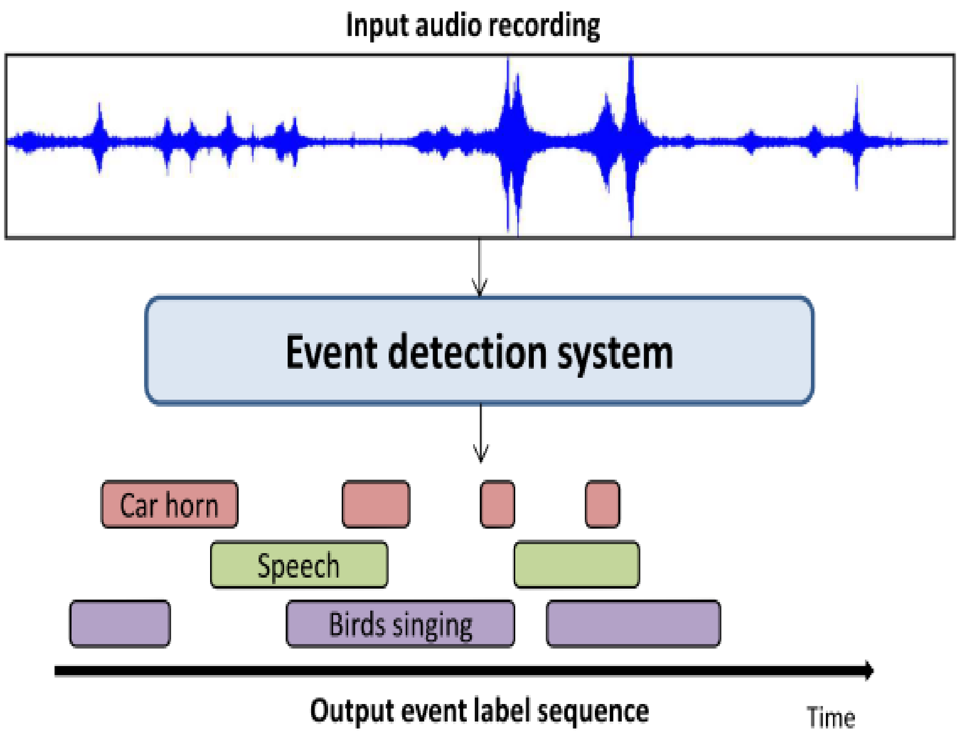
- Identical
- Approximate (live vs studio)

