

# Music Source Separation and its Applications to MIR

Emmanuel Vincent and Nobutaka Ono

INRIA Rennes - Bretagne Atlantique, France  
The University of Tokyo, Japan

Tutorial supported by the VERSAMUS project  
<http://versamus.inria.fr/>

Contributions from Alexey Ozerov, Ngoc Duong, Simon Arberet, Martin Klein-Hennig and Volker Hohmann.



## Part I: General principles of music source separation

- 1 Source separation and music
- 2 Computational auditory scene analysis
- 3 Probabilistic linear modeling
- 4 Probabilistic variance modeling
- 5 Summary and future challenges

## Audio source separation

Many sound scenes are **mixtures** of several concurrent sound sources.

When facing such scenes, humans are able to perceive and focus on individual sources.

Source separation is the problem of **recovering the source signals** underlying a given mixture.

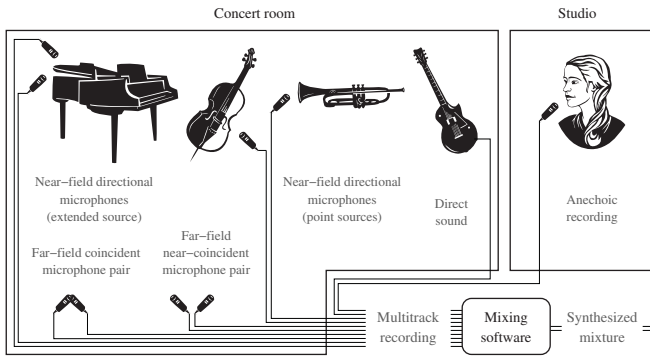
It is a core problem of audio signal processing, with applications such as:

- hearing aids,
- post-production, remixing and 3D upmixing,
- spoken/multimedia document retrieval,
- MIR.

## The data at hand

As an inverse problem, source separation requires some **knowledge**.

Music is among the most difficult application areas of source separation because of the wide variety of sources and mixing processes.



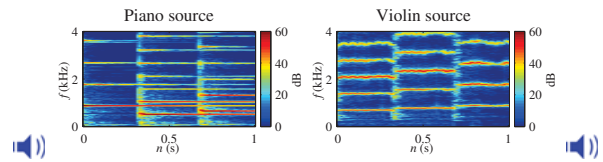
## Music sources

Music sources include acoustical or virtual instruments and singing voice.

Sound is produced by transmission of one or more excitation movements/signals through a resonant body/filter.

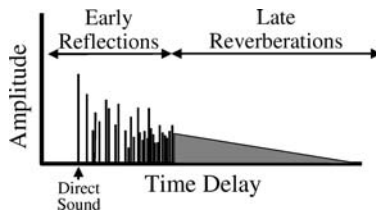
This results in a wide variety of sounds characterized by their:

- **polyphony** (monophonic or polyphonic)
- **temporal shape** (transitory, constant or variable)
- **spectral fine structure** (random or pitched)
- **spectral envelope**



## Effects of microphone recording

For point sources, room acoustics result in **filtering** of the source signal



where the **intensity** and **delay** of direct sound are functions of the **source position** relative to the microphone.

**Diffuse** sources (piano, drums) amount to (infinitely) many point sources.

The mixture signal is equal to the sum of the contributions of all sources at each microphone.

## Software mixing effects

Usual software mixing effects include:

- **compression** and **equalization**
- **panning**, *i.e.* channel-dependent intensity scaling
- **reverb**
- **polarity** and **autopan**

The latter are widely employed to achieve perceptual envelopment, whereby even point sources are mixed diffusely.

Again, the intensity of direct sound is a function of the source position and the mixture signal is equal to the sum of the contributions of all sources in each channel.

## Overview

Hundreds of source separation systems were designed in the last 20 years. . .

. . . but few are yet applicable to real-world music, as illustrated by the 2008 and 2010 Signal Separation Evaluation Campaigns (SiSEC).

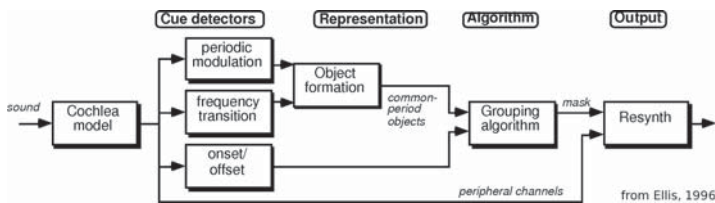
The wide variety of techniques boils down to three modeling paradigms:

- computational auditory scene analysis (CASA),
- probabilistic linear modeling, including independent component analysis (ICA) and sparse component analysis (SCA),
- probabilistic variance modeling, including hidden Markov models (HMM) and nonnegative matrix factorization (NMF).

- 1 Source separation and music
- 2 Computational auditory scene analysis
- 3 Probabilistic linear modeling
- 4 Probabilistic variance modeling
- 5 Summary and future challenges

## Computational auditory scene analysis (CASA)

CASA aims to emulate the human auditory system.



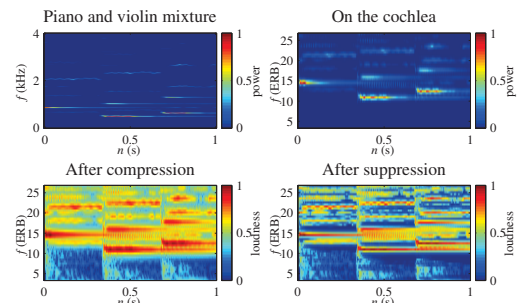
Source formation relies on the *Gestalt* rules of cognition:

- proximity,
- similarity,
- continuity,
- closure,
- common fate.

## Auditory front-end

The sound signal is first converted into an auditory nerve representation via a series of processing steps:

- outer- and middle-ear: filter
- cochlear traveling wave model: filterbank
- haircell model: halfwave rectification + bandwise compression + cross-band suppression

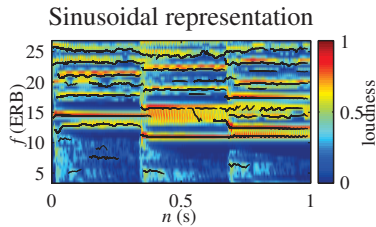


## Sinusoidal+noise decomposition

Many systems further decompose the signal into a collection of **sinusoidal tracks** plus residual noise.

This decomposition is useful to

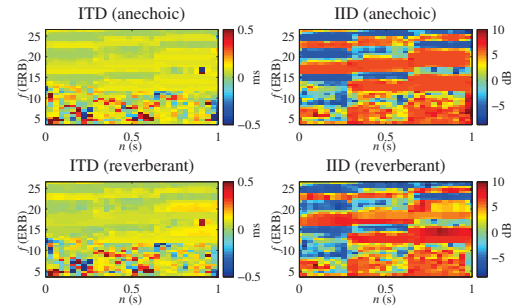
- reduce the number of sound atoms to be grouped into sources,
- enable the exploitation of advanced cues, e.g. amplitude and frequency modulation.



## Spatial cues

Spatial proximity is assessed by comparing the observed

- interchannel time difference (ITD),
- interchannel intensity difference (IID).



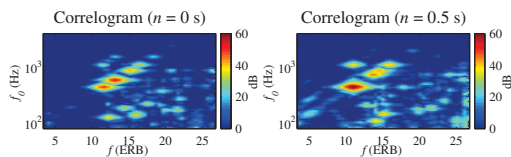
Note: in practice, most systems consider only binaural data, *i.e.* recorded by in-ear microphones.

## Spectral cues

The *Gestalt* rules also translate into e.g.

- common pitch and onset time,
- similar spectral envelope,
- spectral and temporal smoothness,
- lack of silent time intervals,
- correlated amplitude and frequency modulation.

Most effort has been devoted to the estimation of **pitch** by cross-correlation of the auditory nerve representation in each band.



## Learned cues

In addition to the above **primitive cues**, the auditory system relies on a range of **learned cues** to focus on a given source:

- veridical expectation (episodic memory): "I know the lyrics"
- schematic expectation (semantic memory): "The inaudible word after *love you must be babe*"
- dynamic adaptive expectation (short-term memory): "This melody already occurred in the song"
- conscious expectation

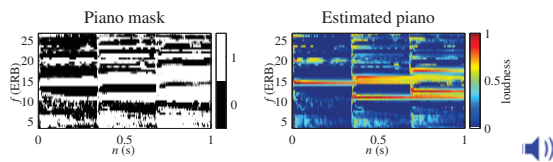
## Source formation and signal extraction

Each time-frequency bin or each sinusoidal track is associated to a single source according to the above cues: this is known as **binary masking**.

Individual cues are ambiguous, e.g.

- the observed IID/ITD may be due to a single source in the associated direction or to several concurrent sources around that direction,
- a given sinusoidal track may be a harmonic of different sources.

Most systems exploit several cues with some **precedence order** or **weighting factors** determined by psycho-acousticians.



## Summary of CASA

Advantages:

- wide range of spectral, spatial and learned cues
- robustness thanks to joint exploitation of several cues

Limitations:

- musical noise artifacts due to binary masking
- suboptimal cues, designed for auditory scene analysis instead of machine source separation
- practical limitation to a few spectral and/or spatial cues, with no general framework for the integration of additional cues
- (historically) bottom-up approach, prone to error propagation, and limitation to pitched sources
- no results within recent evaluation campaigns

## Model-based audio source separation

- 1 Source separation and music
- 2 Computational auditory scene analysis
- 3 Probabilistic linear modeling
- 4 Probabilistic variance modeling
- 5 Summary and future challenges

The alternative top-down approach consists of finding the source signals that best fit the mixture and the expected properties of audio sources.

In a probabilistic framework, this translates into

- building **generative models** of the source and mixture signals,
- inferring latent variables in a **maximum a posteriori (MAP)** sense.

## Linear modeling

The established linear modeling paradigm relies on two assumptions:

- 1 point sources
- 2 low reverberation

Under assumption 1, the sources and the mixing process can be modeled as **single-channel source signals** and a **linear filtering process**.

Under assumption 2, this filtering process is equivalent to complex-valued multiplication in the **time-frequency domain** via the short-time Fourier transform (STFT).

In each time-frequency bin  $(n, f)$

$$\mathbf{X}_{nf} = \sum_{j=1}^J S_{jnf} \mathbf{A}_{jf}$$

$\mathbf{X}_{nf}$ : vector of mixture STFT coeff.  
 $J$ : number of sources  
 $S_{jnf}$ :  $j$ th source STFT coeff.  
 $\mathbf{A}_{jf}$ :  $j$ th mixing vector

## Priors over the mixing vectors

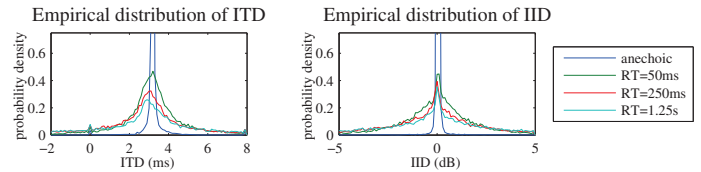
The mixing vectors  $\mathbf{A}_{jf}$  encode the apparent sound direction in terms of

- ITD  $\tau_{jf}$ ,
- IID  $g_{jf}$ .

