Singing voice separation

Gerard Erruz & Manaswi Mishra

Motivation

- Isolate vocal contribution from other sounds in the mixed audio file
- Content-based and blindly performed. No extra information is used apart from original audio file
- Applications:
  - Karaoke creation
  - Singer identification
  - Lyrics recognition and alignment
  - Synthetic music composition
  - Remixing

MIREX tasks description

- Input: audio file containing the mix with a vocal sound in it.
- Output: new audio file containing extracted vocal track (minimized background noise/accompaniment)
- Specifications:
  - 16-bit
  - Monaural
  - 44.1kHz sample rate
  - 30 seconds long

Datasets

Name	Characteristics	Limitations	Availability	URL
Demixing Secrets Dataset (DSD100)	100 tracks - professionally mixed - different styles	-----	Free Download	http://bit.ly/2kdeeMM
IKALA Dataset	252 excerpts 30 sec long	- Only Chinese music (6 voices) - not pro mix	On Request	http://bit.ly/2kyHzlX
Medley DB	70 tracks - multiple genres - pro mix	- limited tracks with vocals	On Request	http://bit.ly/2kiqcHz
MIR1K	- 1000 songs - Chinese pop music	- voice and accompaniment on separate channels	Free Download	http://bit.ly/2kdlqZm

Literature review (pre-2014)

- Supervised systems
  - maps signal into feature space
  - detect singing voice segments
  - apply separation techniques

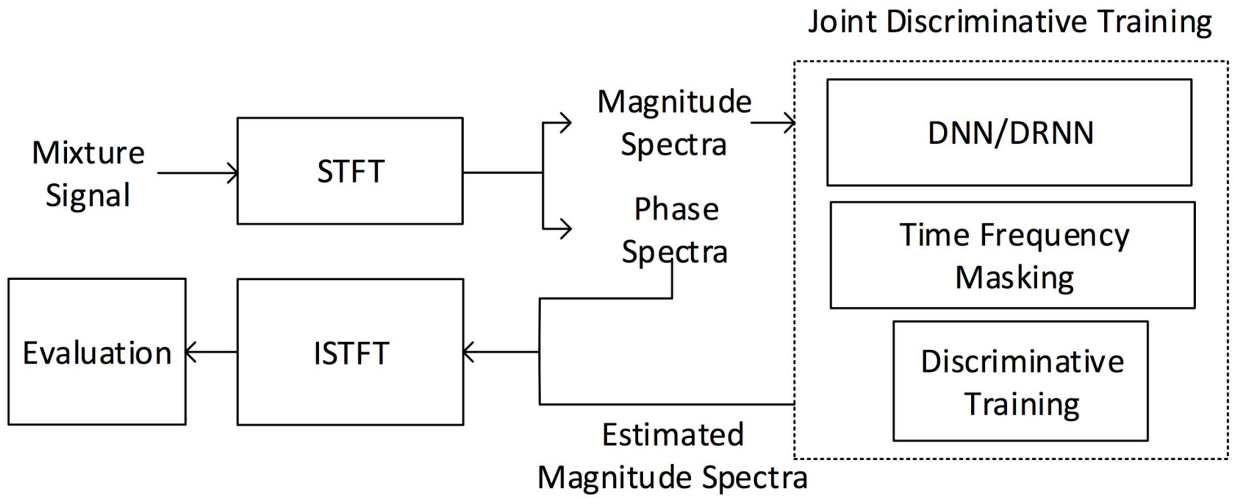
  - NMF [Vembu 2005] [D.D. Lee 2001]
  - Adaptive Bayesian Modeling [A. Ozerov 2007]
  - Pitch Based Interference [Yipeng Li 2007]

- Unsupervised systems
  - Source Filter model [J. L. Durrieu 2010]
  - Autocorrelation based method [Z Rafii 2011]

Literature review (MIREX)

- Singing-Voice Separation using Deep Recurrent Neural Networks (Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, Paris Smaragdis)