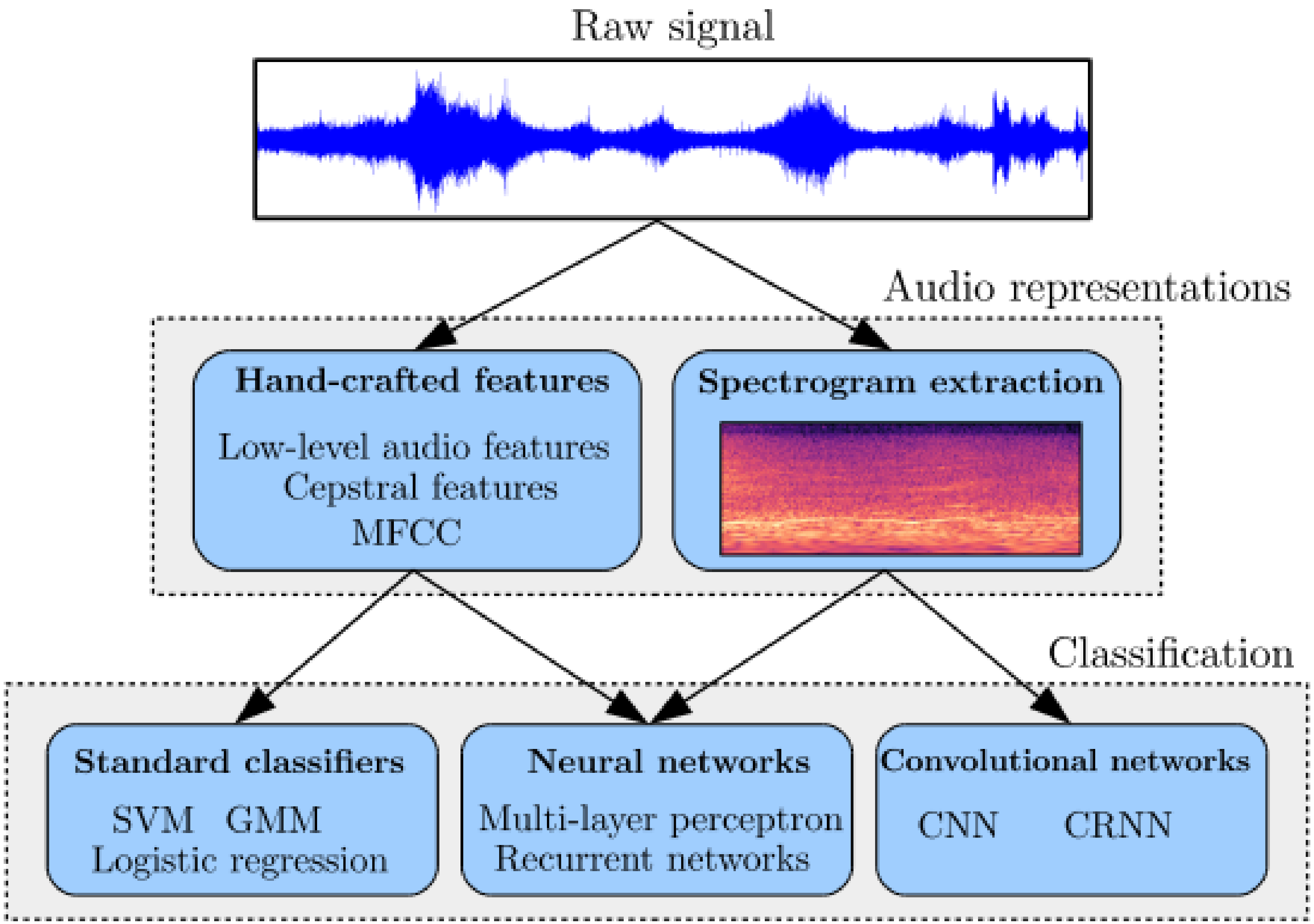
- Real time demonstrator
- For music recommendation, second screen applications, …