For non-echoic mixtures, ITDs and IIDs are **constant over frequency** and related to the direction of arrival (DOA)  $\theta_j$  of each source

$$\mathbf{A}_{jf} \propto \begin{pmatrix} 1 \\ g_j e^{-2i\pi f \tau_j} \end{pmatrix}$$

For echoic mixtures, ITDs and IIDs follow a **smeared distribution**  $P(\mathbf{A}_{jf}|\theta_j)$



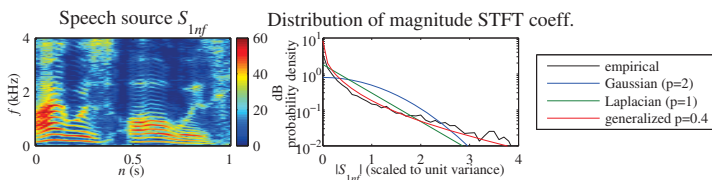
## i.i.d. priors over the source STFT coefficients

Most systems assume that the sources have random spectra, *i.e.* their STFT coefficients  $S_{jnf}$  are **independent and identically distributed (i.i.d.)**.

The magnitude STFT coefficients of audio sources are **sparse**: at each frequency, few coefficients have large values while most are close to zero.

This property is well modeled by the **generalized exponential distribution**

$$P(|S_{jnf}| | \rho, \beta_f) = \frac{\rho}{\beta_f \Gamma(1/\rho)} e^{-\left| \frac{S_{jnf}}{\beta_f} \right|^\rho} \quad \begin{array}{l} \rho: \text{shape parameter} \\ \beta_f: \text{scale parameter} \end{array}$$



Note: coarser binary activity priors have also been employed.

## Inference algorithms

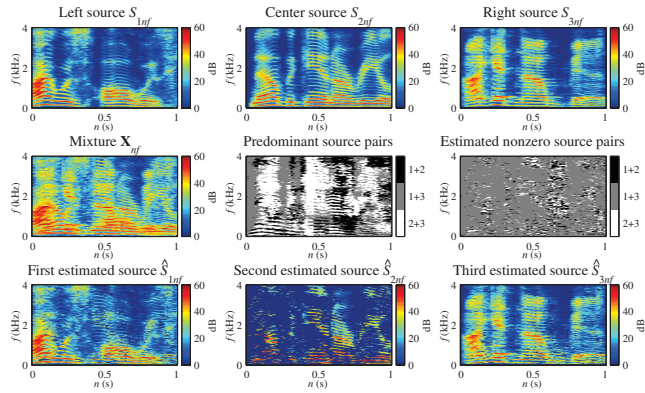
Given the above priors, source separation is typically achieved by joint MAP estimation of the source STFT coefficients  $S_{jnf}$  and other latent variables  $(\mathbf{A}_{jf}, g_j, \tau_j, \rho, \beta_j)$  via **alternating nonlinear optimization**.

This objective is called **sparse component analysis (SCA)**.

For typical values of  $\rho$ , the MAP source STFT coefficients are **nonzero for at most two sources** in a stereo setting.

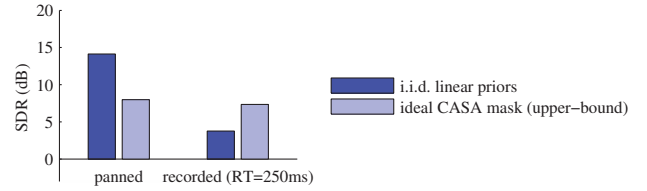
When the number of sources is  $J = 2$ , SCA is renamed **nongaussianity-based frequency-domain independent component analysis (FDICA)**.





### Practical illustration of separation using i.i.d. linear priors



Time-frequency bins dominated by the center source are often erroneously associated with the two other sources.

### SiSEC results on toy mixtures of 3 sources



Panned mixture   
 Estimated sources using i.i.d. linear priors   
 Recorded reverberant mixture   
 Estimated sources using i.i.d. linear priors 

### Summary of probabilistic linear modeling

Advantages:

- top-down approach
- separation of more than one source per time-frequency bin

Limitations:

- restricted to mixtures of non-reverberated point sources
- separation of at most two sources per time-frequency bin
- musical noise artifacts due to the ambiguities of spatial cues
- no straightforward framework for the integration of spectral cues

- 1 Source separation and music
- 2 Computational auditory scene analysis
- 3 Probabilistic linear modeling
- 4 Probabilistic variance modeling
- 5 Summary and future challenges

## Idea 1: from sources to mixture components

Diffuse or semi-diffuse sources cannot be modeled as single-channel signals and not even as finite dimensional signals.

Instead of considering the signal produced by each source, one may consider its contribution to each channel of the mixture signal.

Source separation becomes the problem of estimating the **multichannel mixture components** underlying the mixture.

In each time-frequency bin  $(n, f)$

$$\mathbf{x}_{nf} = \sum_{j=1}^J \mathbf{c}_{jnf}$$

$\mathbf{x}_{nf}$ : vector of mixture STFT coeff.  
 $J$ : number of sources  
 $\mathbf{c}_{jnf}$ :  $j$ th mixture component

## Idea 2: translation and phase invariance

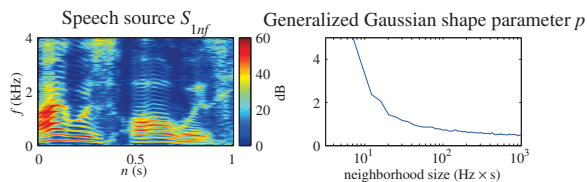
In order to overcome the ambiguities of spatial cues, additional spectral cues are needed as shown by CASA.

Most audio sources are **translation- and phase-invariant**: a given sound may be produced at any time with any relative phase across frequency.

## Variance modeling

Variance modeling combines these two ideas by modeling the STFT coefficients of individual mixture components by a **circular multivariate distribution whose parameters vary over time and frequency**.

The non-sparsity of source STFT coefficients over small time-frequency regions suggests the use of a **non-sparse distribution**.



## Choice of the distribution

For historical reasons, several distributions have been preferred in a mono context, which can equivalently be expressed as **divergence** functions over the source magnitude/power STFT coefficients:

- Poisson  $\leftrightarrow$  Kullback-Leibler divergence aka I-divergence
- tied-variance Gaussian  $\leftrightarrow$  Euclidean distance
- log-Gaussian  $\leftrightarrow$  weighted log-Euclidean distance

These distributions do not easily generalize to multichannel data.



## The multichannel Gaussian model

The zero-mean Gaussian distribution is a simple multichannel model.

$$P(\mathbf{C}_{jnf} | \boldsymbol{\Sigma}_{jnf}) = \frac{1}{\det(\pi \boldsymbol{\Sigma}_{jnf})} e^{-\mathbf{C}_{jnf}^H \boldsymbol{\Sigma}_{jnf}^{-1} \mathbf{C}_{jnf}} \quad \boldsymbol{\Sigma}_{jnf}: j\text{th component covariance matrix}$$

The covariance matrix  $\boldsymbol{\Sigma}_{jnf}$  of each mixture component can be factored as the product of a **scalar nonnegative variance**  $V_{jnf}$  and a **mixing covariance matrix**  $\mathbf{R}_{jf}$  respectively modeling spectral and spatial properties

$$\boldsymbol{\Sigma}_{jnf} = V_{jnf} \mathbf{R}_{jf}$$

Under this model, the mixture STFT coefficients also follow a Gaussian distribution whose covariance is the sum of the component covariances

$$P(\mathbf{X}_{nf} | V_{jnf}, \mathbf{R}_{jf}) = \frac{1}{\det(\pi \sum_{j=1}^J V_{jnf} \mathbf{R}_{jf})} e^{-\mathbf{X}_{nf}^H (\sum_{j=1}^J V_{jnf} \mathbf{R}_{jf})^{-1} \mathbf{X}_{nf}}$$

## General inference algorithm

Independently of the priors over  $V_{jnf}$  and  $\mathbf{R}_{jf}$ , source separation is typically achieved in two steps:

- joint MAP estimation of all model parameters using the **expectation maximization (EM)** algorithm,
- MAP estimation of the source STFT coefficients conditional to the model parameters by **multichannel Wiener filtering**

$$\hat{\mathbf{C}}_{jnf} = V_{jnf} \mathbf{R}_{jf} \left( \sum_{j'=1}^J V_{j'nf} \mathbf{R}_{j'f} \right)^{-1} \mathbf{X}_{nf}.$$

## Rank-1 priors over the mixing covariances

The mixing covariances  $\mathbf{R}_{jf}$  encode the apparent spatial direction and spatial spread of sound in terms of

- ITD,
- IID,
- normalized interchannel correlation a.k.a. **interchannel coherence**.

For non-reverberated point sources, the interchannel coherence is equal to one, *i.e.*  $\mathbf{R}_{jf}$  has **rank 1**

$$\mathbf{R}_{jf} = \mathbf{A}_{jf} \mathbf{A}_{jf}^H$$

The priors  $P(\mathbf{A}_{jf} | \theta_j)$  used with linear modeling can then be simply reused.

## Full-rank priors over the mixing covariances

For reverberated or diffuse sources, the interchannel coherence is smaller than one, *i.e.*  $\mathbf{R}_{jf}$  has **full rank**.

The theory of statistical room acoustics suggests the **direct+diffuse model**

$$\mathbf{R}_{jf} \propto \lambda_j \mathbf{A}_{jf} \mathbf{A}_{jf}^H + \mathbf{B}_f$$

$\lambda_j$ : direct-to-reverberant ratio  
 $\mathbf{A}_{jf}$ : direct mixing vector  
 $\mathbf{B}_f$ : diffuse noise covariance

with

$$\mathbf{A}_{jf} = \sqrt{\frac{2}{1+g_j^2}} \begin{pmatrix} 1 \\ g_j e^{-2i\pi f \tau_j} \end{pmatrix}$$

$\tau_j$ : ITD of direct sound  
 $g_j$ : IID of direct sound

$$\mathbf{B}_f = \begin{pmatrix} 1 & \text{sinc}(2\pi f d/c) \\ \text{sinc}(2\pi f d/c) & 1 \end{pmatrix}$$

$d$ : microphone spacing  
 $c$ : sound speed

### i.i.d. priors over the source variances

Baseline systems rely again on the assumption that the sources have random spectra and model the source variances  $V_{jnf}$  as i.i.d. and locally constant within small time-frequency regions.

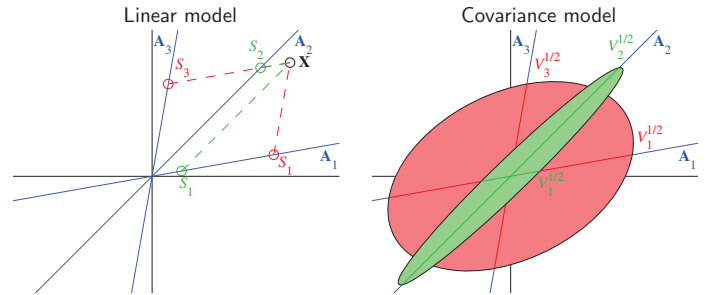
When these follow a mildly sparse prior, it can be shown that the MAP variances are nonzero for up to four sources.

Discrete priors constraining the number of nonzero variances to one or two have also been employed.

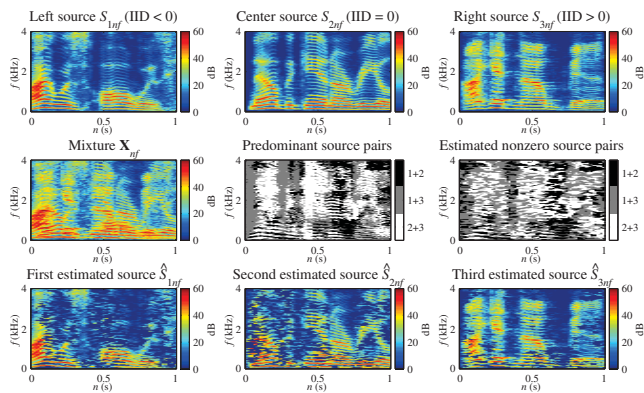
When the number of sources is  $J = 2$ , this model is also called nonstationarity-based FDICA.