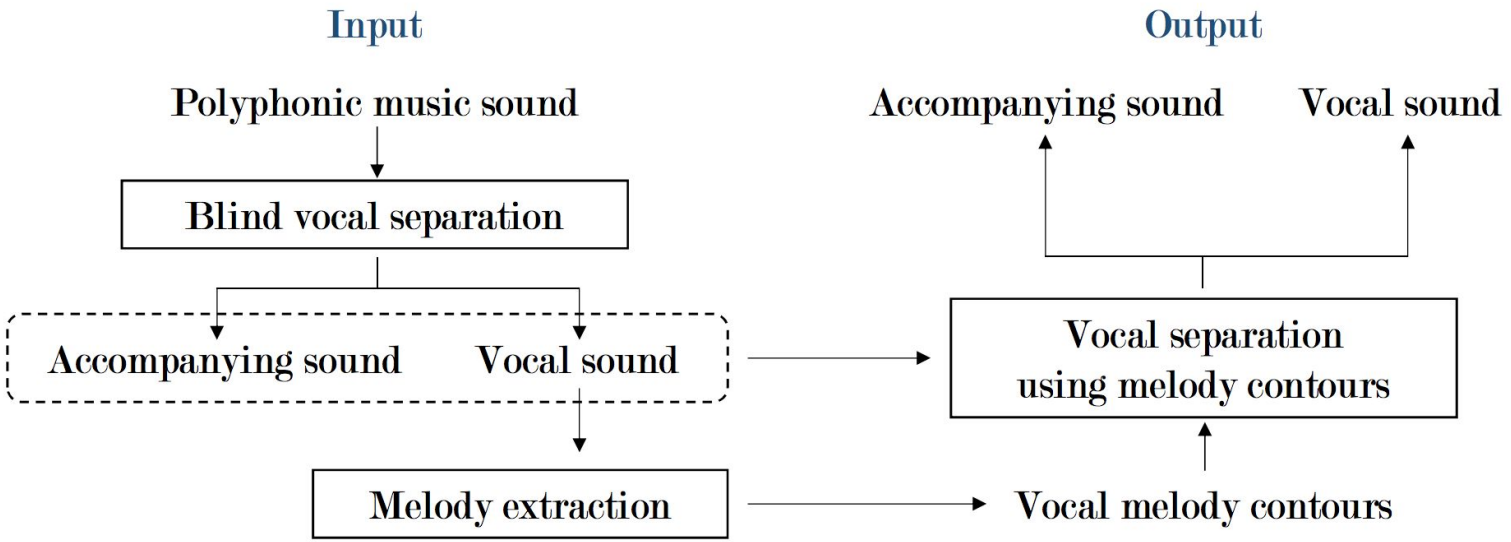
  http://www.music-ir.org/mirex/abstracts/2014/HKHS2.pdf



Literature review (MIREX)

- MIREX2015: Singing Voice Separation (Yukara Ikemiya, Katsutoshi Itoyama, Kazuyoshi Yoshii)

