Source Separation Methods for under-determined sound mixtures

Mathieu Lagrange
Analyse / Synthèse Team, IRCAM
Mathieu.lagrange@ircam.fr

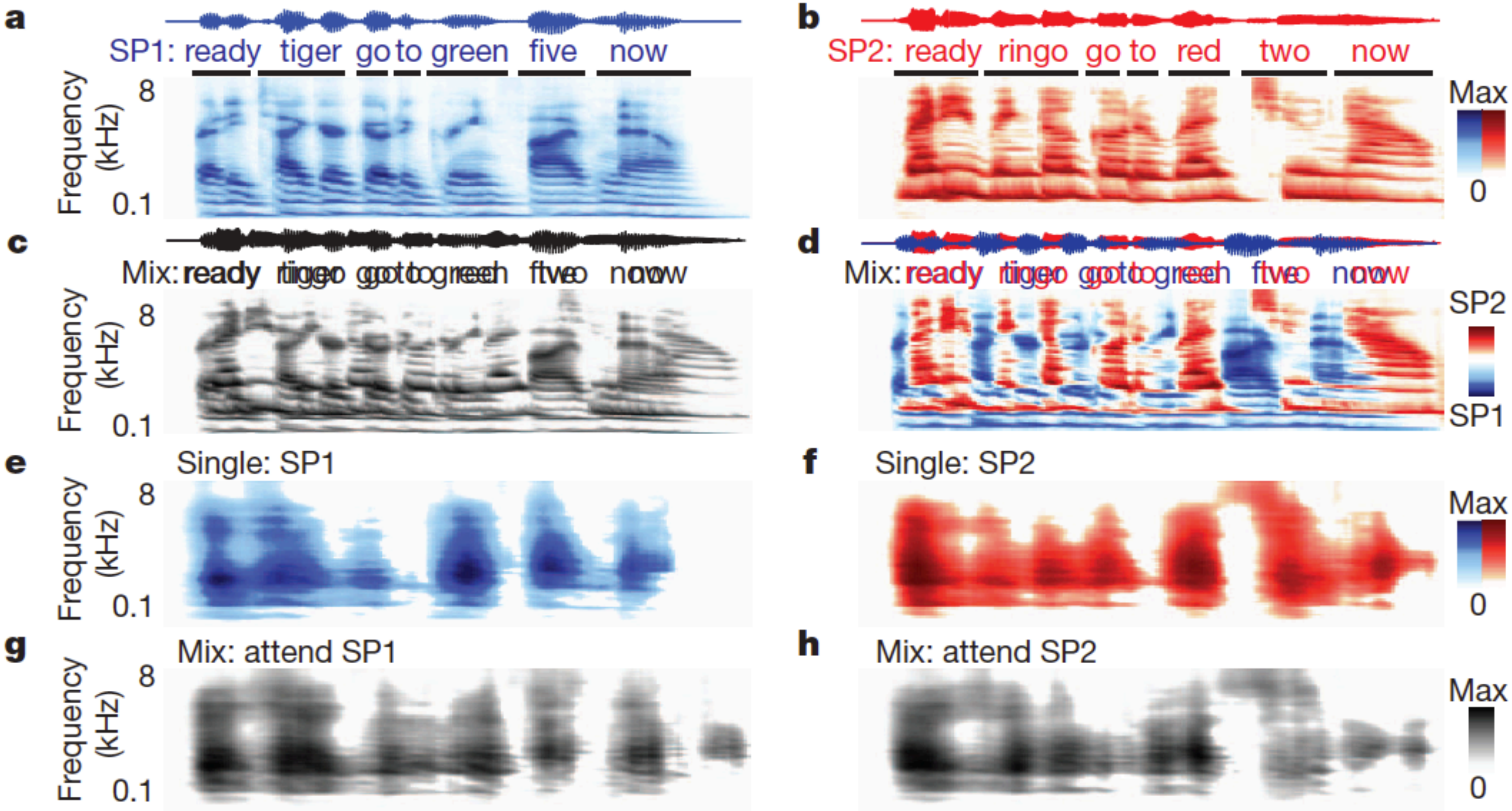
ATIAM

Analysis of Sound Mixture

- We aim at performing
  - Auditory Scene Analysis
  - Computationally
  - But like human do
  - Humans focus on one source
- Task
  - Source separation ?
  - Source classification ?
  - Something in-between ?
    - What then ?

Human separate, really ?

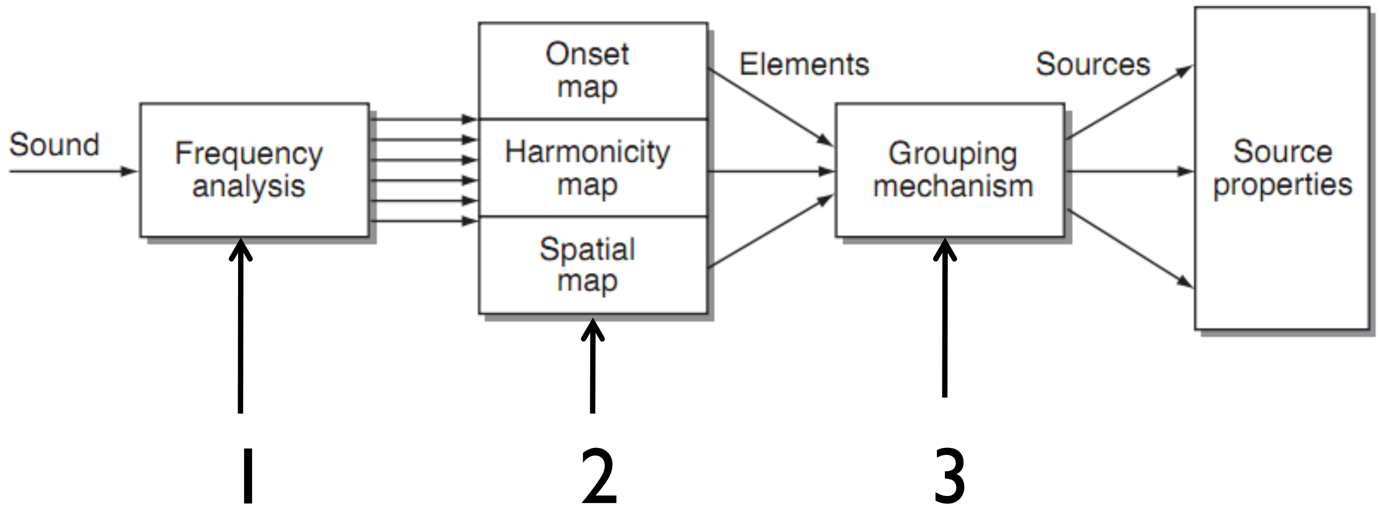
- It seems so:



(Fig. from Mesgarani Nature'12

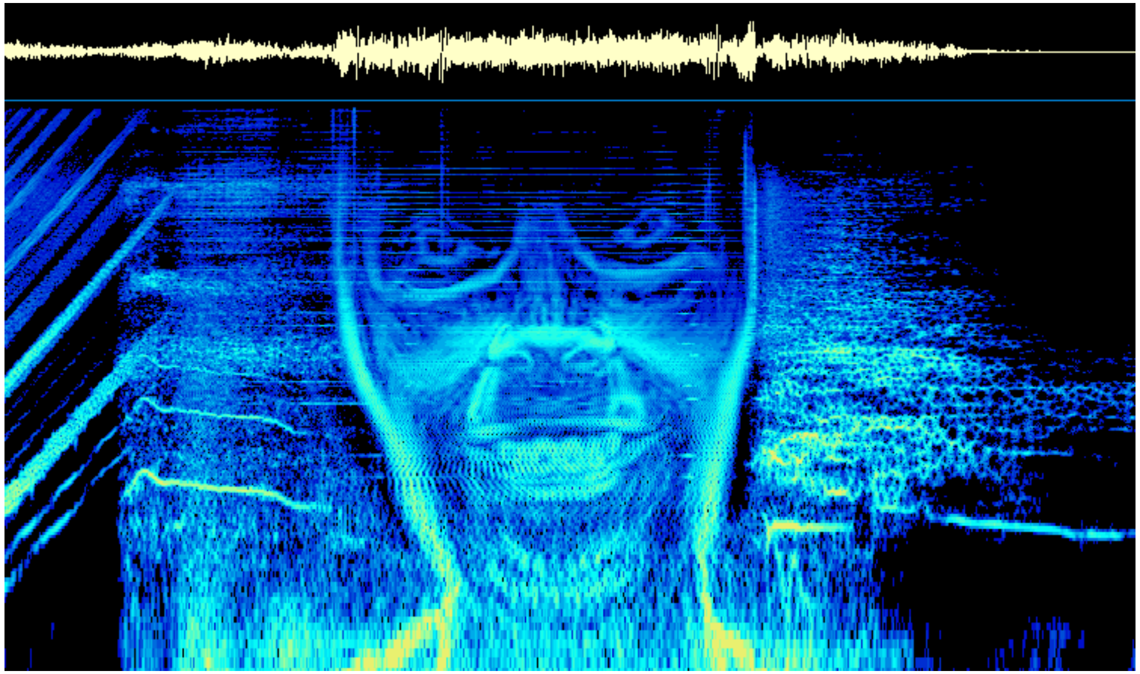
Computational ASA (CASA)