### Benefit of exploiting interchannel coherence

Interchannel coherence helps resolving some ambiguities of ITD and IID and identify the predominant sources more accurately.



### Practical illustration of separation using i.i.d. variance priors



### Spectral priors based on template spectra

Variance modeling enables the design of phase-invariant spectral priors.

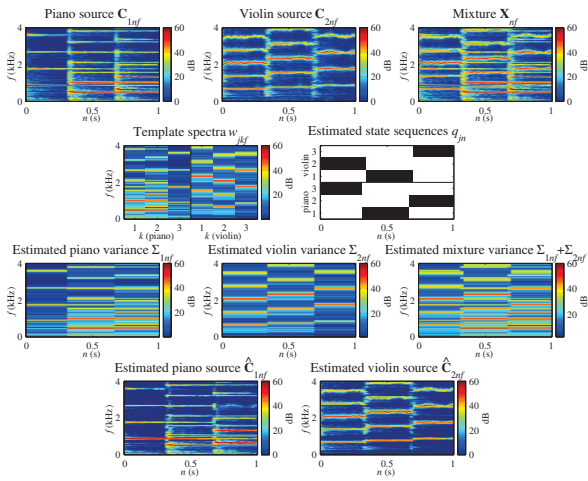
The Gaussian mixture model (GMM) represents the variance  $V_{jnf}$  of each source at a given time by one of  $K$  template spectra  $w_{jkf}$  indexed by a discrete state  $q_{jn}$

$$V_{jnf} = w_{jq_{jn}f} \text{ with } P(q_{jn} = k) = \pi_{jk}$$

Different strategies have been proposed to learn these spectra:

- speaker-independent training on separate single-source data,
- speaker-dependent training on separate single-source data,
- MAP adaptation to the mixture using model selection or interpolation,
- MAP inference from a coarse initial separation.

### Practical illustration of separation using template spectra



### Spectral priors based on basis spectra

The GMM does not efficiently model polyphonic musical instruments.

The variance  $V_{jnf}$  of each source is then better represented as the linear combination of  $K$  basis spectra  $w_{jkf}$  multiplied by time-varying scale factors  $h_{jkn}$

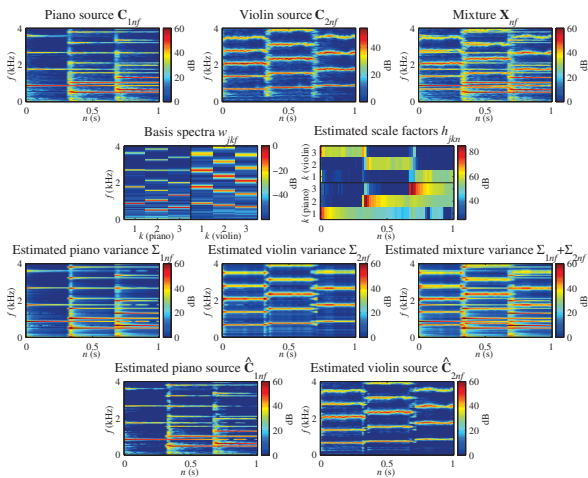
$$V_{jnf} = \sum_{k=1}^K h_{jkn} w_{jkf}$$

This model is also called nonnegative matrix factorization (NMF).

Again, a range of strategies have been used to learn these spectra:

- instrument-dependent training on separate single-source data,
- MAP adaptation to the mixture using uniform priors,
- MAP adaptation to the mixture using trained priors.

### Practical illustration of separation using basis spectra



### Constrained template/basis spectra

MAP adaptation or inference of the template/basis spectra is often needed due to

- the lack of training data,
- the mismatch between training and test data.

However, it is often inaccurate: additional constraints over the spectra are needed to further reduce overfitting.

## Harmonicity and spectral smoothness constraints

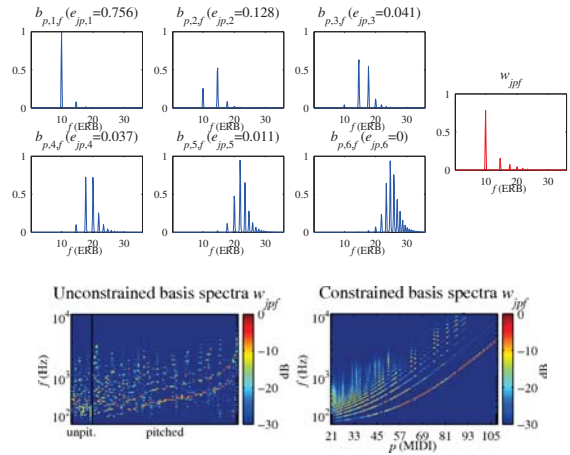
For instance, harmonicity and spectral smoothness can be enforced by

- associating each basis spectrum with some a priori pitch  $p$
- modeling  $w_{jpf}$  as the sum of fixed narrowband spectra  $b_{p,f}$  representing adjacent partials at harmonic frequencies scaled by spectral envelope coefficients  $e_{jp,l}$

$$w_{jpf} = \sum_{l=1}^{L_p} e_{jpl} b_{plf}.$$

Parameter estimation now amounts to estimating the active pitches and their spectral envelopes instead of their full spectra.

## Practical illustration of harmonicity constraints



## Further constraints

Further constraints that have been implemented in this context include

- source-filter model of instrumental timbre,
- inharmonicity and tuning.

Probabilistic priors are also popular:

- state transition priors

$$P(q_{jn} = k | q_{j,n-1} = l) = \pi_{jkl}$$

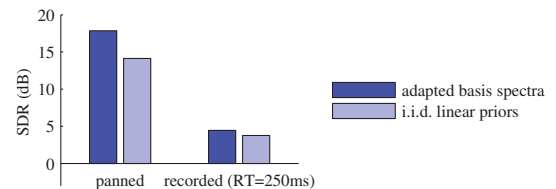
- spectral continuity priors (for percussive sounds)

$$P(V_{jnf} | V_{jn,f-1}) = \mathcal{N}(V_{jnf}; V_{jn,f-1}, \sigma_{perc})$$

- temporal continuity priors (for sustained sounds)

$$P(V_{jnf} | V_{j,n-1}, f) = \mathcal{N}(V_{jnf}; V_{j,n-1}, f, \sigma_{sust})$$

## SiSEC results on toy mixtures of 3 sources



Panned mixture

Estimated sources using adapted basis spectra

Estimated sources using i.i.d. linear priors



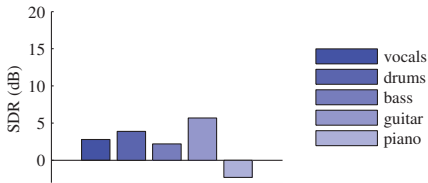
Recorded reverberant mixture

Estimated sources using adapted basis spectra

Estimated sources using i.i.d. linear priors



## SiSEC results on professional mixtures



*Tamy* (2 sources)

Estimated sources using adapted basis spectra



*Bearlin* (10 sources)

Estimated sources using adapted basis spectra



## Summary of probabilistic variance modeling

Advantages:

- top-down approach
- virtually applicable to any mixture, including to diffuse sources
- no hard constraint on the number of sources per time-frequency bin
- fewer musical noise artifacts by joint exploitation of spatial, spectral and learned cues
- principled modular framework for the integration of additional cues

Limitations:

- remaining musical noise artifacts
- current implementations limited to a few spectral and/or spatial cues... but this is gradually changing!

- 1 Source separation and music
- 2 Computational auditory scene analysis
- 3 Probabilistic linear modeling
- 4 Probabilistic variance modeling
- 5 Summary and future challenges

## Summary principles of model-based source separation

Most model-based source separation systems rely on modeling the STFT coefficients of each source as a function of

- a scalar variable ( $S_{jnf}$  or  $V_{jnf}$ ) encoding spectral cues,
- a vector or matrix variable ( $\mathbf{A}_{jf}$  or  $\mathbf{R}_{jf}$ ) encoding spatial cues.

Robust source separation requires priors over both types of cues:

- spectral cues alone cannot discriminate sources with similar pitch range and timbre,
- spatial cues alone cannot discriminate sources with the same DOA.

A range of informative priors have been proposed, relating for example

- $S_{jnf}$  or  $V_{jnf}$  to discrete or continuous latent states,
- $\mathbf{A}_{jf}$  or  $\mathbf{R}_{jf}$  to the source DOAs.

Variance modeling outperforms linear modeling.

## Conclusion and remaining challenges

To sum up, source separation is a **core problem of audio signal processing with huge potential applications**.

Existing systems are **gradually finding their way into the industry**, especially for applications that can accommodate

- a certain amount of musical noise artifacts, such as MIR,
- partial user input/feedback, such as post-production.

We believe that these two **limitations could be addressed in the next 10 years** by exploiting the full power of probabilistic modeling, especially by:

- integrating more and more spatial and spectral cues,
- making a better use of learned cues, using training data or repeated sounds

## References

D.L. Wang and G.J. Brown, Eds.  
*Computational Auditory Scene Analysis: Principles, Algorithms and Applications*  
Wiley/IEEE Press, 2006.

E. Vincent, M.G. Jafari, S.A. Abdallah, M.D. Plumbley, and M.E. Davies  
Probabilistic modeling paradigms for audio source separation  
in *Machine Audition: Principles, Algorithms and Systems*  
IGI Global, 2010.

2008 and 2010 Signal Separation Evaluation Campaigns  
<http://sisec.wiki.irisa.fr/>

# Music Source Separation and its Applications to MIR

Nobutaka Ono and Emmanuel Vincent

The University of Tokyo, Japan

INRIA Rennes - Bretagne Atlantique, France

Tutorial supported by the VERSAMUS project

<http://versamus.inria.fr/>

Contributions from Shigeki Sagayama, Kenichi Miyamoto, Hirokazu Kameoka,  
Jonathan Le Roux, Emiru Tsunoo, Yushi Ueda, Hideyuki Tachibana,  
Geroge Tzanetakis, Halfdan Rump, Other members of IPC Lab#1

## Outline

- Introduction
- Part I: Brief Introduction of State-of-the-arts
  - Singer/Instrument Identification
  - Audio Tempo Estimation
- Part II: Harmonic/Percussive Sound Separation
  - Motivation and Formulation
  - Open Binary Software
- Part III: Applications of HPSS to MIR Tasks
  - Audio Chord Estimation
  - Melody Extraction
  - Audio Genre Classification
- Conclusions

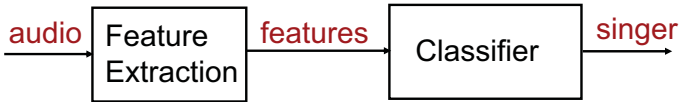
## Introduction

- Focus of the second half of this tutorial is to clarify
  - What **source separation** has been used for MIR?
  - How does it **improve** performance of MIR tasks?
- Examples:
  - **Multi pitch estimation**  
Task itself is tightly coupled with source separation.
  - **Audio genre classification**  
How source separation is useful?  
Not straightforward.

