Demixing Commercial Music Productions via Human-Assisted Time-Frequency Masking

MarC Vinyes, Jordi Bonada, Alex Loscos

Universitat Pompeu Fabra - Music Technology Group

May 21, 2006 (corrected)

Overview

Overview

1. Introduction
   - Problem
   - Signal Processing Solution
2. Algorithm
   - Overview
   - Generation of candidate solutions
   - Solution selection
   - Sound examples
3. In terms of previous research
4. Conclusions

Overview

1. Introduction
   - Problem
   - Signal Processing Solution
2. Algorithm
   - Overview
   - Generation of candidate solutions
   - Solution selection
   - Sound examples
3. In terms of previous research
4. Conclusions

Overview

Index

Context

- End-user goal: Demix commercial music productions
- Formulation of similar problems

  Audio Blind Source Separation: Sources?

  Computational Auditory Scene Analysis: Auditory streams?

Our formulation of the problem

Audio Blind Separation:

- Original mixed audio $(out_L, out_R) \longrightarrow$ Audio signals (s_i^L, s_i^R)
- Restrictions on (s_i^L, s_i^R):
  1. $(\sum_i^n s_i^L, \sum_i^n s_i^R)$ perceived similarly to (out_L, out_R)
  2. $s_i^L, s_i^R \; i = 1..n$ should mean something to a human (examples: tracks, instruments, auditory streams, physical sources, notes, chords, noises...)

Our formulation of the problem

Audio Blind Separation:

- Original mixed audio $(out_L, out_R) \longrightarrow$ Audio signals (s_i^L, s_i^R)
- Restrictions on (s_i^L, s_i^R):
  1. $(\sum_i^n s_i^L, \sum_i^n s_i^R)$ perceived similarly to (out_L, out_R)
  2. $s_i^L, s_i^R \; i = 1..n$ should mean something to a human (examples: tracks, instruments, auditory streams, physical sources, notes, chords, noises...)

What do we need?

1. **Generator of candidate solutions**
   A method to synthesize s_i^L,s_i^R from out_L,out_R.
2. **Solution selection criteria**
   A way to set the parameters of this method in order to obtain the desired meaninfgul solutions.

What do we need?

1. **Generator of candidate solutions**
   A method to synthesize s_i^L,s_i^R from out_L,out_R.
2. **Solution selection criteria**
   A way to set the parameters of this method in order to obtain the desired meaninfgul solutions.

Index

Main points

1. Method = TFM (Time Frequency Masking)
2. Meaningful solutions = Audio tracks used to produce the mix

Main points

1. Method = TFM (Time Frequency Masking)
2. Meaningful solutions = Audio tracks used to produce the mix

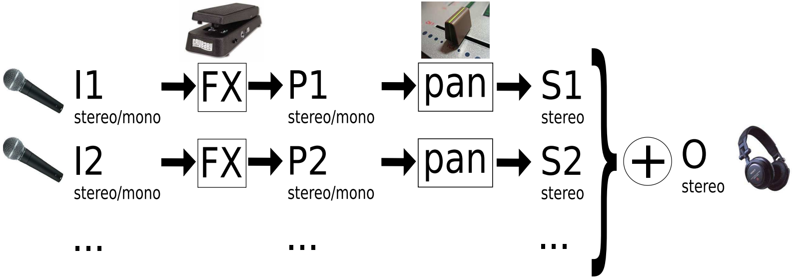
TFM explained

1. Signal splitted into overlapped frames of fixed size in time.
2. FFT
3. Binary mask applied

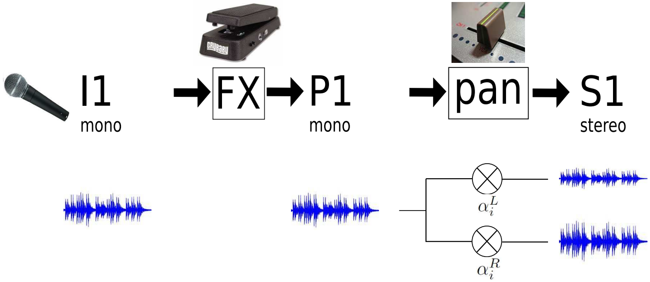
0.3+0.1j	0.2-j	0.2+0.2j	...	
↓	↓	↓		
1	0	1	...	Binary Mask
↓	↓	↓		
0.3+0.1j	0	0.2+0.2j	...	

4. IFFT
5. Overlap-and-add process.

Assumptions on how commercial music is produced



What is a “mono” track?



Let $p_i \in [0, 1]$ the value of the panning knob of track i:

$$\begin{cases} \alpha_i^L = \cos\left(p_i \cdot \pi/2\right) \\ \alpha_i^R = \sin\left(p_i \cdot \pi/2\right) \\ p_i = \arctan\left(\alpha_i^R/\alpha_i^L\right) \cdot 2/\pi \end{cases}$$

Mathematical characterizations of a “mono” track

- **Pan**:
  $p_i = \arctan \left| \frac{DFT_p(s_i^R)[f]}{DFT_p(s_i^L)[f]} \right| \cdot 2/\pi \ \forall f \in [0 \ldots N/2]$
- **IPD (Interchannel Phase-Difference)**:
  $|Arg(DFT_p(s_i^L)[f]) - Arg(DFT_p(s_i^R)[f])| = 0 \ \forall f \in [0 \ldots N/2]$