- How do people analyze sound mixtures ?
  - break mixture into small elements (in time-freq)
  - elements are grouped in to sources using cues
  - sources have aggregate attributes



1. Frequency Analysis (FA)

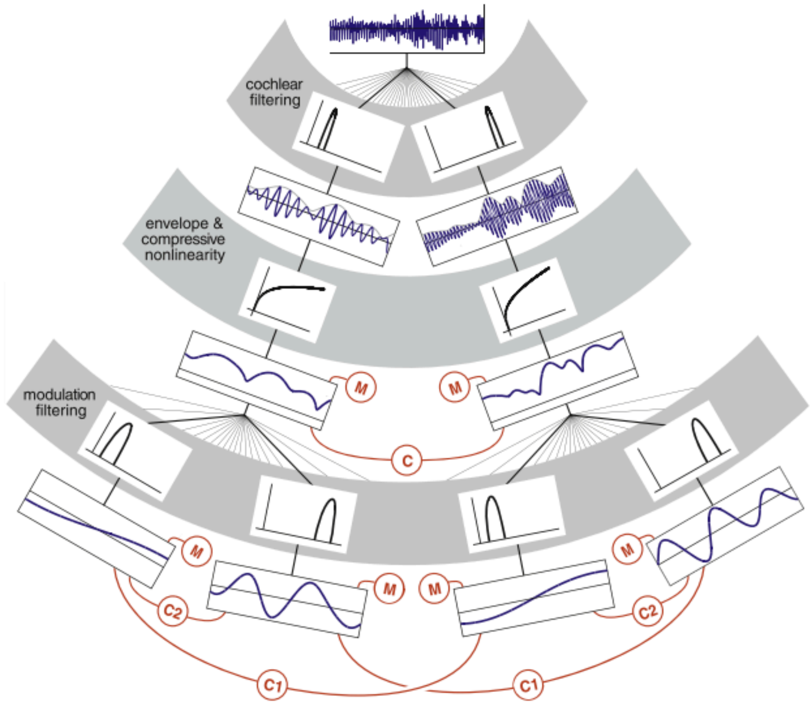
- Fourier based analysis
  - The Short-Term Fourier Transform (STFT)
  - By far the most widely used



(Fig. from Aphex Twin)

1. Frequency Analysis (FA)

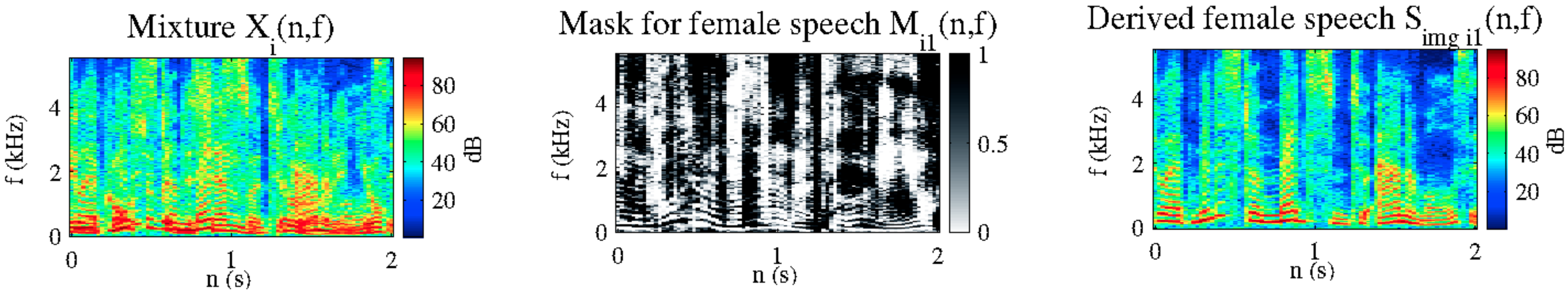
- Perception inspired front-ends
  - Like the Correlogram
  - Designed to imitate what is known about the physiology of the inner ear
  - Usually composed of
    - A cascade of filterbanks
    - Interleaved with non linear operators



(Fig. from [McDermott11])

How to use FA for grouping ?

- Source Separation: a masking problem
- Goal: find a mask M that retrieves one source when used to filter a given time-frequency representation.



$$\boxed{\hat{\mathbf{S}}_n(r,k) = \mathbf{M}_{mn}(r,k) \circ \mathbf{X}_m(r,k)}$$

$\circ$ is the Hadamard (element-wise) product

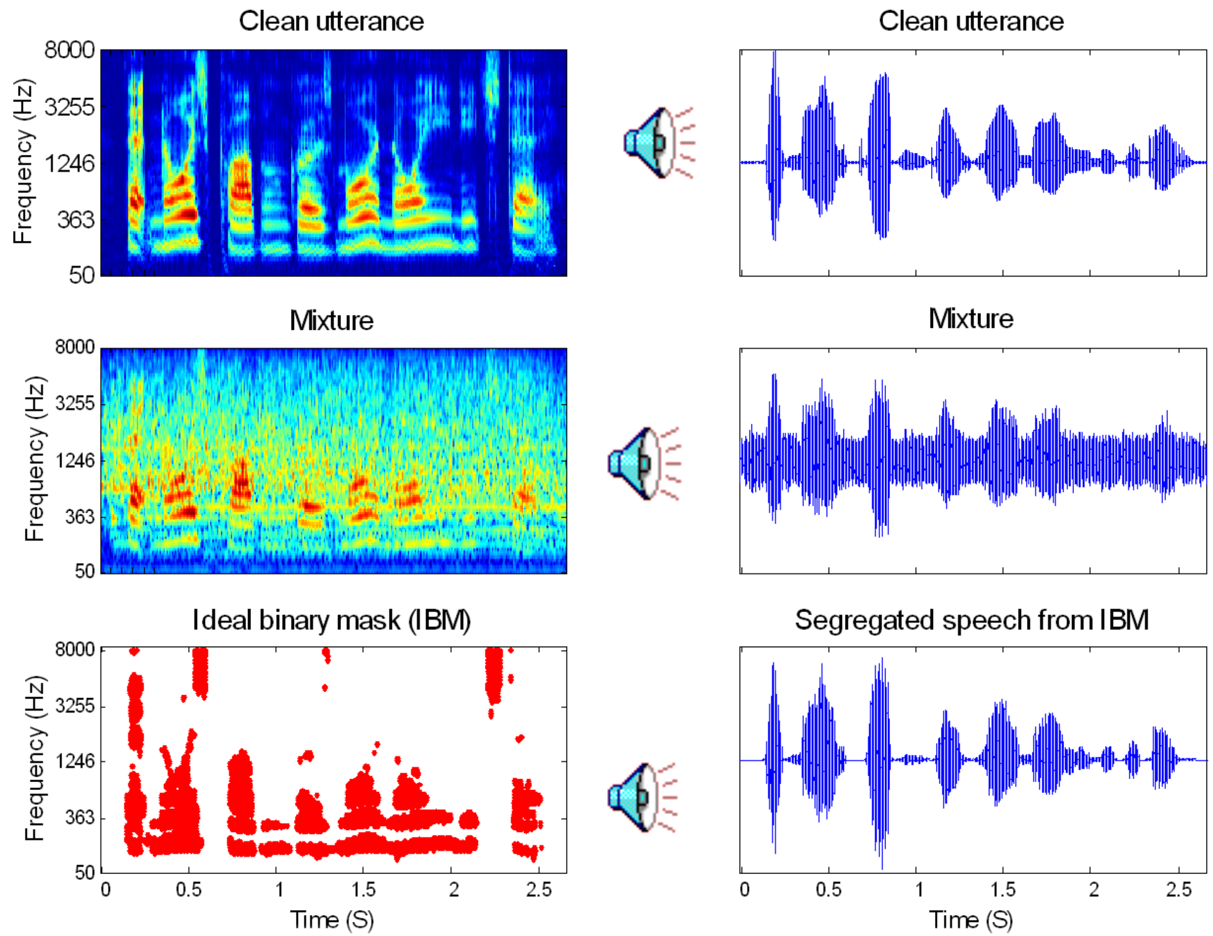
- What about the phase ?
  - Keep the one of the mixture