## Part I: Brief Introduction of State-of-the-arts

# Singer Identification

- Task: Identify a singer from music audio with accompaniment
- Typical approach



# Accompaniment Sound Reduction [Fujihara2005]

- Pre-dominant F0 based voice separation

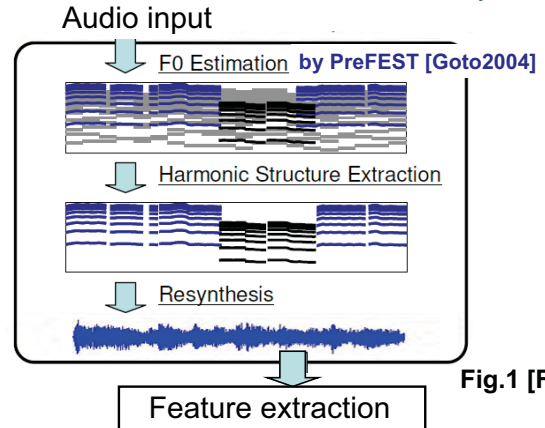


Fig.1 [Fujihara2005]

# Reliable Frame Selection [Fujihara2005]

- Only reliable frame is used for classification

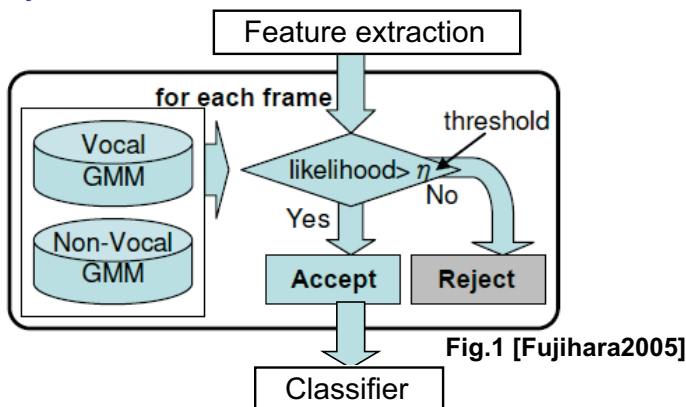


Fig.1 [Fujihara2005]

# Evaluation by Confusing Matrix

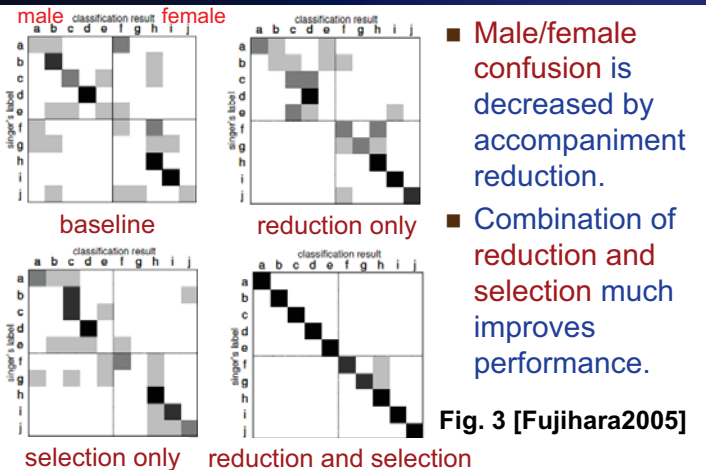


Fig. 3 [Fujihara2005]

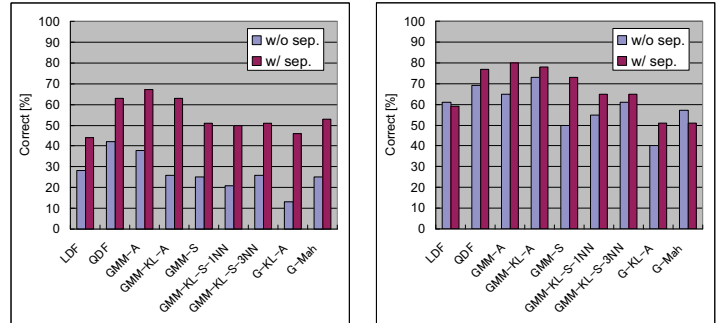
- Male/female confusion is decreased by accompaniment reduction.
- Combination of reduction and selection much improves performance.



## Vocal Separation Based on Melody Transcriber

- **Melody-F0-based Vocal Separation** [Mesaros2007]
  - Estimate melody-F0 by melody transcription system [Ryynanen2006].
  - Generate harmonic overtones at multiple of estimated F0.
  - Estimate amplitudes and phases of overtones based on cross correlation between original signal and complex exponentials.
- They evaluate the effect of separation in singer identification performance using by different classifiers.

## Evaluation by Identification Rate



Singing to Accompaniment Ratio: -5dB Singing to Accompaniment Ratio: 15dB  
Generated by Table 1 and 2 [Mesaros2007]

Performance is much improved, especially in low singing-to-accompaniment ratio.

## Instrument Identification

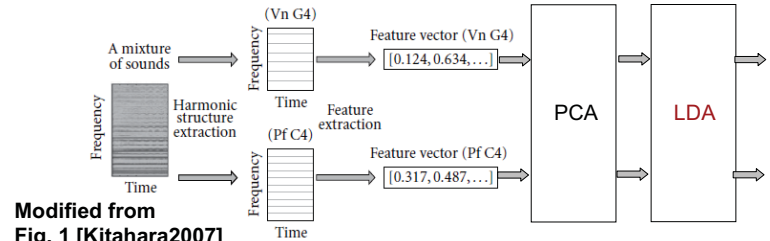
- **Task: Determine instruments present in music piece**
- **Typical approach**

```

graph LR
    audio --> S[Separation to Notes]
    S -- spectrogram of notes --> FE[Feature Extraction]
    FE -- features --> C[Classifier]
    C -- instrument --> out
            
```
- **Important Issue**
  - Source separation is **not perfect**. How to reduce errors?

## Feature Weighting [Kitahara2007]

- Feature vectors of each instrument are collected from polyphonic music for training.
- Robustness of each feature is evaluated by ratio of intra-class variance to inter-class variance: Applying Linear discriminant analysis (LDA) for feature weighting.



Modified from Fig. 1 [Kitahara2007]

## Effectiveness of Feature Weighting

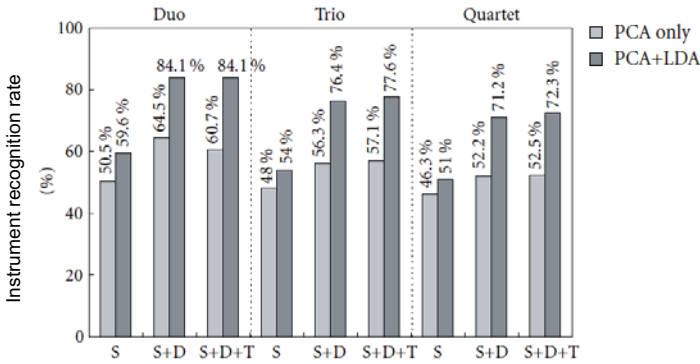
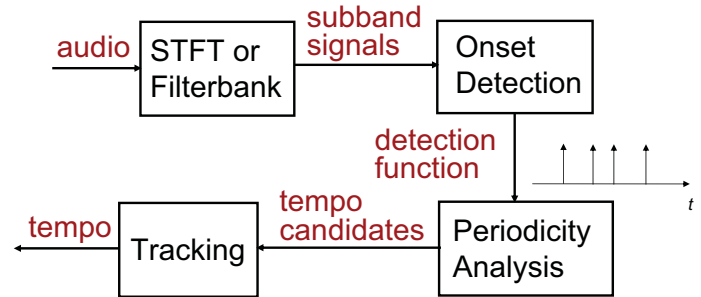


Fig. 6 [Kitahara2007]

Feature weighting by LDA improves recognition rate.

## Audio Tempo Estimation

- Task: Extract tempo from musical audio
- Typical approach:



## Applying Harmonic+Noise Model

- Harmonic+Noise model is applied before calculating detection function [Alonso2007]

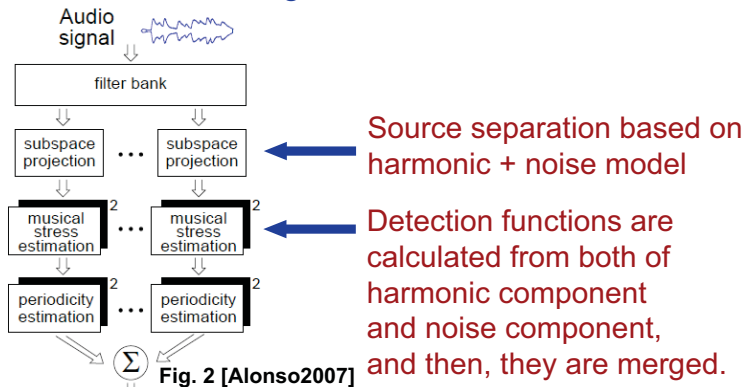


Fig. 2 [Alonso2007]

## Influence of S+N Model

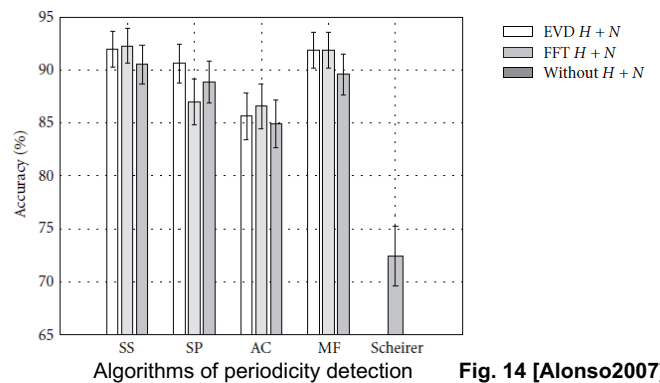
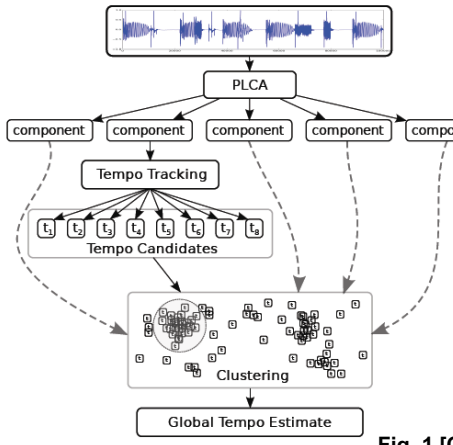


Fig. 14 [Alonso2007]

Separation based on H+N model shows better results.

# Applying PLCA



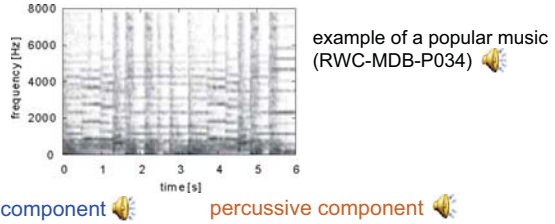
- PLCA (Probabilistic Latent Component Analysis), NMF-like method is applied.
- It increases much candidates of tempo.
- They report its effectiveness.