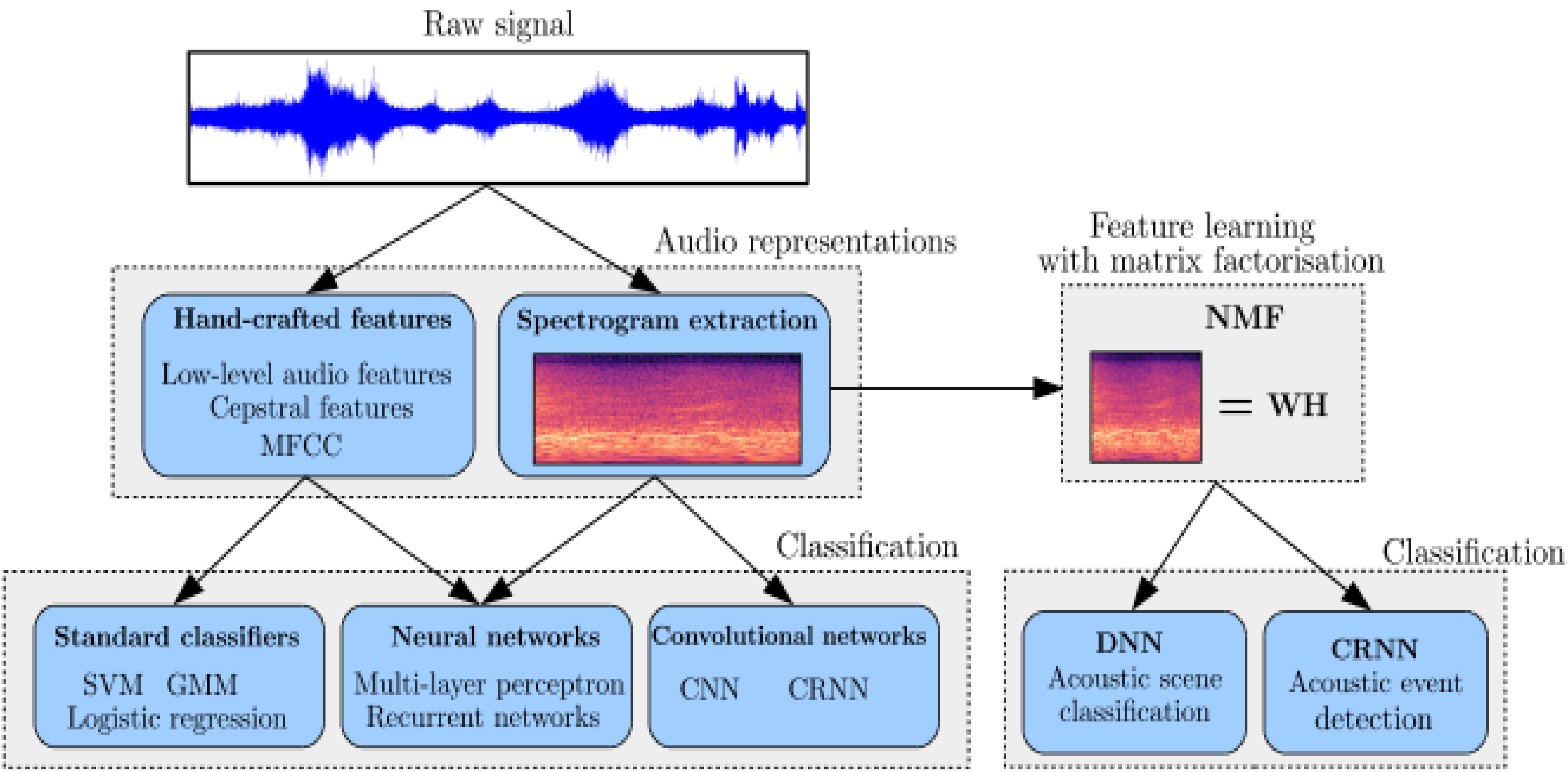
Sébastien Fenet, Yves Grenier, Gaël Richard: An Extended Audio Fingerprint Method with Capabilities for Similar Music Detection. ISMIR 2013: 569-574