The Ideal Binary Mask (IBM)

- The IBM
  - Is an “oracle” separation method, that is we know something (everything ?) we need for separating the sources.

- It provides
  - An upper bound for masking based approaches
  - A way to understand issues with the front end
    - Time/frequency resolution tradeoff
    - Issues with the phase

Demonstration of the IBM

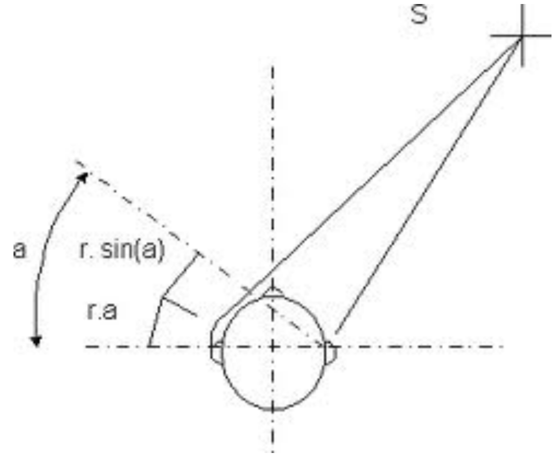


(Fig. from [Hanse 04])

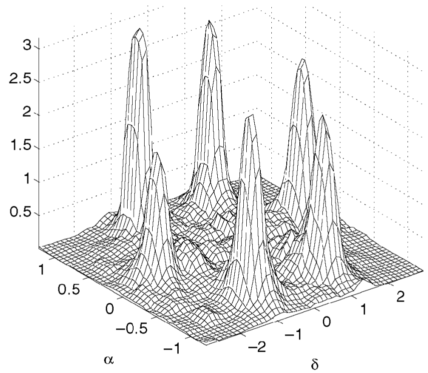
- Utterance: "That noise problem grows more annoying each day"
- Interference: Crowd noise with music (0 SNR)

2. Cues (Binaural Case)

- Have spatial location cues
  - Termed Interchannel or Interaural
  - Phase and Intensity Differences: IPD and IID
  - Warning: profesionaly mastered audio does not preserve them.



- DUET (Degenerate Unmixing Estimation Technique) [Yilmaz&Rickard04]
  - Histogram of IPD and IID
  - Binary Mask created by selecting bins around histogram peaks.

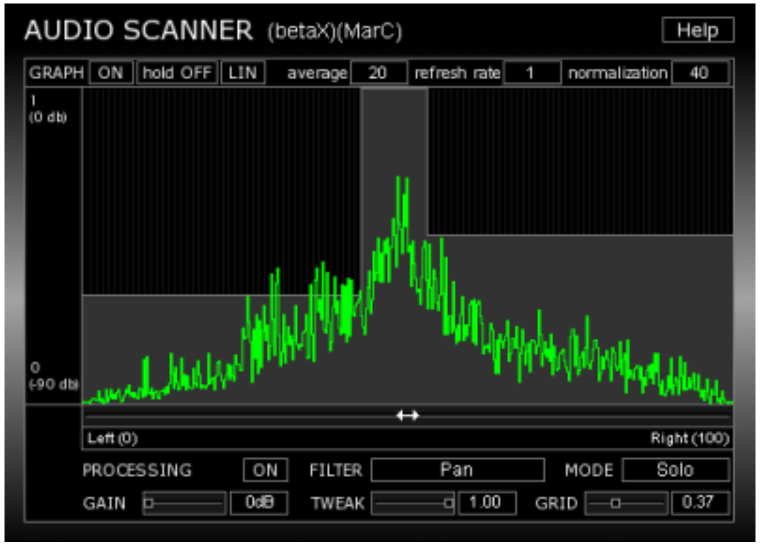


(Fig. from [Yilmaz&Rickard04])

[Yilmaz&Rickard04] Ö. Yilmaz and S. Rickard. Blind Separation of Speech Mixtures via Time-Frequency Masking. *IEEE Trans. on Signal Processing*. Vol. 52(7), July 2004

2. Cues (Binaural Case)

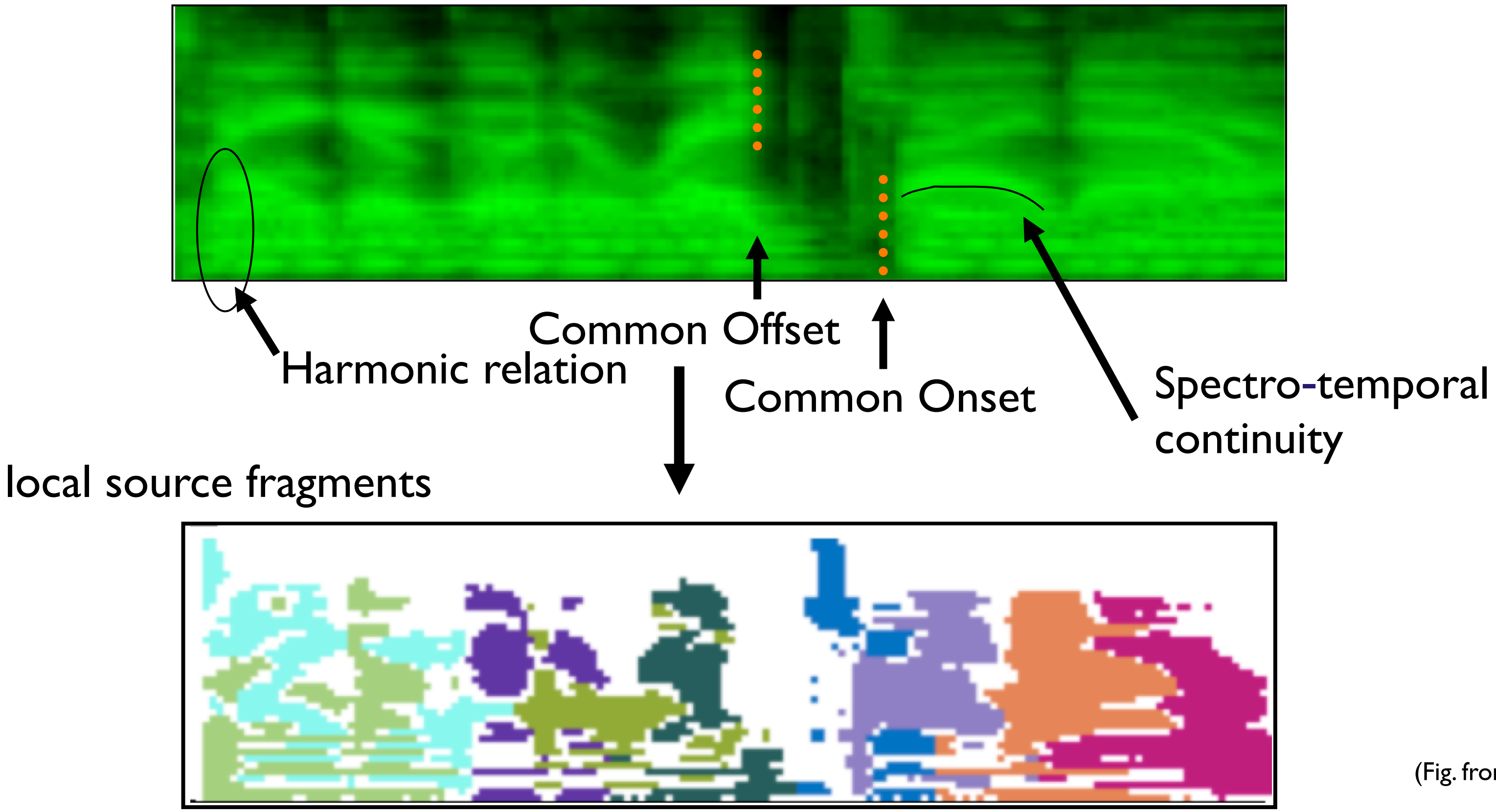
- Human-assisted time-frequency masking [Vinyes06]
  - Human-assisted selection of the time-frequency bins out of the DUET-like histogram for creating the unmixing mask
  - Implementation as a VST plugin ("Audio Scanner")