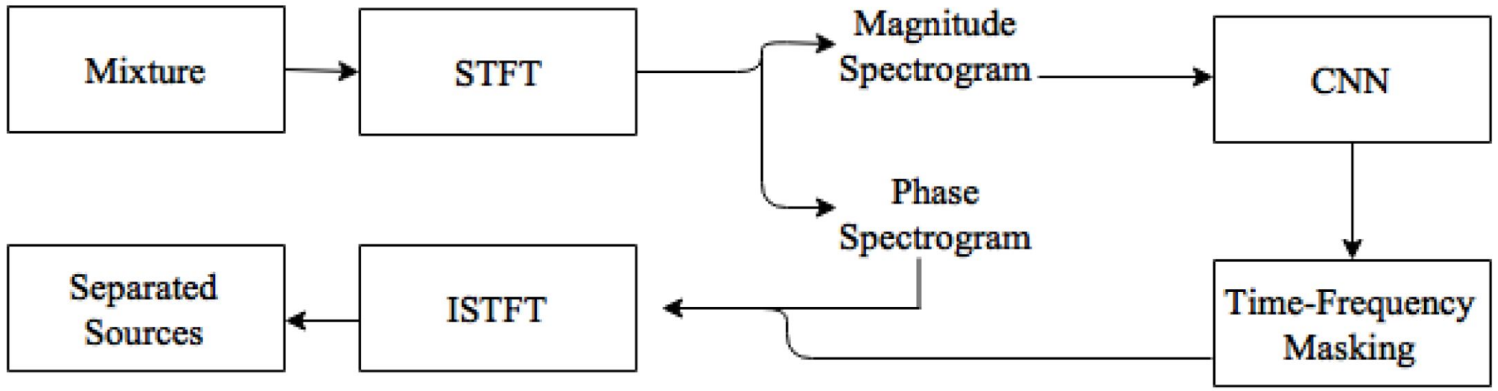
  http://www.music-ir.org/mirex/abstracts/2015/IIY3.pdf



Literature review (MIREX)

- MIREX 2016 Submission for Singing Voice Separation (Marius Miron, Pritish Chandna)

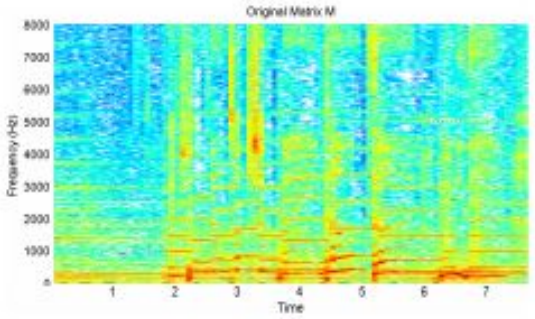
  http://mtg.upf.edu/node/3680



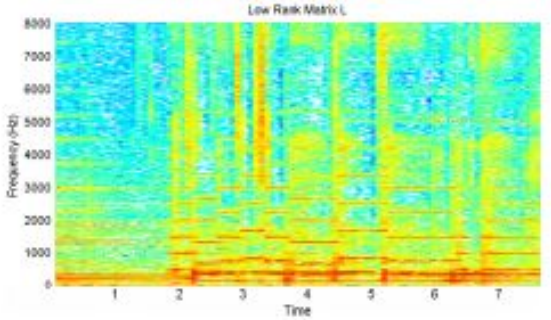
Implementations I

- Flexible Audio Source Separation Toolbox (FASST)
  - NMF, GMM, Source-Filter models for separation
  - https://github.com/wslihgt/pyfasst

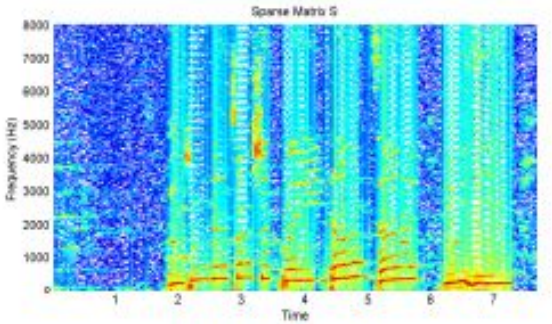
- Po-Sen Huang (2012)
  - Robust Principal Component Analysis
  - https://github.com/posenhuang/singingvoiceseparationrpca