SOUND EVENTS AND ACOUSTIC SCENE RECOGNITION



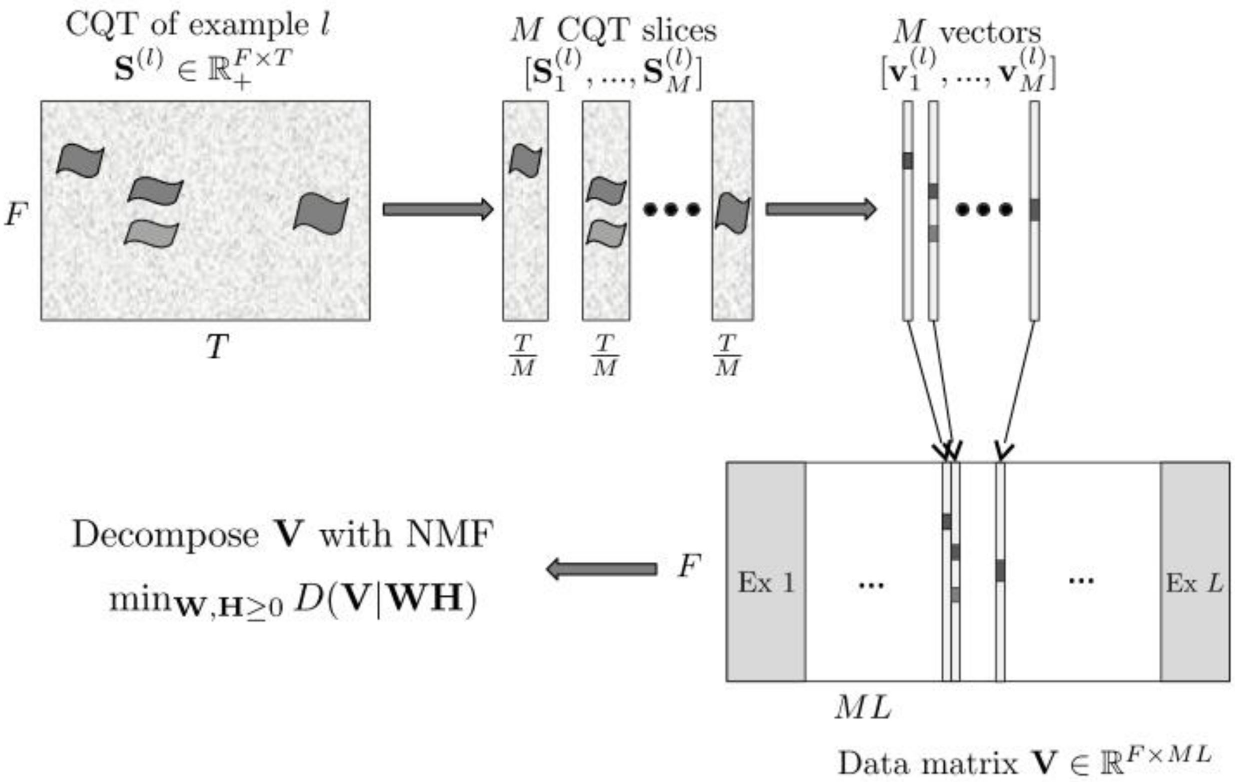



Raw signal
Audio representations
Hand-crafted features
Low-level audio features
Cepstral features
MFCC
Spectrogram extraction
Classification
Standard classifiers
SVM GMM
Logistic regression
Neural networks
Multi-layer perceptron
Recurrent networks
Convolutional networks
CNN CRNN



V. Bisot, R. Serizel, S. Essid, G. Richard, "Feature Learning with Matrix Factorization Applied to Acoustic Scene Classification", IEEE/ACM Transactions on Audio, Speech, and Language Processing, (2017), Special Issue on Sound Scene and Event Analysis.

From time-frequency representations to dictionary learning



Nonnegative matrix factorization

$\min_{\mathbf{W},\mathbf{H}\geq 0} D(\mathbf{V}|\mathbf{WH})$ with $\mathbf{W} \in \mathbb{R}_{+}^{F\times K}$ and $\mathbf{H} \in \mathbb{R}_{+}^{K\times N}$

Dictionary learning with NMF

$$\underbrace{\mathbf{V} \approx \mathbf{W} \times \mathbf{H}}_{\min\limits_{\mathbf{W},\mathbf{H}\geq 0} D(\mathbf{V}|\mathbf{WH})}$$