[Vinyes06] M. Vinyes, J. Bonada and A. Loscos. Demixing Commercial Music Productions via Human-Assisted Time-Frequency Masking. *120th AES convention,* Paris, France, 2006.

2. Cues (Monaural case)

- Most ASA cues can be considered
- But the most important cue is pitch



(Fig. from [Barker 11])

2. Cues (Monaural case)

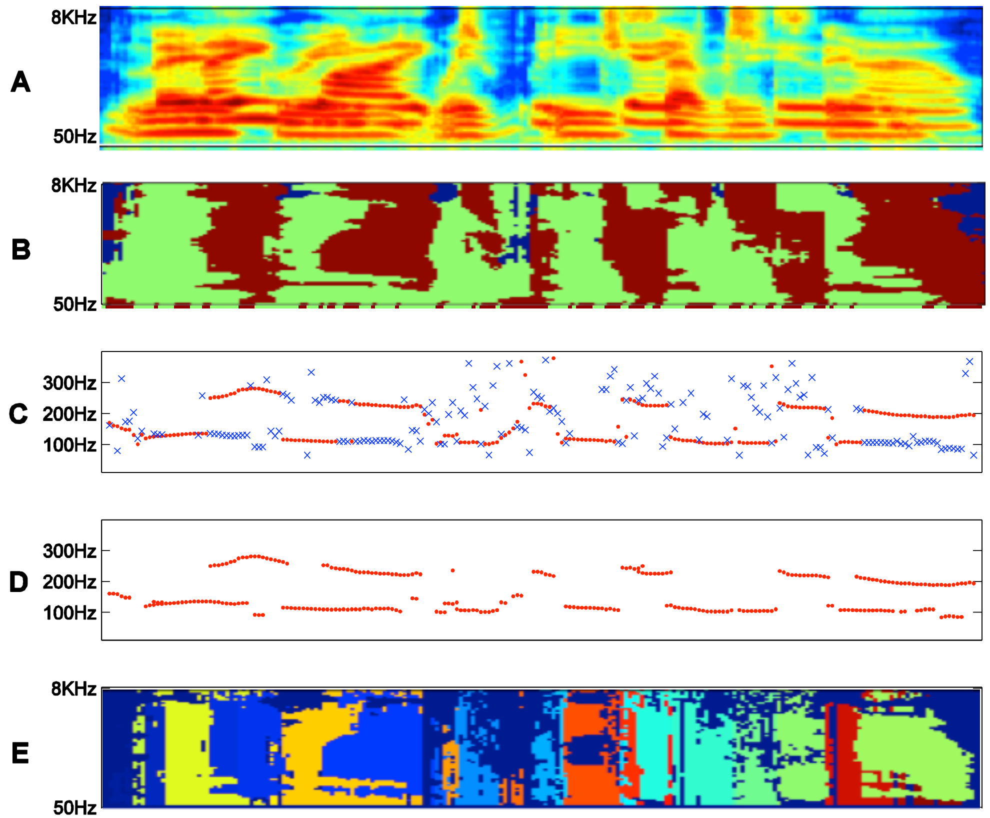
Filterbank output

`Ideal' segmentation

Pitch candidates

Pitch tracking

Harmonic fragments



(Fig. from [Barker 11])

3. Grouping

- Bottom up approaches
  - Statistical (Blind) approaches (NMF)
  - Clustering approaches based on ASA cues (CASA)
- Top down approaches
  - Model based approach
  - Dictionary based approach
- Combination between the two
  - Model based approach

Nonnegative Matrix Factorization (NMF)

- Given a nonnegative matrix V of dimensions FxN, NMF is the problem of finding a factorization

$$\mathbf{V} \approx \mathbf{WH}$$

- where **W** and **H** are nonnegative matrices of dimensions FxK and KxN, respectively.

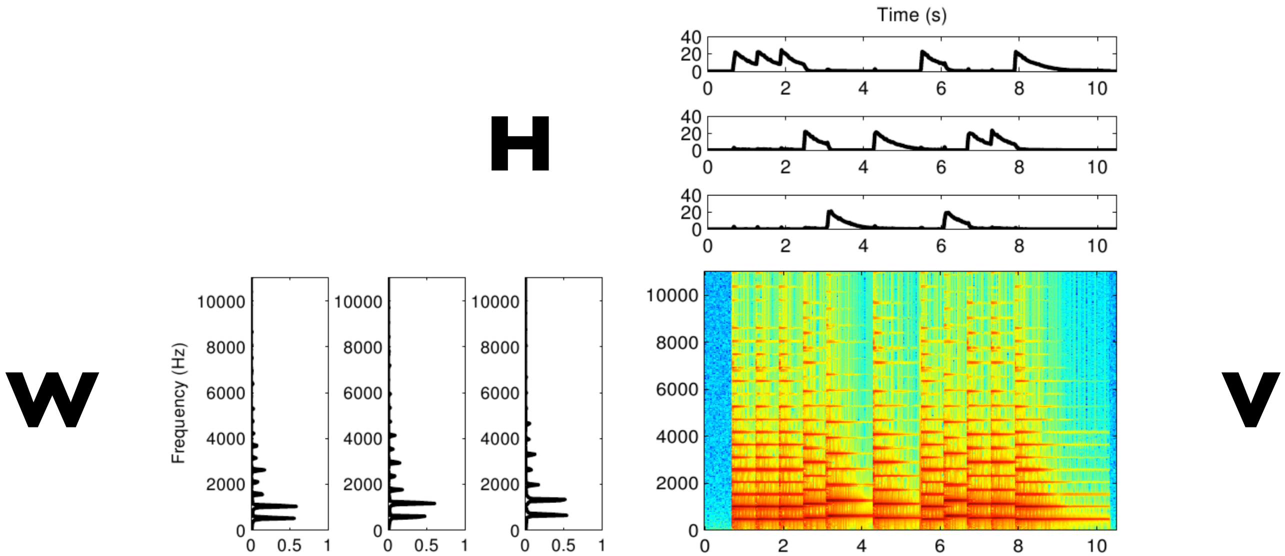
- Use for transcription:
  - P. Smaragdis and J.C. Brown. Non-Negative Matrix Factorization for Polyphonic Music Transcription. *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA),* New Paltz, USA, 2003.

- Use for separation:
  - B. Wang and M. D. Plumbley. Musical Audio Stream Separation by Non-Negative Matrix Factorization. *Proc. UK Digital Music Research Network (DMRN) Summer Conf.,* 2005.