Fig. 1 [Chordia2009]

# Part II: Harmonic/Percussive Sound Separation

# Motivation and Goal of HPSS

- **Motivation:** Music consists of two different components



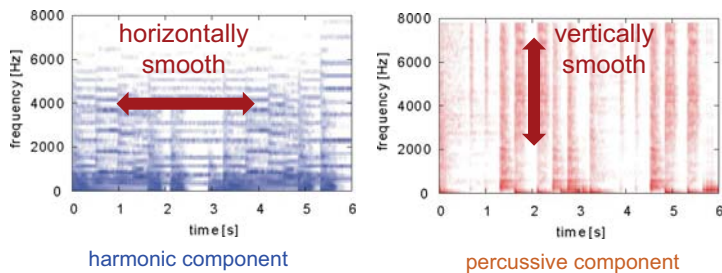
- **Goal:** Separation of a monaural audio signal into harmonic and percussive components
- **Target:** MIR-related tasks
  - multi-pitch analysis, chord recognition... H-related
  - beat tracking, rhythm recognition... P-related

# Related Works to H/P Separation

- **Source separation into multiple components followed by classification**
  - ICA and classification [Uhle2003]
  - NMF and classification [Helen2005]
- **Steady + Transient model**
  - Adaptive phase vocoder
  - Subspace projection
  - Matching pursuit
  - ...etc

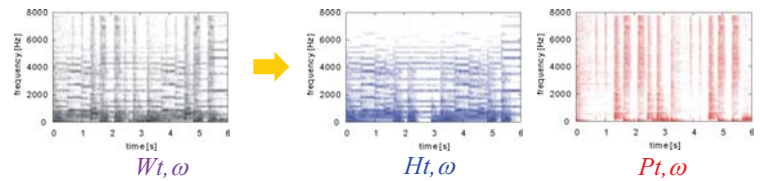
Good review is provided in [Daudet2005]
- **Bayesian NMF [Dikmen2009]**

## Point: Anisotropy of Spectrogram



## H/P Separation Problem

- **Problem:**  
Find  $H_{t,\omega}$  and  $P_{t,\omega}$  from  $W_{t,\omega}$  on power spectrogram



- **Requirements:**
  - 1)  $H_{t,\omega}$ : horizontally smooth
  - 2)  $P_{t,\omega}$ : vertically smooth
  - 3)  $H_{t,\omega}$  and  $P_{t,\omega}$ : non-negative
  - 4)  $H_{t,\omega} + P_{t,\omega}$ : should be close to  $W_{t,\omega}$

## Formulation of H/P Separation (1/2)

- **Formulation as an Optimization Problem:**

- Objective function to minimize

$$J(\mathbf{H}, \mathbf{P}) = \underbrace{D(\mathbf{W}, \mathbf{H} + \mathbf{P})}_{\text{Closeness cost}} + \underbrace{C_H(\mathbf{H}) + C_P(\mathbf{P})}_{\text{Smoothness cost}}$$

- Under constraints:

- $H_{t,\omega} \geq 0$
- $P_{t,\omega} \geq 0$

In MAP estimation context, they are corresponding likelihood term and prior term, respectively.

## Formulation of H/P Separation (2/2)

- **Closeness cost function: l-divergence**

$$D(\mathbf{W}, \mathbf{H} + \mathbf{P}) = \sum_{\omega, \tau} \left\{ W_{\omega, \tau} \log \frac{W_{\omega, \tau}}{H_{\omega, \tau} + P_{\omega, \tau}} - W_{\omega, \tau} + H_{\omega, \tau} + P_{\omega, \tau} \right\}$$

- **Smoothness cost function: Square of difference**

$$C_H(\mathbf{H}) = \sum_{\omega, \tau} \frac{1}{2\sigma_H^2} (H_{\omega, \tau-1}^\gamma - H_{\omega, \tau}^\gamma)^2$$

$$C_P(\mathbf{H}) = \sum_{\omega, \tau} \frac{1}{2\sigma_P^2} (P_{\omega-1, \tau}^\gamma - P_{\omega, \tau}^\gamma)^2$$

Weights to control two smoothness

$\gamma = 0.5$   
for scale invariance

- A variance modeling-based separation using
  - Poisson observation distribution
  - Gaussian continuity priors

[Miyamoto2008, Ono2008, etc]

# Update Rules

Update alternatively two kinds of variables:

H and P:

$$H_{\omega,\tau} \leftarrow \left( \frac{b_{H_{\omega,\tau}} + \sqrt{b_{H_{\omega,\tau}}^2 + 4a_{H_{\omega,\tau}}c_{H_{\omega,\tau}}}}{2a_{H_{\omega,\tau}}} \right)^2$$

$$P_{\omega,\tau} \leftarrow \left( \frac{b_{P_{\omega,\tau}} + \sqrt{b_{P_{\omega,\tau}}^2 + 4a_{P_{\omega,\tau}}c_{P_{\omega,\tau}}}}{2a_{P_{\omega,\tau}}} \right)^2$$

Auxiliary variables:

$$m_{P_{\omega,\tau}} = \frac{P_{\omega,\tau}}{H_{\omega,\tau} + P_{\omega,\tau}}$$

$$m_{H_{\omega,\tau}} = \frac{H_{\omega,\tau}}{H_{\omega,\tau} + P_{\omega,\tau}}$$

$$a_{P_{\omega,\tau}} = \frac{2}{\sigma_P^2} + 2$$

$$a_{H_{\omega,\tau}} = \frac{2}{\sigma_H^2} + 2$$

$$b_{P_{\omega,\tau}} = \frac{(\sqrt{P_{\omega-1,\tau}} + \sqrt{P_{\omega+1,\tau}})}{\sigma_P^2}$$

$$b_{H_{\omega,\tau}} = \frac{(\sqrt{H_{\omega,\tau-1}} + \sqrt{H_{\omega,\tau+1}})}{\sigma_H^2}$$

$$c_{P_{\omega,\tau}} = 2m_{P_{\omega,\tau}}W_{\omega,\tau}$$

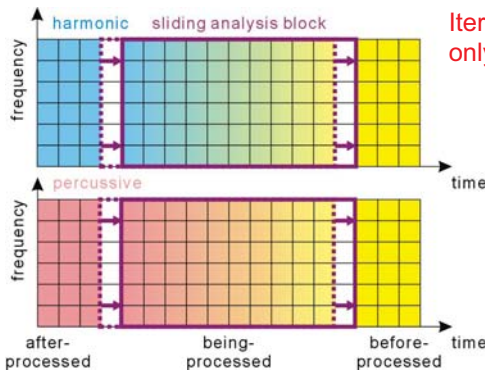
$$c_{H_{\omega,\tau}} = 2m_{H_{\omega,\tau}}W_{\omega,\tau}$$

# Separated Examples

Music piece	original	H	P
RWC-MDB-P-7 "PROLOGUE "			
RWC-MDB-P-12 "KAGE-ROU "			
RWC-MDB-P-18 "True Heart"			
RWC-MDB-P-25 "tell me"			
RWC-MDB-J-16 "Jive "			

# Real-Time Implementation

Sliding Block Analysis



Iterations are applied only within sliding block

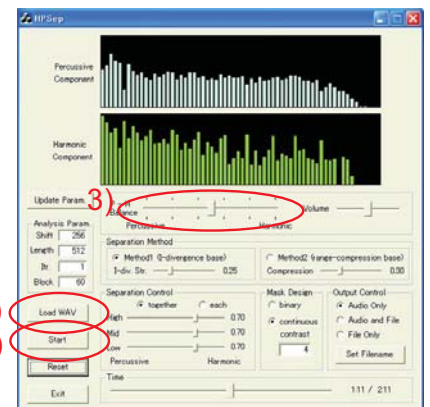
# Open Software: Real-time H/P equalizer

Available at <http://www.hil.t.u-tokyo.ac.jp/software/HPSS/>

Control H/P balance of audio signal in real time

Simple instructions:

- 1) Click "Load WAV" button and choose a WAV-formatted audio file.
- 2) Click "Start" button, and then, audio starts.
- 3) Slide H/P balance bar as you like and listen how the sound changes.

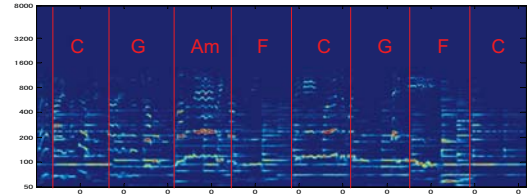


# Audio Chord Detection

## Part III: Applications of HPSS to MIR Tasks

- **Task:** Estimate chord sequence and its segmentation from music audio

### III-1: Audio Chord Detection

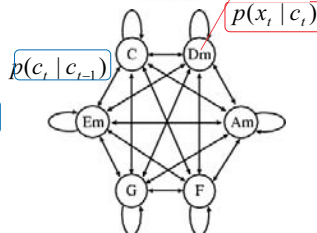
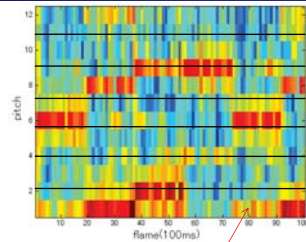


## Typical Approach: Chroma Feature + HMM

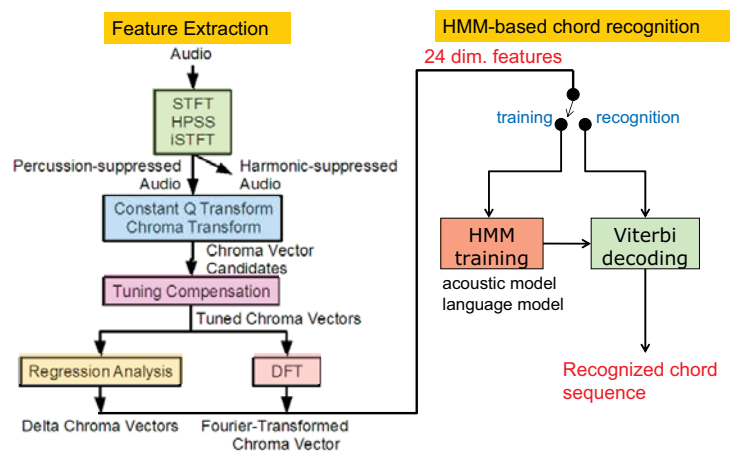
- **Feature:** chroma [Fujishima1999]
  - Chroma observation probability  $p(x_t | c_t)$
- **Transition:** chord progression
  - Bigram probability  $p(c_t | c_{t-1})$
- **Maximum a Posteriori Chord Estimation** [Sheh2003]
  - Viterbi algorithm for ...

$$\operatorname{argmax}_c p(x_0 | c_0) p(c_0) \prod_{t=1}^T p(x_t | c_t) p(c_t | c_{t-1})$$

Initial prob. emission transition

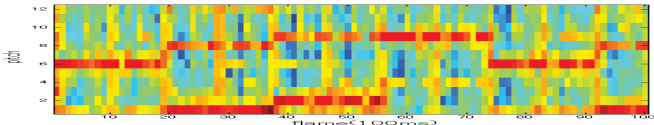


## Feature-refined System [Ueda2009]

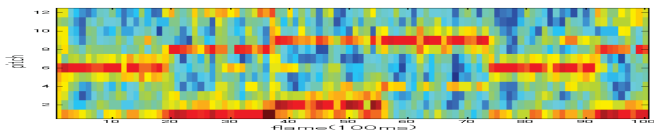


## Suppressing Percussive Sounds

- Percussive sounds are harmful in chord detection

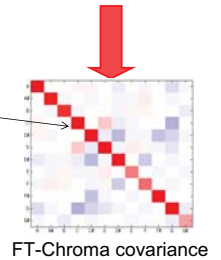
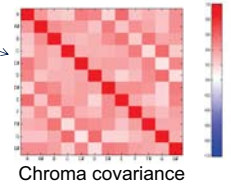