UNSUPERVISED NMF FOR ACOUSTIC SCENE RECOGNITION

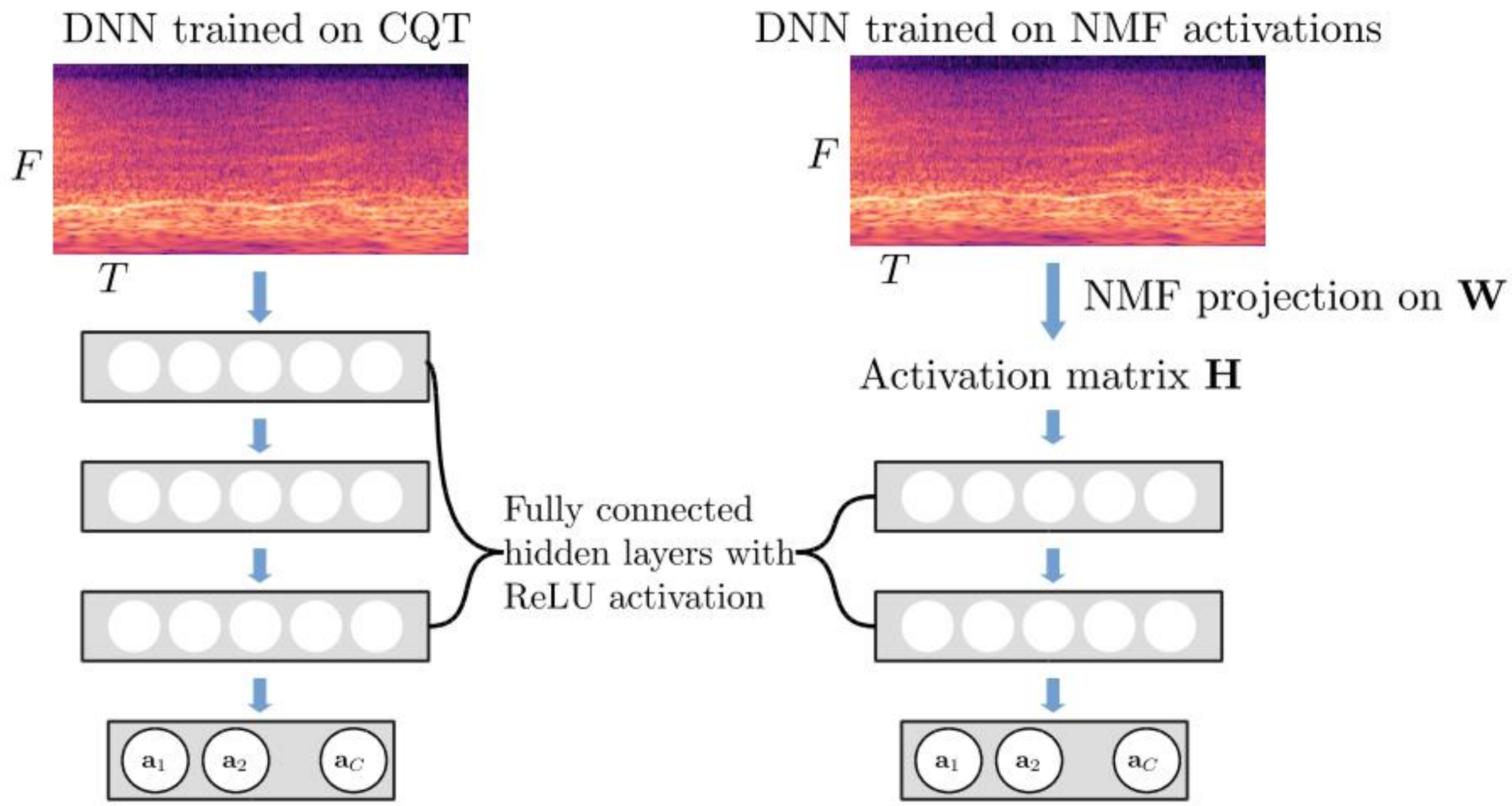
Nonnegative matrix factorization

$\min_{\mathbf{W},\mathbf{H}\geq 0} D(\mathbf{V}|\mathbf{WH})$ with $\mathbf{W} \in \mathbb{R}_+^{F\times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K\times N}$

Feature extraction $\longrightarrow$ project on learned dictionary

$$\underbrace{\boxed{\mathbf{V}} \approx \boxed{\mathbf{W}} \times \boxed{\mathbf{H}}}_{\min\limits_{\mathbf{H}\geq 0} D(\mathbf{V}|\mathbf{WH})}$$