NMF

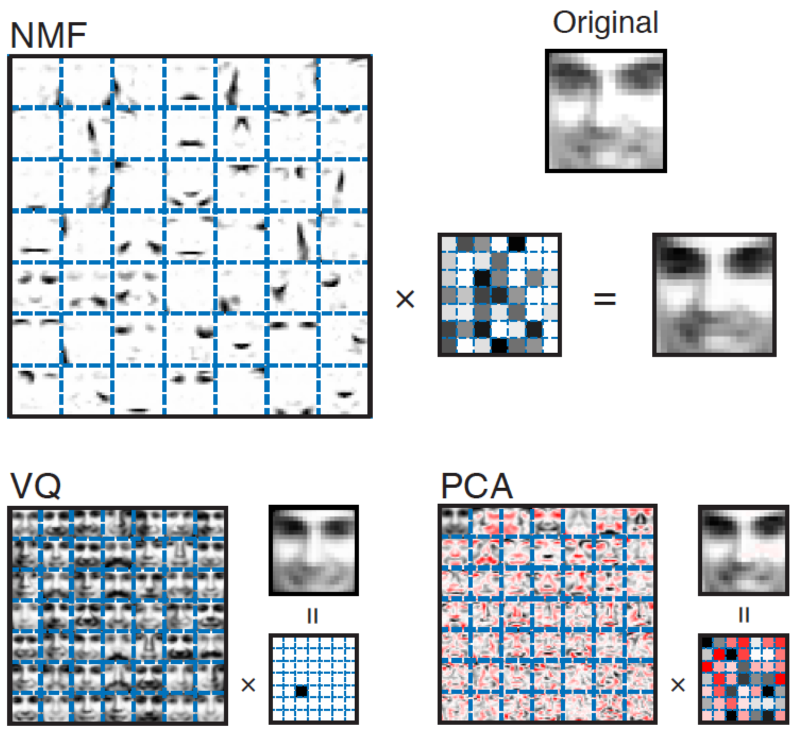
- Along VQ, PCA or ICA, NMF provides an unsupervised linear representation of data

$\mathbf{v}_n$	$\approx$	$\mathbf{W}$	$\mathbf{h}_n$
data vector		"explanatory variables" "basis", "dictionary" "patterns"	"regressors" "expansion coefficients" "activation coefficients"



NMF for Vision

- By representing signals as a sum purely additive, non- negative sources, we get a parts-based representation [Lee'99]



[Lee'99] Lee and Seung, Learning the parts of objects by nonnegative matrix factorization, Nature, 1999, 41

Update Rules for NMF

- Multiplicative (Lee & al)
  - Minimize a cost function with positivity constraints

$$||A - B||^2 = \sum_{ij} (A_{ij} - B_{ij})^2$$

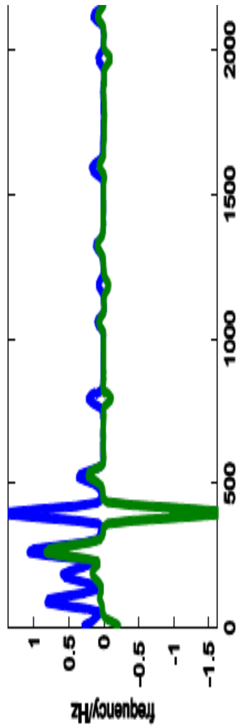
  - Update Rules

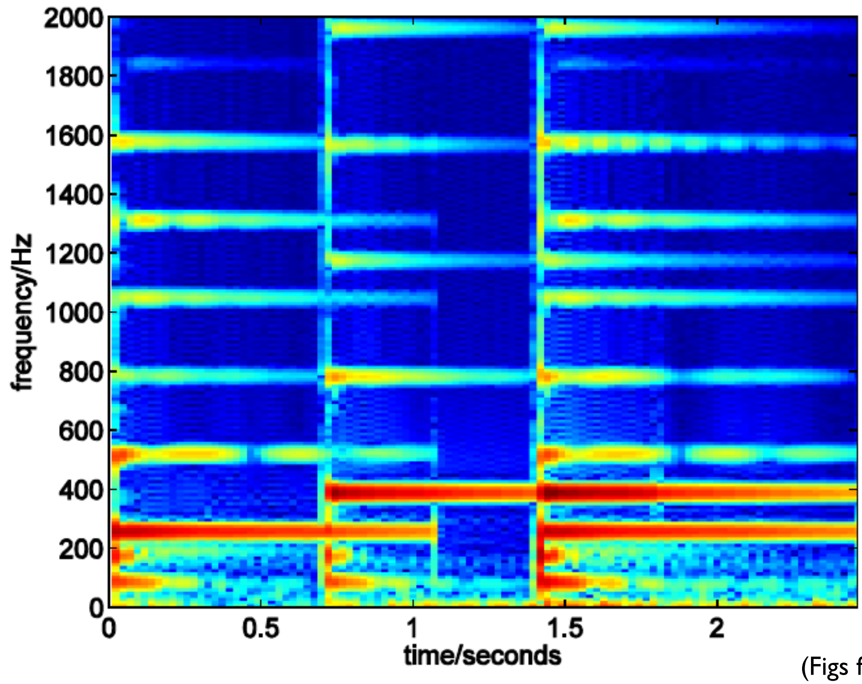
$$H_{a\mu} \leftarrow H_{a\mu} \frac{(W^T V)_{a\mu}}{(W^T W H)_{a\mu}} \qquad W_{ia} \leftarrow W_{ia} \frac{(V H^T)_{ia}}{(W H H^T)_{ia}}$$

  - Theorem: under the update rules, the cost function is
    - Non increasing
    - Invariant iif @ stationary point

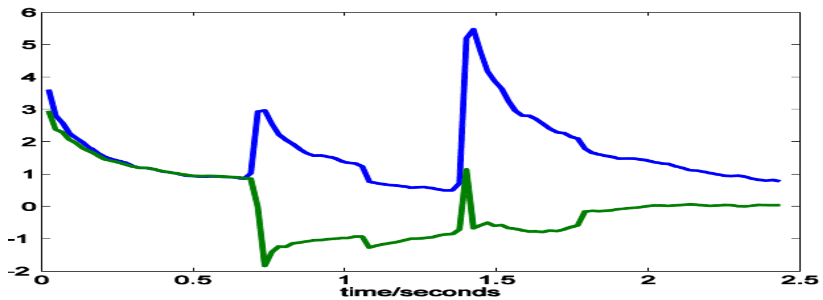
[Lee'01] Lee and Seung, Algorithms for Non-negativeMatrix Factorization, Nips, 2001