Approximate track time-frequency orthogonality

- Orthogonality favoured with a high resolution DFT: Blackmann-Harris -92dB, N=8192

$$DFT_p(out^R)[f] \simeq DFT_p(s_k^R)[f] + \sum_{i \neq k} \cdot DFT_p(s_i^R)[f]$$

$$\text{where } \left| \sum_{i \neq k} \cdot DFT_p(s_i^R)[f] \right| \ll 1$$

- Consequences:

Pan $\arctan \left| \frac{DFT_p(out^R)[f]}{DFT_p(out^L)[f]} \right| \cdot 2/\pi = p_k \pm \Delta$

IPD $|Arg(DFT_p(out^L)[f]) - Arg(DFT_p(out^R)[f])| = 0 \pm \Delta$

Approximate track time-frequency orthogonality

- Orthogonality favoured with a high resolution DFT: Blackmann-Harris -92dB, N=8192

$$DFT_p(out^R)[f] \simeq DFT_p(s_k^R)[f] + \sum_{i \neq k} \cdot DFT_p(s_i^R)[f]$$

$$\text{where} \quad \left| \sum_{i \neq k} \cdot DFT_p(s_i^R)[f] \right| \ll 1$$

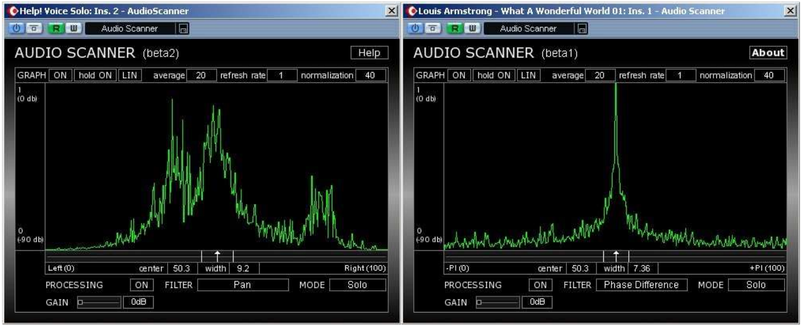
- Consequences:

Pan $\arctan \left| \frac{DFT_p(out^R)[f]}{DFT_p(out^L)[f]} \right| \cdot 2/\pi = p_k \pm \Delta$

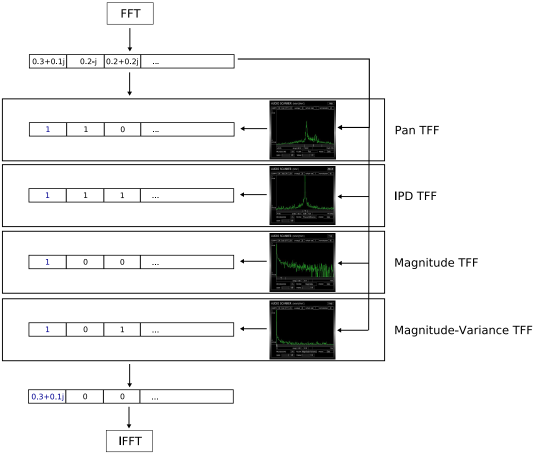
IPD $|Arg(DFT_p(out^L)[f]) - Arg(DFT_p(out^R)[f])| = 0 \pm \Delta$

Manual range selection

- We accumulate the energy contributions of the DFT coefficients for each possible estimated Pan and IPD.



TF Mask set in several steps: Time-Frequency Filters



Time-Frequency Filters Summarized

Selected range	Segregated signals
Pan	“mono” tracks with different pan
IPD	“mono” and “stereo” tracks
Magnitude	signals with narrow or wide spectrum
Magnitude-variance	attacks and steady sounds

Sound examples

- Beatles - Help (voice isolation/removal)
- Pearl Jam - Better Man (remixing)
- Explosions in the Sky - Memorial (snare extraction)

Index

Main references

YilRic Özgür Yilmazz and Scott Rickard. *Blind separation of speech mixtures via time-frequency masking.* IEEE Transactions on Signal Processing, 2003.

Ave Carlos Avendano. *Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications.* Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2004.

BarLaw Dan Barry, Bob Lawlor, and Eugene Coyle. *Sound source separation: Azimuth discrimination and resynthesis.* Proc. of the 7th Int.Conference on Digital Audio Effects (DAFX 04), 2004.

Previous approaches

- Time-Frequency Masking [YilRic], [Ave]

Cue	*Chosen* Δ	*Authors*
2D (IID,IDD)	Maximum likelihood	[YilRic]
1D (mapped IID)	Gaussian window	[Ave]

- Time-Frequency resynthesis (ADRess) [BarLaw]

Cue	*Chosen* Δ	*Authors*
Cancellation rule	Manual	[BarLaw]

Contributions

- IID mapping improvement. Original mixer knob resolution reproduced.
- Improved graphical interface to manually select Δ
- New criteria to set the TF Mask:
  IPD, Magnitude, Magnitude-variance.
- Introduction of the concept of cascaded TFFs. Flexibility to combine those criteria.

Index

Conclusions

- We are close to achieve good quality demixing of commercial music productions.
  - TFM is a good choice.
  - We only need to find better TFFs.
- Presented algorithm
  - The present characterizations are still not robust although the human-assisted approach helps.
  - Still an experimental tool for voice removal and remixing.
  - Excelent tool to analyse music recordings.