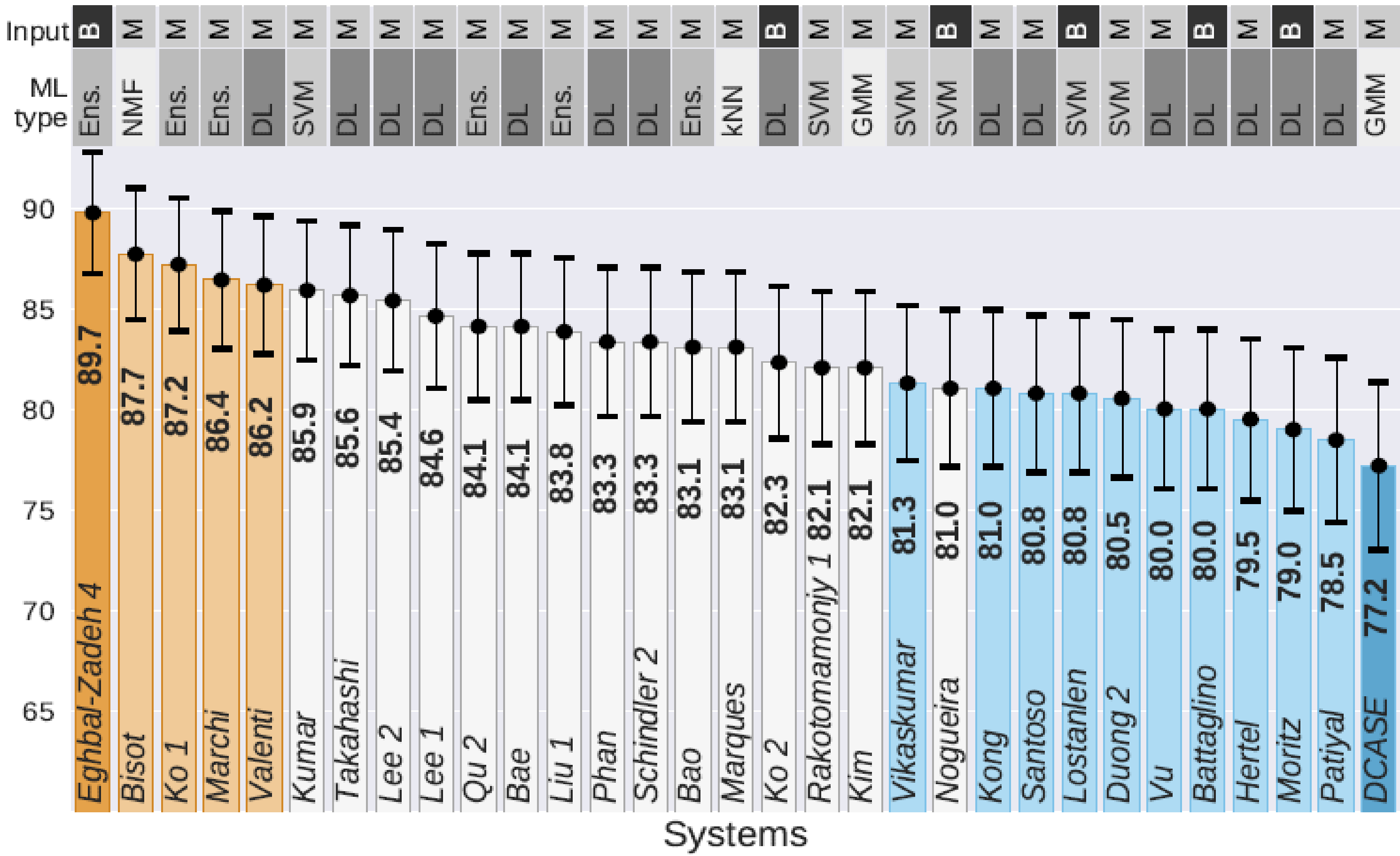
EXAMPLE WITH DNN: ACOUSTIC SCENE RECOGNITION