ICA on spectrograms

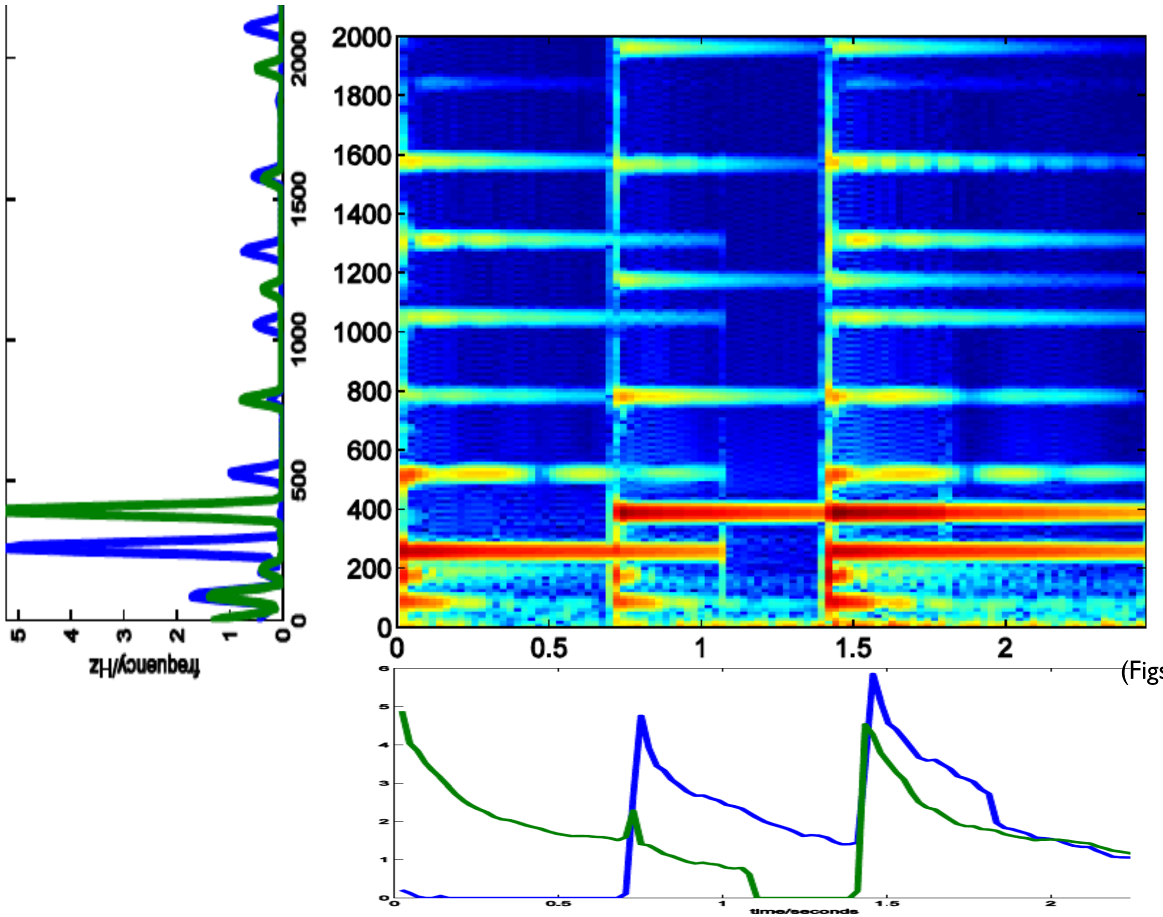




(Figs from Virtanen)



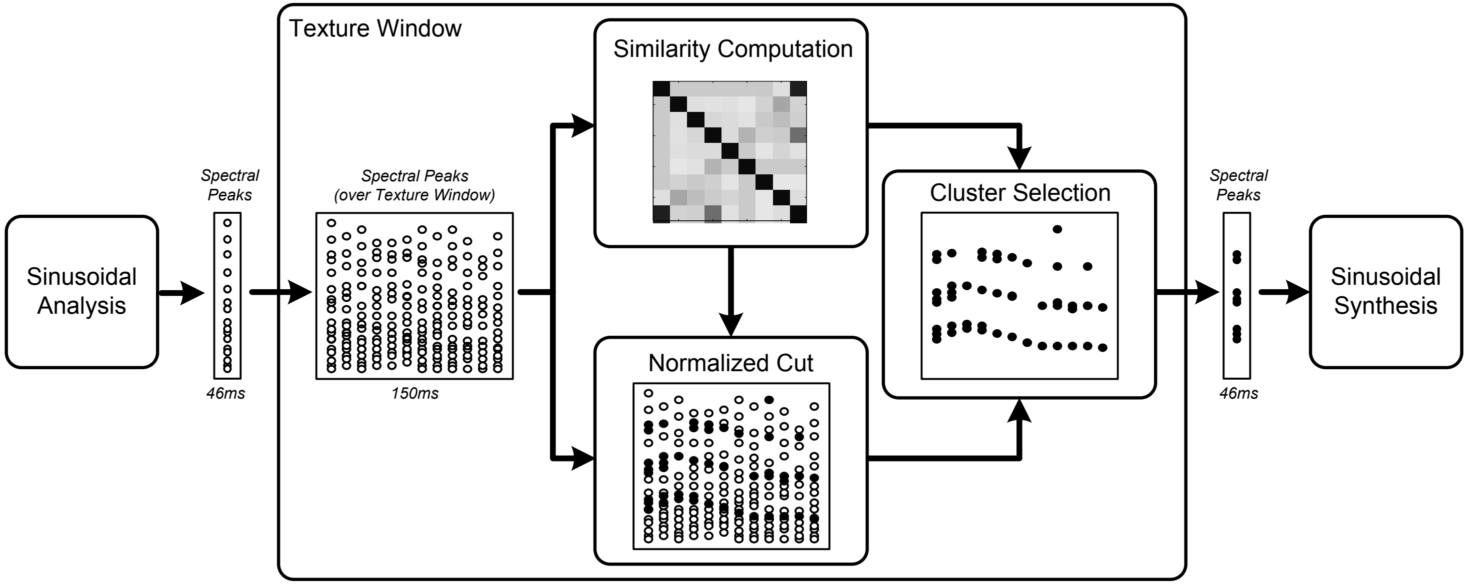
NMF on spectrograms



(Figs from Virtanen)

CASA

- How can we use the different cues ?
  - Earlier approach: consider the cues in sequence.
  - Sequentiality is brittle due to the propagation of errors
- All at once



Top down approaches

- Prior knowledge can be represented as an abstract model of some events of interest
  - Recognition:
    - Example: GMM models of spoken digits like in speech recognition
    - In this case, the background can be dealt with numerous approaches
      - Noisy training
      - Multi-condition training
  - Separation:
    - Example: separation of the singing voice in a music signal
    - Need model for
      - the singing voice
      - The music

(Fig. from [Barker 11])

GMM – Based Source Separation

- Given a mixture

$$x(n) = v(n) + m(n)$$

- Represented in the spectral domain

$$X_t(f) = V_t(f) + M_t(f)$$

- Following simple algebra

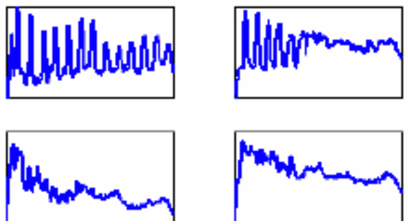
$$\boxed{p(V_t) = \sum_i \omega_{v,i} N\left(V_t; \bar{0}, \Sigma_{v,i}\right)}$$

Local Power Spectral Density (PSD) →

$$\Sigma_{v,i} = \begin{pmatrix} \sigma^2_{v,i}(1) & 0 & \cdots & 0 \\ 0 & \sigma^2_{v,i}(2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2_{v,i}(F) \end{pmatrix}$$

