Source Separation Methods for under-determined sound mixtures

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

IRCAM
Centre Pompidou

ATIAM
Analysis of Sound Mixture

• We aim at performing
  o Auditory Scene Analysis
  o Computationally
  o But like human do
  o Humans focus on one source

• Task
  o Source separation ?
  o Source classification ?
  o Something in-between ?
    o What then ?
Human separate, really?

- It seems so:

(Fig. from Mesgarani Nature’12)
Computational ASA (CASA)

• How do people analyze sound mixtures?
  o break mixture into small elements (in time-freq)
  o elements are grouped into sources using cues
  o sources have aggregate attributes
I. Frequency Analysis (FA)

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

(Fig. from Aphex Twin)
I. 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.

\[ \hat{S}_n(r, k) = M_{mn}(r, k) \odot X_m(r, k) \]

- 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

- 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: professionaly 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.

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”)

---

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 $F \times N$, NMF is the problem of finding a factorization

$$V \approx WH$$

- where $W$ and $H$ are nonnegative matrices of dimensions $F \times K$ and $K \times N$, respectively.

- Use for transcription:

- Use for separation:
NMF

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

\[ v_n \approx \sum W_{i} h_{i} \]

- \( v_n \): data vector
- \( W \): "explanatory variables" or "basis", "dictionary" or "patterns"
- \( h_{i} \): "regressors" or "expansion coefficients" or "activation coefficients"
By representing signals as a sum purely additive, non-negative sources, we get a parts-based representation [Lee’99].
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
    \[ H_{a\mu} \leftarrow H_{a\mu} \frac{(W^T V)_{a\mu}}{(W^T WH)_{a\mu}} \quad W_{ia} \leftarrow W_{ia} \frac{(V H^T)_{ia}}{(WHHT)_{ia}} \]
  - Theorem: under the update rules, the cost function is
    - Non increasing
    - Invariant if @ stationary point

ICA on spectrograms
NMF on spectrograms
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
  
  o 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

  o 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

\[ p(V_i) = \sum_i \omega_{v,i} N(V_i; \bar{v}_i, \Sigma_{v,i}) \]

Local Power
Spectral Density
(PSD)

\[ \Sigma_{v,i} = \begin{pmatrix} \sigma_{v,i}^2(1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{v,i}^2(F) \end{pmatrix} \]

Voice GMM

\[ \lambda_v = \{ \omega_{v,i}^2, \Sigma_{v,i} \}_{i} \]

(Fig. from [Ozerov 05])
GMM – Based Source Separation

\[ x = v + m \]

(Fig. from [Ozerov 05])
GMM – Based Source Separation

Mathieu Lagrange. Blind Source Separation. 25

(Fig. from [Ozerov 05])
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:
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
  o ON rationale: remove what we can infer from the scene, and model the remaining
  o Semi Supervised NMF:
    o $X = FG + HU$
      • $F$: prior knowledge
      • $H$: model new events

• Reference:
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:
  o The accompaniment does not change throughout the song
  o The singer starts singing at about 1 minute

• Proposed approach
  o Model the accompaniment as the audio signal of the beginning of the song
  o Model the singing voice as the audio signal around 1 minute
  o Compare songs represented as
    o spectral features
    o MFCC’s

• Binary Masking:
  o 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?