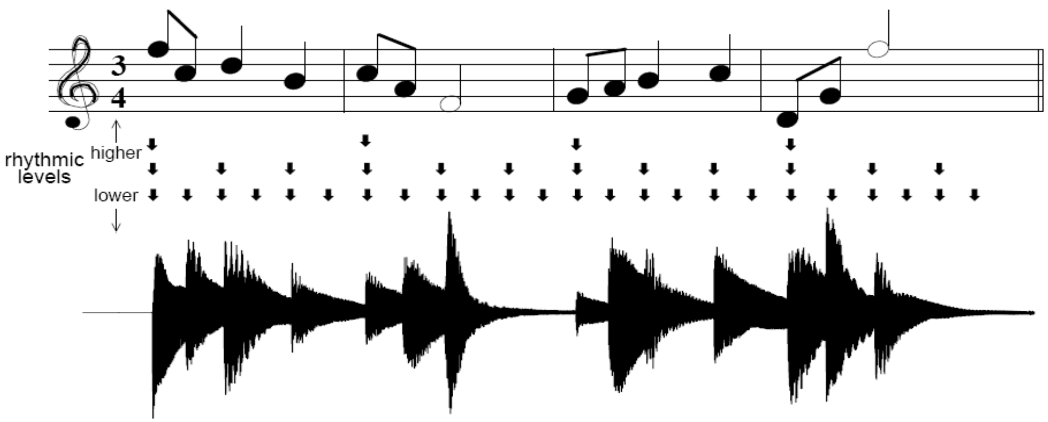
V. Bisot & al., "Feature Learning with Matrix Factorization Applied to Acoustic Scene Classification", IEEE/ACM Transactions on Audio, Speech, and Language Processing, (2017),

V. Bisot & al., Leveraging deep neural networks with nonnegative representations for improved environmental sound classification IEEE International Workshop on Machine Learning for Signal Processing MLSP, Sep 2017, Tok

TYPICAL PERFORMANCES OF ACOUSTIC SCENE RECOGNITION (CHALLENGE DCASE 2016)



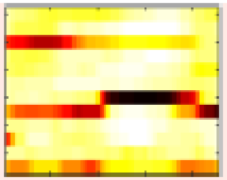
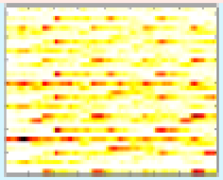
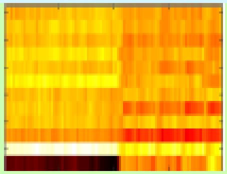
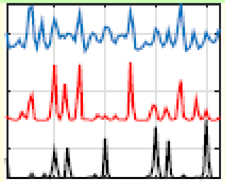
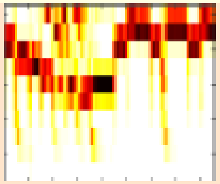
- *A Mesaros & al. Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 challenge IEEE/ACM Transactions on Audio, Speech, and Language Processing 26 (2), 379-393*