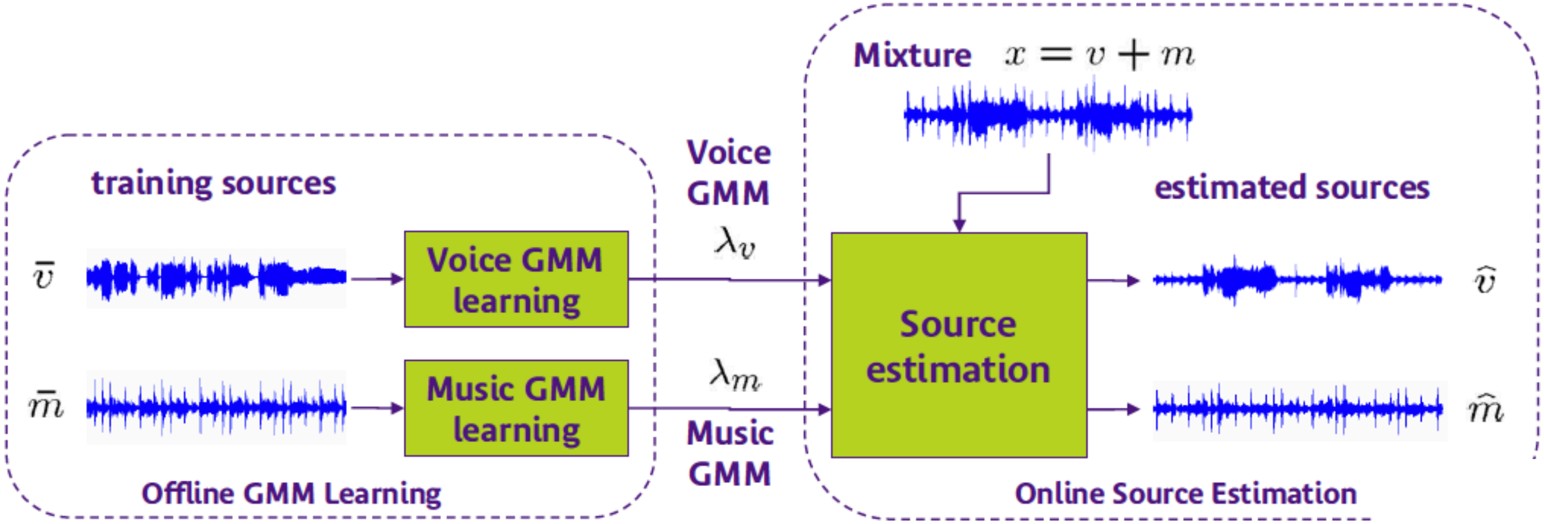
Voice GMM

$$\lambda_v = \left\{\omega_{v,i}, \Sigma_{v,i}\right\}_i$$



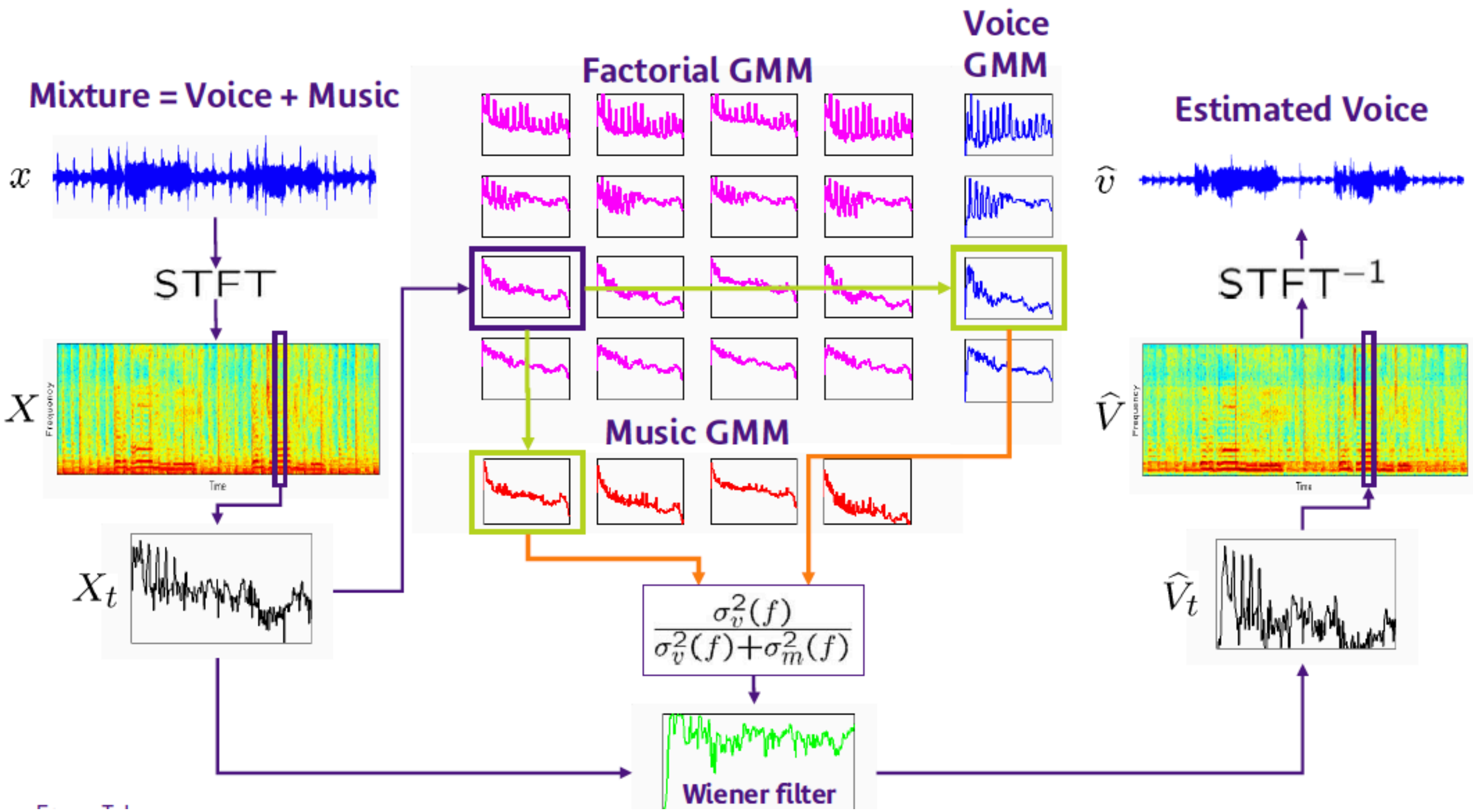
(Fig. from [Ozerov 05])

GMM – Based Source Separation



(Fig. from [Ozerov 05])

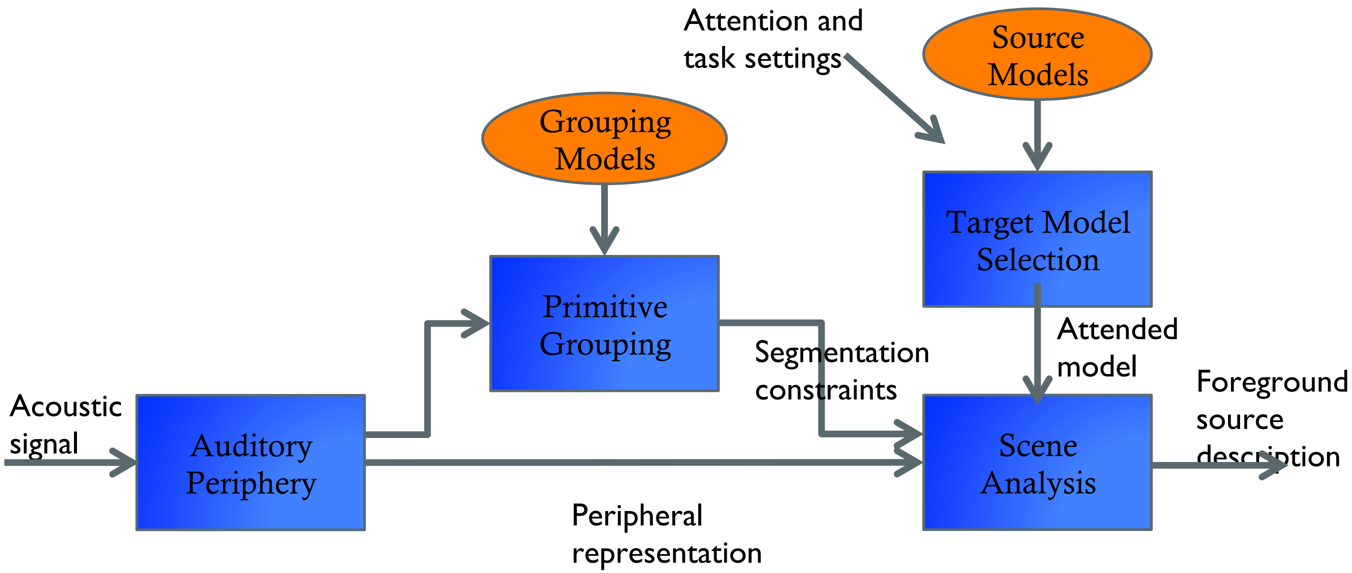
GMM – Based Source Separation



(Fig. from [Ozerov 05])