↓ Emphasize harmonic components by HPSS



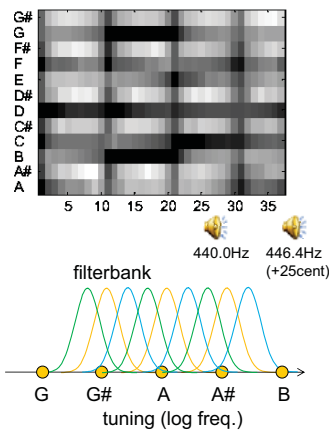
## Fourier-transformed Chroma

- Covariance matrix of chroma
  - Highly correlated components: diagonal-only approximation infeasible
    - Caused by harmonic overtones or some pitches performed at the same time
    - Results in large number of parameters
- Covariance matrix is near circulant
  - Assuming ...
    - Harmonic overtones of all pitches have the same structure
    - The amount of occurrence of the same intervals is the same
  - Circulant matrix diagonalized by DFT
- Diagonal approximation of FT-Chroma covariance
  - Reduces the number of model parameters (statistically robust)



## Tuning Compensation

- Tuning difference among songs
  - Neglecting this may blur chroma features
- Choose best tuning from multiple candidates
  - Find maximum chroma energy (sum of all bins of chroma)
  - Assume: tuning does not change within a song

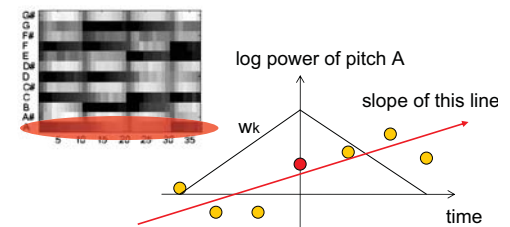


## Delta Chroma Features

- Improve chord boundary accuracy
  - by features representing chord boundaries
- Chord tones largely changes at chord boundary
  - Delta chroma: derivative of chroma features
  - Cf. Delta cepstrum (MFCC): Effective features of speech recognition
- Calculated by regression analysis of  $\delta$  sample points [Sagayama&Itakura1979]
  - Robust to noise

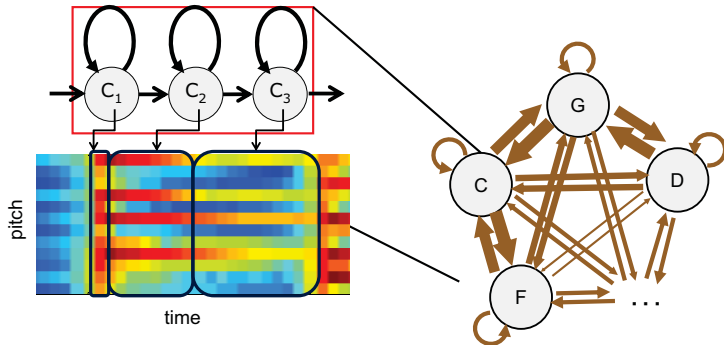
$$\Delta C(i, n) = \frac{\sum_{k=-\delta}^{\delta} k w_k C(i, t+k)}{\sum_{k=-\delta}^{\delta} k^2 w_k}$$

$$i = 1, \dots, 12$$



## Multiple States per Chord

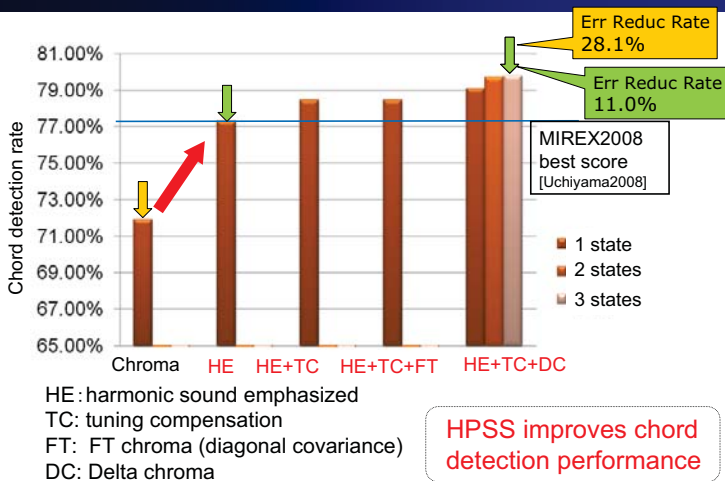
- Chroma changes from “onset” to “release”
  - capture the change by having multiple states per chord
  - tradeoff between data size and the number of states



## Experimental Evaluation

- Test Data
  - 180 songs (12 albums) of The Beatles (chord reference annotation provided by C. Harte)
  - 11.025 kHz sampling, 16bit, 1ch, WAV file
  - Frequency range: 55.0Hz-1661.2Hz (5 octaves)
- Labels
  - 12 × major/minor = 24 chords + N (no chord)
- Evaluation
  - Album filtered 3-fold cross validation
    - 8 albums for training, 4 albums for testing
  - Frame Recognition Rate = (#correct frames) / (#total frames)
  - Sampled every 100ms

## Chord Detection Results



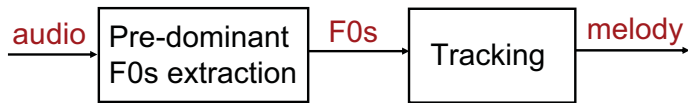
## Part III: Applications of HPSS to MIR Tasks

### III-2: Melody Extraction



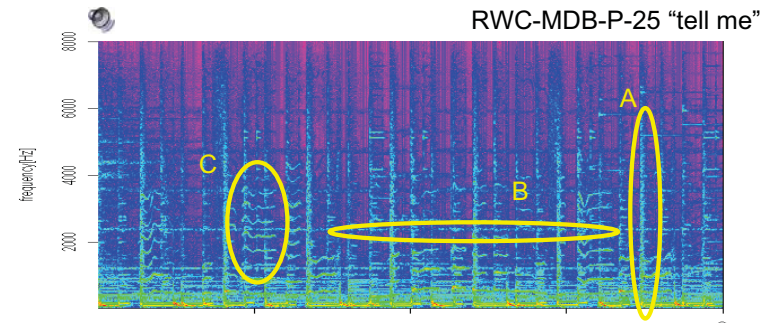
## Melody Extraction

- **Task:** Identify a melody pitch contour from polyphonic musical audio
- **Typical approach:**



- Singing voice enhancement will be useful pre-processing.

## Singing Voice in Spectrogram

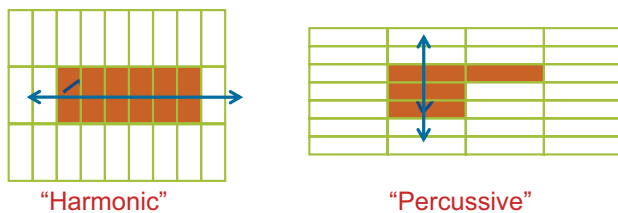


- A. Vertical component: Percussion
- B. Horizontal component: Harmonic instrument (piano, guitar, etc..)
- C. Fluctuated component: Singing voice

## Is voice harmonic or percussive?

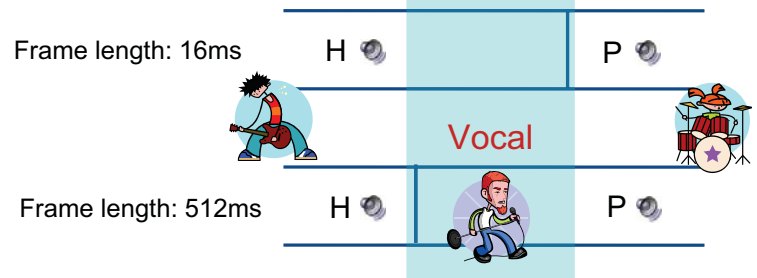
Depends on spectrogram resolution (frame-length)

- On **short-frame** STFT domain, voice appears as "H" (time direction clustered).
- On **long-frame** STFT domain, voice appears as "P" (frequency direction clustered).

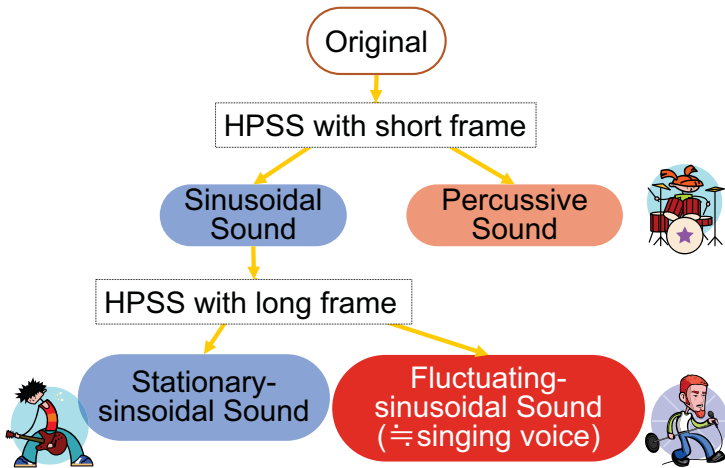


## HPSS results with different frame length

Example

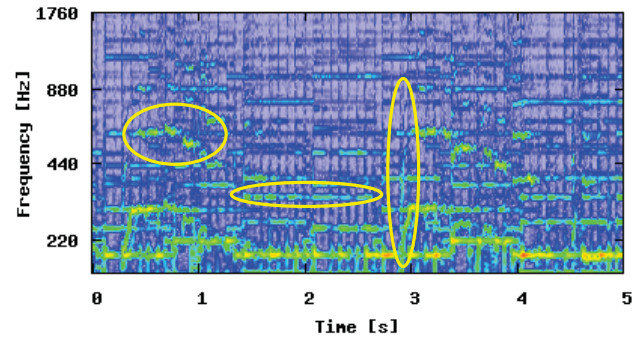


## Two-stage HPSS [Tachibana2010]



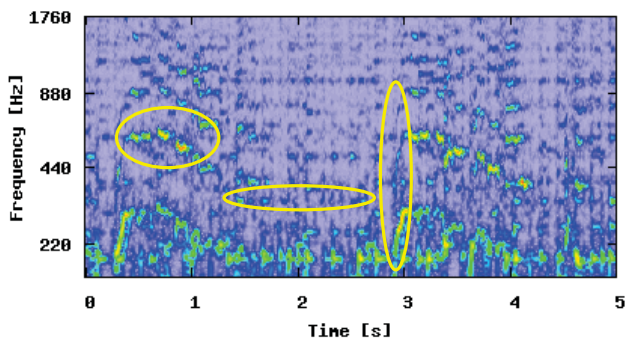
## Spectrogram Example

Original signal (from LabROSA dataset)



## Spectrogram Example

Voice-enhanced signal (by two-stage HPSS)

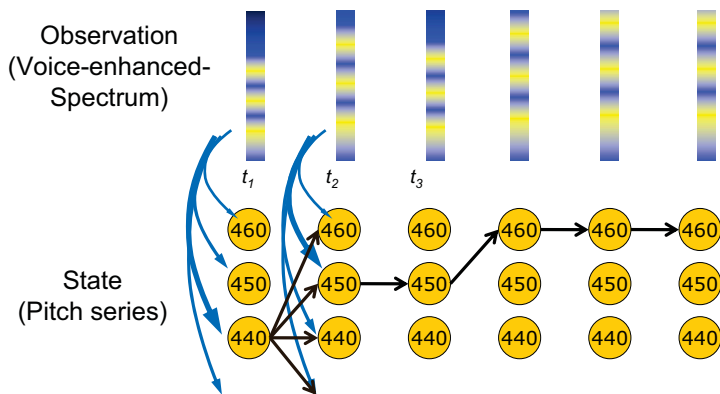


## Separation Examples

title	original	Extracted Vocal	Vocal Cancelled*	Genre
"tell me"				F, R&B
"Weekend"				F, Euro beat
"Dance Together"				M, Jazz
"1999"				M, Metal rock
"Seven little crows"				F, Nursery rhyme
"La donna è mobile" from Verdi's opera "Rigoletto"				M, Classical