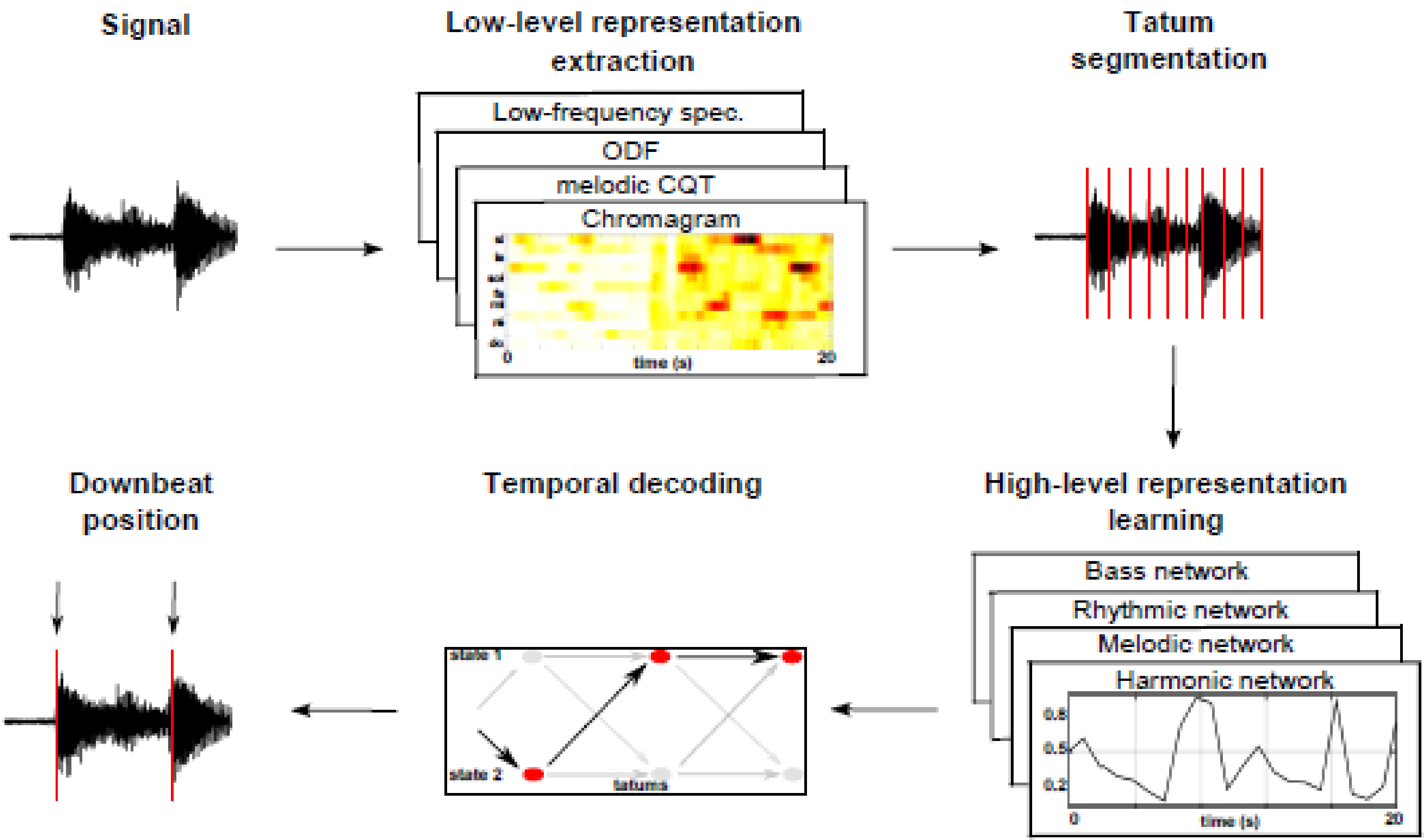
Major topics:

- Music transcription (Multiple F0 estimation, Beat/Downbeat detection, instrument classification, …),
- Music recommendation
- Source separation,
- Multimodal music processing

MIR: AN EXAMPLE WITH DOWNBEAT ESTIMATION
(DURAND & AL. 2017)

Cue	Examples	Input
Harmony	Chord change, Cadence	
Melody	Melodic pattern, pivot notes	
Timbre	Section change, new instrument	
Rhythm	Bar-length rhythm patterns	
Bass content	Bass, Double bass and kick drum highlight downbeats	

MIR: AN EXAMPLE WITH DOWNBEAT ESTIMATION
(DURAND & AL. 2017)



S Durand & al., "Robust Downbeat Tracking Using an Ensemble of Convolutional Networks", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol 25, N°1, 2017

Examples at the output of each network

https://simondurand.github.io/dnn_audio.html

Other audio example

JBB (Tatum)

JBB (Downbeat)

Exemple (Tatum)

Exemple (Downbeat)

Some Current trends:

EEG-driven music processing

With technicolor

Context-driven Style transfer

With technicolor

Conditional audio generation

With Sony CSL

Context-aware music Recommendation

with DEEZER

Text-Informed Lead Vocal Extraction

with Audionamix SEPARATE2CREATE

5 new PhD grants within project «New Frontiers in Music Information Processing (MIP-Frontiers), funded by the European Unions Horizon 2020 research and innovation program under the Marie Skodowska-Curie grant agreement No. 765068.