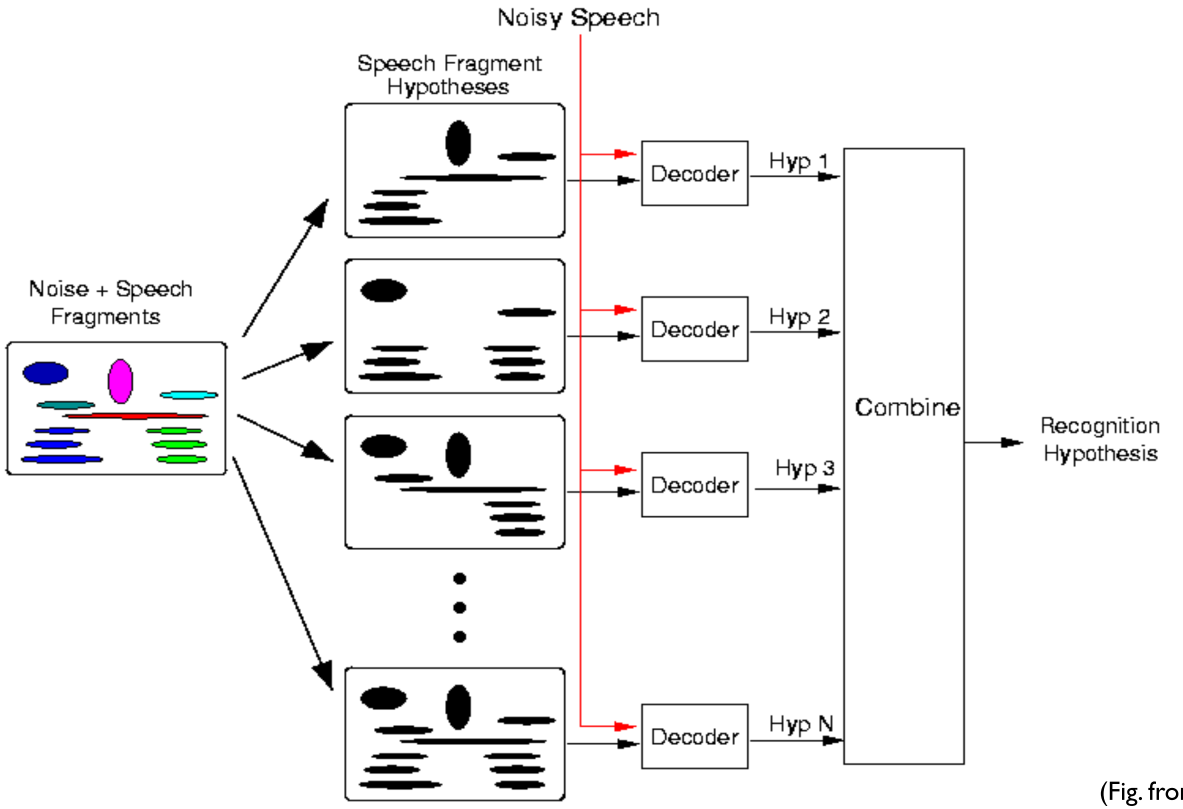
Combining Bottom-up and Top-Down

- Combining bottom up and top down approaches is
  - the dream goal
  - Is difficult



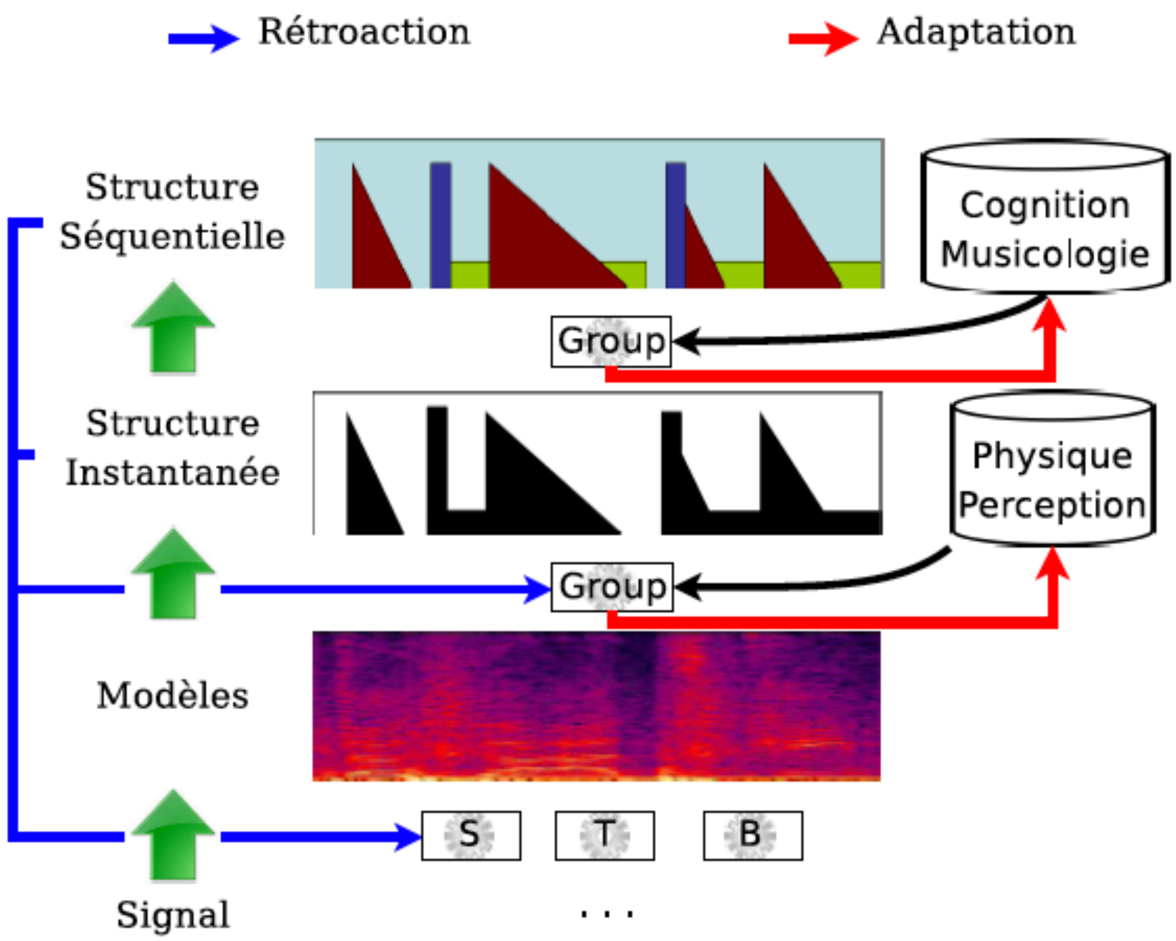
Combining Bottom-up and Top-Down

- One good example
  - Fragment-based spoken digit decoding
  - A simple (but terribly inefficient) implementation:



(Fig. from [Barker 11])

To summarize



Live coding in Matlab

- You can find the source here:
  - http://recherche.ircam.fr/equipes/analyse-synthese/lagrange/teaching/atiam11/coursAtiam2011Ibm.m
  - http://recherche.ircam.fr/equipes/analyse-synthese/lagrange/teaching/atiam11/coursAtiam2011Nmf.m
- You will need some external dependencies, web locations are provided in the code
- The code uses cell mode, please look at the Matlab documentation for usage

Research question (Master Subject)

- Can those computational frameworks such as NMF be considered for implementing important aspects of ASA ?
- Proposition: consider Semi Supervised NMF for implementing the Old+New heuristic
  - ON rationale: remove what we can infer from the scene, and model the remaining
  - Semi Supervised NMF:
    - X = FG + HU
      - F: prior knowledge
      - H: model new events
- Reference:
  - Supervised and Semi-Supervised Separation of Sounds from Single-Channel Mixtures [Smaragdis 07] http://www.merl.com/reports/docs/TR2007-062.pdf

CASA for singer similarity

- Aim: discover an application of CASA for MIR
- Testbed: Music similarity by singer
  - 2 songs are defined as similar if they have the same lead-singer
  - Evaluation metric : ranking
  - First method:
    - Extract some features from the spectral representation of the songs
    - Compare them
    - Check if the closest ones are from the same singer
  - Problem: even though the lead singer is prominent, the spectral properties of the observed signal are most of the time a non linear combination of the singer and the accompaniment.
  - Question: can we use some knowledge about ASA to minimize the impact of the accompaniment ?

CASA for singer similarity

- Assumptions:
  - The accompaniment does not change throughout the song
  - The singer starts singing at about 1 minute
- Proposed approach
  - Model the accompaniment as the audio signal of the beginning of the song
  - Model the singing voice as the audio signal around 1 minute
  - Compare songs represented as
    - spectral features
    - MFCC's
- Binary Masking:
  - Only consider spectral bins where amplitude of the mixture is larger than the accompaniment model.

CASA for singer similarity

- Dealing with missing data
  - Marginalization: only consider the non-zero spectral components during comparison
    - Loose a lot of data when many zeros are present
    - Feature representation is less powerful (can't use MFCCs)
  - Imputation: replace zero values by default ones
    - Can use any feature representation
    - What are the default values to consider ?