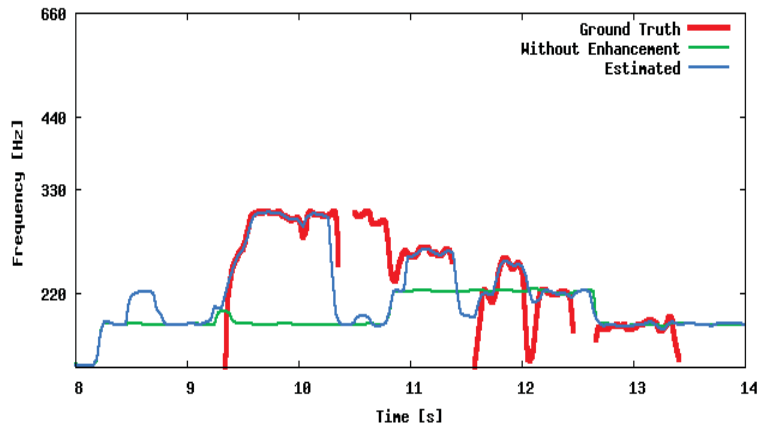
## Melody Tracking by DP [Tachibana2010]

- Estimating hidden states by dynamic programming



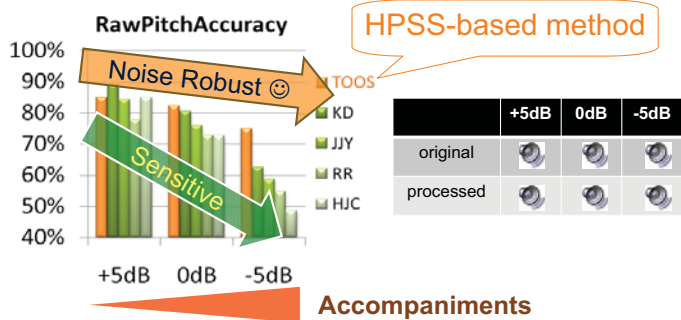
## Example of Melody Tracking

- train06.wav, distributed by LabROSA database



## Results in MIREX 2009

- Data: 379 songs, mixed in +5 dB, 0dB, and -5 dB.



Robustness to large singer-to-accompaniment ratio is greatly improved.

## Part III: Applications of HPSS to MIR Tasks

### III-3: Audio Genre Classification

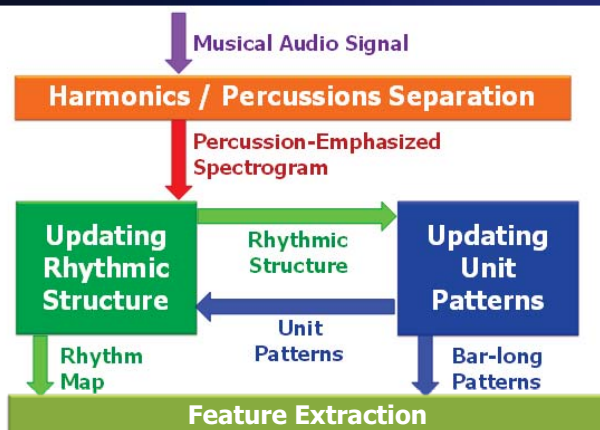
## Audio Genre Classification

- Task: estimate genre from music audio
  - Blues, classical, jazz, rock, ...
- Typical approach



- Example of features [Tzanetakis2001]
  - Timbral information (MFCC, etc.)
  - Melodic information
  - Statistics about periodicities: Beat histogram

## New Features I: Percussive Patterns



[Tsunoo2009]

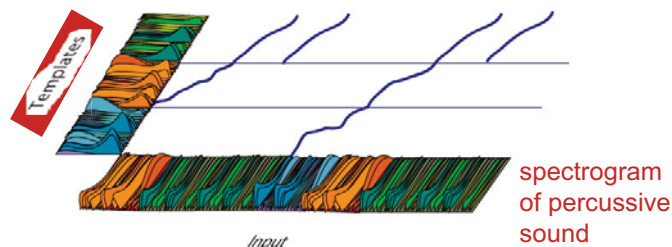
## Motivation for Bar-long Percussive Patterns

- Bar-long percussive patterns (temporal information) are frequently characteristic of a particular genre
- Difficulties
  - 1) Mixture of harmonic and percussive components
  - 2) Unknown bar-lines
  - 3) Tempo fluctuation
  - 4) Unknown multiple patterns

A A A A B A A A C C C C

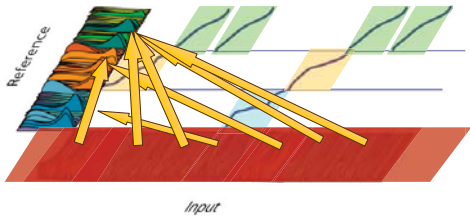
## Rhythmic Structure Analysis by One-pass DP algorithm

- Assume that correct bar-line unit patterns are given.
- Problem: tempo fluctuation and unknown segmentation
  - Analogous to continuous speech recognition problem
  - One-pass dynamic programming algorithm can be used to segment



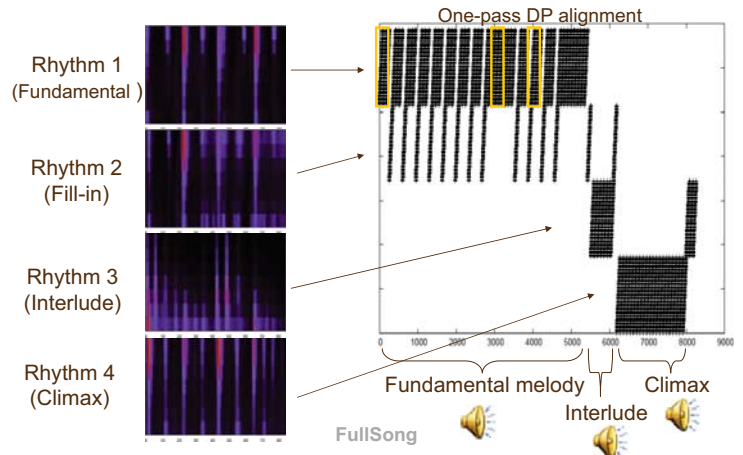
## Dynamic Pattern Clustering [Tsunoo2009]

- Actually, unit patterns also should be estimated.
  - Chicken-and-egg problem
  - Analogous to unsupervised learning problem
- Iterative algorithm based on *k*-means clustering
  - Segment spectrogram using one-pass DP algorithm
  - Update unit patterns by averaging segments
- Convergence is guaranteed mathematically



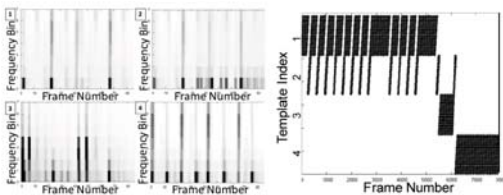
57

## Example of "Rhythm Map"

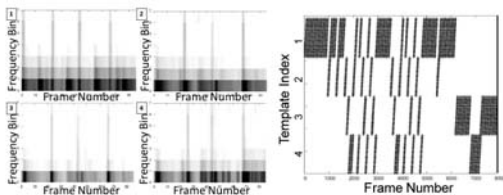


## Necessity of HPSS in Rhythm Map

With HPSS



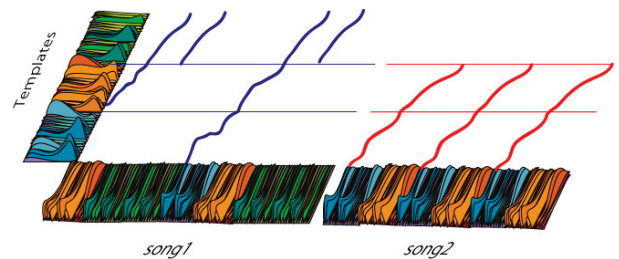
Without HPSS



Rhythm patterns and structures are not extracted without HPSS!

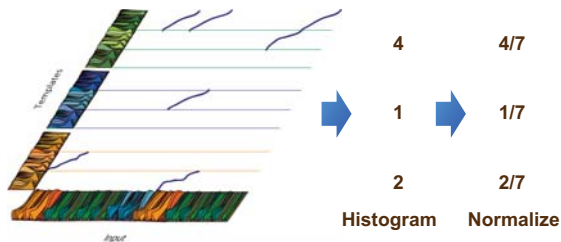
## Extracting Common Patterns to a Particular Genre

- Apply to a collection of music pieces
  - Alignment calculation by one-pass DP algorithm
    - Use same set of templates
  - Updating templates by *k*-means clustering
    - Use whole music collection of a particular genre
- Iteration



## Features and Classifiers

- Feature Vectors: Genre-pattern Occurrence Histogram (normalized)
- Classifier: Support Vector Machine (SVM)



61

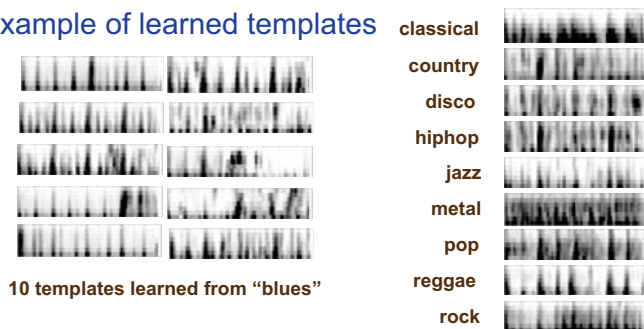
## Experimental Evaluation

- Dataset
  - (standard)
    - GTZAN dataset
    - 22050Hz sampling, 1ch
    - 30 seconds clips
    - 10 genres
      - {blues, classical, country, disco, hiphop, jazz, metal, pop, reggae, rock}
    - 100 songs per genre: total 1000 songs
  - (rhythm-intensive)
    - Ballroom dataset
    - 22050Hz sampling, 1ch
    - 30 seconds clips
    - 8 styles
      - {chacha, foxtrot, quickstep, rumba, samba, tango, viennese waltz, waltz}
    - 100 songs per style: total 800 songs
- Evaluation
  - 10-fold cross-validation
  - Classifier: linear SVM (toolkit "Weka" used)

## Extracted Percussive Patterns

- Pattern set
  - Divided the datasets into 2 parts and obtained 2 sets of 10 templates for each genre

- Example of learned templates



10 templates learned from "blues"

## Genre Classification Accuracy

- Percussive pattern feature only

Features [number of dim.]	GTZAN dataset	Ballroom dataset
Baseline (Random)	10.0%	12.5%
Rhythmic (from template set #1) [10/8]	43.6%	54.0%
Rhythmic (from template set #2) [10/8]	42.3%	55.125%

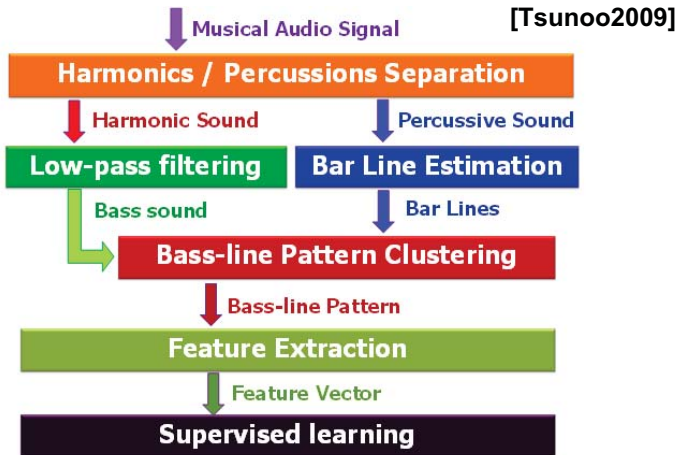
- Merged with timbral features

- Statistic features such as MFCC, etc. (68 dim.) [Tzanetakis 2008]
- Performed well on audio classification tasks in MIREX 2008

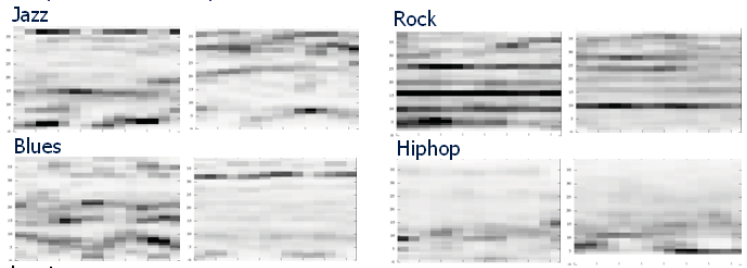
Features [number of dim.]	GTZAN dataset	Ballroom dataset
Existing (Timbre) [68]	72.4%	57.625%
Merged (from template set #1) [78/76]	76.1%	70.125%
Merged (from template set #2) [78/76]	76.2%	69.125%

Classification accuracy is improved by combining percussive pattern features.

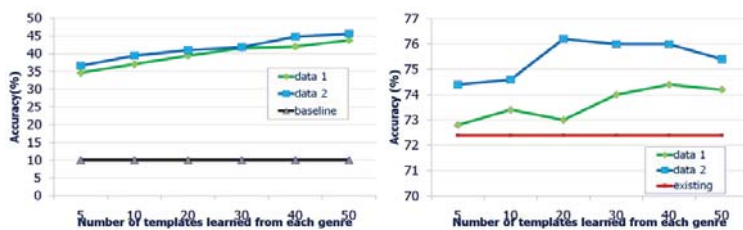
## New Features II: Bass-line Patterns



## Examples of Extracted Bass-line Patterns



## Genre Classification Accuracy



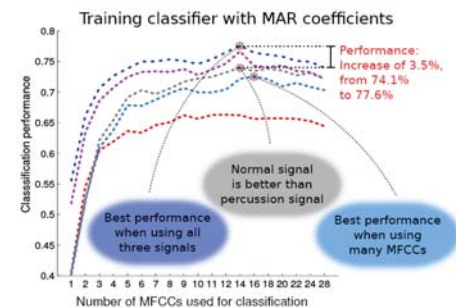
Classification accuracy with only bass-line features

Classification accuracy merged with timbre features

Features	GTZAN dataset	Ballroom dataset
Baseline (random classifier)	10.0%	10.0%
Only bass-line (400 dim.)	42.0%	44.8%
Existing (timbre, 68 dim.)	72.4%	72.4%
Merged (468 dim.)	74.4%	76.0%

## Another Application of HPSS [Rump2010]

- Autoregressive MFCC Model applied to Genre Classification
- HPSS increases the number of channels mono -> three (original, harmonic, percussive) and improves performance



## Conclusions

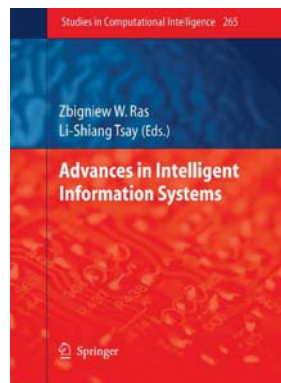
- Source separation techniques used to MIR
  - F0-based harmonic separation
  - Non-negative matrix factorization or PLCA
  - Sinusoid + Noise model
  - Harmonic/percussive sound separation
- Source separation is useful
  - To enhance specific components
  - To increase the number of channels and the dimension of feature vectors
  - To generate new features

## Future Works

- Application of source separation to other MIR tasks
  - Cover song identification, audio music similarity,...
- Improvement of separation performance itself by exploiting musicological knowledge
- Using spatial (especially stereo) information
  - Current works are limited to monaural separation
- Feature weighting technique for overcoming errors due to imperfect source separation

## Reference Book Chapter

- *Advances in Music Information Retrieval, ser. Studies in Computational Intelligence, Z. W. Ras and A. Wiczorkowska, Eds. Springer, 274*
  - N. Ono, K. Miyamoto, H. Kameoka, J. Le Roux, Y. Uchiyama, E. Tsunoo, T. Nishimoto and S. Sagayama, "Harmonic and Percussive Sound Separation and its Application to MIR-related Tasks," pp.213-236



## Available Separation Softwares

- Harmonic Percussive Sound Separation (HPSS)
  - <http://www.hil.t.u-tokyo.ac.jp/software/HPSS/>
- ICA Central: Early software restricted to mixtures of two sources
  - <http://www.tsi.enst.fr/icacentral/algos.html>
- SiSEC Reference Software: Linear modeling-based software for panned or recorded mixtures
  - <http://sisec2008.wiki.irisa.fr/tiki-index.php?page=Under-determined+speech+and+music+mixtures>
- QUAERO Source Separation Toolkit: Modular variance-modeling based software implementing a range of structures: GMM, NMF, source-filter model, harmonicity, diffuse mixing, etc
  - To be released Fall 2010: watch the music-ir list for an announcement!



## Advertisement: LVA/ICA 2010

**LVA ICA 2010<sup>th</sup>** Celebrating the anniversary  
September 27-30, St. Malo, France

**180 degree panoramic sea view**  
**81 contributed papers**  
**42 liters of coffee**  
**4 keynotes:**  
Hervé Lohman, University of Nice, France  
Christophe Mallat, Ecole Polytechnique, France  
Mark Srinivasan, University of Glasgow, UK  
Eric Sennou, Tel Aviv University, Israel

**2 panel sessions:**  
Harmonic, vocal, music and remaining challenges  
The future of blind variable analysis and signal separation.

**2 hours of private visit to Mont-Saint Michel**  
... 1 unique conference !

<http://lva2010.inria.fr>

- **LVA/ICA 2010** is held will be held in St. Malo, France on September 27-30, 2010.
- **More than 20 papers** on music and audio source separation will be presented.

## References

### ■ Singer/Instrument Identification

- H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata and H. Okuno, "Singer Identification Based on Accompaniment Sound Reduction and Reliable Frame Selection," *Proc. ISMIR*, 2005.
- M. Goto, "A real-time music-scene description system: predominant-F0 estimation," *Speech Communication*, vol. 43, no. 4, pp. 311–329, 2004.
- A. Mesaros, T. Virtanen and A. Klapuri, "Singer identification in polyphonic music using vocal separation and pattern recognition methods," *Proc. ISMIR*, pp. 375-378, 2007.
- M. Ryyanen and A. Klapuri, "Transcription of the Singing Melody in Polyphonic Music", *Proc. ISMIR*, 2006.
- T. Kitahara, M. Goto, K. Komatani, T. Ogata and H. G. Okuno, "Instrument identification in polyphonic music: feature weighting to minimize influence of sound overlaps," *EURASIP Journal on Applied Signal Processing*, vol. 2007, 2007, article ID 51979.

## References

### ■ Audio Tempo Estimation

- M. Alonso, G. Richard and B. David, "Accurate tempo estimation based on harmonic + noise decomposition," *EURASIP Journal on Advances in Signal Processing* Volume 2007 (2007), Article ID 82795
- P. Chordia and A. Rae, "Using Source Separation to Improve Tempo Detection," *Proc. ISMIR*, pp. 183-188, 2009.

### ■ Related Works to H/P Separation

- C. Uhle, C. Dittmar, and T. Sporer, "Extraction of drum tracks from polyphonic music using independent subspace analysis," *Proc. ICA*, pp. 843-847, 2003.
- M. Helen and T. Virtanen, "Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine," *Proc. EUSIPCO*, Sep. 2005.
- L. Daudet, "A Review on Techniques for the Extraction of Transients in Musical Signals," *Proc. CMMR*, pp. 219-232, 2005.
- O. Dikmen, A. T. Cemgil, "Unsupervised Single-channel Source Separation Using Bayesian NMF," *Proc. WASPAA*, pp. 93-96, 2009.

## References

### ■ Harmonic/Percussive Sound Separation

- K. Miyamoto, H. Kameoka, N. Ono and S. Sagayama, "Separation of Harmonic and Non-Harmonic Sounds Based on Anisotropy in Spectrogram," *Proc. ASJ*, pp.903-904, 2008. (in Japanese)
- N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka and S. Sagayama, "Separation of a Monaural Audio Signal into Harmonic/Percussive Components by Complementary Diffusion on Spectrogram," *Proc. EUSIPCO*, 2008.
- N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka and S. Sagayama, "A Real-time Equalizer of Harmonic and Percussive Components in Music Signals," *Proc. of ISMIR*, pp.139-144, 2008.
- N. Ono, K. Miyamoto, H. Kameoka, J. Le Roux, Y. Uchiyama, E. Tsunoo, T. Nishimoto and S. Sagayama, "Harmonic and Percussive Sound Separation and its Application to MIR-related Tasks," *Advances in Music Information Retrieval, ser. Studies in Computational Intelligence, Z. W. Ras and A. Wiczorkowska, Eds. Springer*, 274, pp.213-236, Feb., 2010.

## References

### ■ Applications of HPSS to MIR Tasks

- Y. Ueda, Y. Uchiyama, T. Nishimoto, N. Ono and S. Sagayama, "HMM-Based Approach for Automatic Chord Detection Using Refined Acoustic Features," *Proc. ICASSP*, pp.5518-5521, 2010.
- J. Reed, Y. Ueda, S. M. Siniscalchi, Y. Uchiyama, S. Sagayama, C. -H. Lee, "Minimum Classification Error Training to Improve Isolated Chord Recognition," *Proc. ISMIR*, pp.609-614, 2009.
- H. Tachibana, T. Ono, N. Ono and S. Sagayama, "Melody Line Estimation in Homophonic Music Audio Signals Based on Temporal-Variability of Melodic Source," *Proc. ICASSP*, pp.425-428, 2010.
- H. Rump, S. Miyabe, E. Tsunoo, N. Ono and S. Sagayama, "On the Feature Extraction of Timbral Dynamics," *Proc. ISMIR*, 2010.



## References

### ■ Applications of HPSS in MIR Tasks

- E. Tsunoo, N. Ono and S. Sagayama, "Rhythm Map: Extraction of Unit Rhythmic Patterns and Analysis of Rhythmic Structure from Music Acoustic Signals," *Proc. ICASSP*, pp.185-188, 2009.
- E. Tsunoo, G. Tzanetakis, N. Ono and S. Sagayama, "Audio Genre Classification Using Percussive Pattern Clustering Combined with Timbral Features," *Proc. ICME*, pp.382-385, 2009.
- E. Tsunoo, N. Ono and S. Sagayama, "Musical Bass-Line Pattern Clustering and Its Application to Audio Genre Classification," *Proc. ISMIR*, pp.219-224, 2009.
- E. Tsunoo, T. Akase, N. Ono and S. Sagayama, "Music Mood Classification by Rhythm and Bass-line Unit Pattern Analysis," *Proc. ICASSP*, pp.265-268, 2010.