(a) Original Matrix M



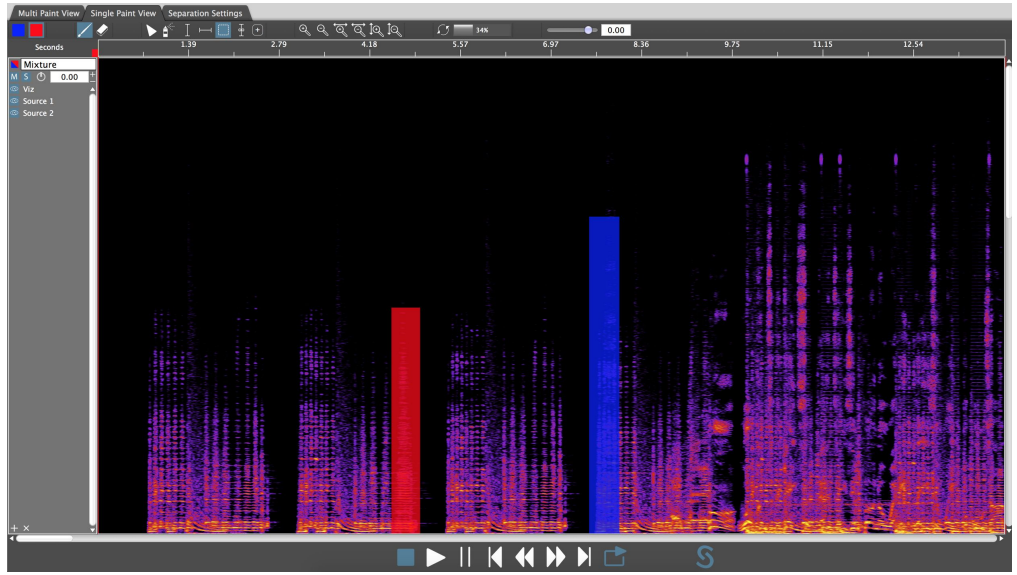
(b) Low-Rank Matrix L



(c) Sparse Matrix M

Implementations II

- ISSE
  - Semi-supervised interactive source separation
  - http://isse.sourceforge.net



- DeepConvSep
  - Using Convolutional Neural Networks (CNN)
  - Aimed to any sound (not only singing voice)
  - https://github.com/MTG/DeepConvSep/

Evaluation

For evaluation, MIREX propose to compute different comparison values between voice contribution and background sounds:

- **GNSDR** = Global Normalized Signal-to-Distortion Ratio

$$GNSDR = \frac{\sum_{i=1}^{100} NSDR_i}{100}$$

- **SDR** = Signal-to-Distortion Ratio

$$\mathrm{SDR} := 10 \log_{10} \frac{\|s_{\mathrm{dist}}\|^2}{\|e_{\mathrm{interf}} + e_{\mathrm{noise}} + e_{\mathrm{artif}}\|^2}$$

- **SIR** = Signal-to-Interference Ratio

$$\mathrm{SIR} := 10 \log_{10} \frac{\|s_{\mathrm{dist}}\|^2}{\|e_{\mathrm{interf}}\|^2},$$

- **SAR** = Signal-to-Artifacts Ratio

$$\mathrm{SAR} := 10 \log_{10} \frac{\|s_{\mathrm{dist}} + e_{\mathrm{interf}} + e_{\mathrm{noise}}\|^2}{\|e_{\mathrm{artif}}\|^2}.$$

Approach for evaluation

- DSD100 Matlab: SISEC (and MIREX) proposed evaluation script for DSD100 dataset

https://github.com/faroit/dsd100mat

- Output variables:
  - SDR, SIR, SAR
  - Global and Local Evaluations
  - Several values per song (averages for overlapping 30 second frames)

- Need for subjective evaluation methods

Current work

- pyFASST
  - Tested on short song samples (30 secs)
  - 3 types of model parameters tested for tuning

- DeepConvSep
  - Whole DSD100 separated sources
  - Voice showing good empirical results

Data Augmentation

Robustness Tests

- Mixing with parameterized noise
  - iKALA and MiR1K (synthetic mixes)

Variability in Training

- Circular shifting voiced components and mixing
- Synthetic mixing at different ratios.

Further work

- Run evaluation script using UPF cluster: http://hpc.dtic.upf.edu
- Compare results from pyFASST and DeepConvSep
- Compute separation with data augmentation modifications (noise, different mixing ratios)
- Compare evaluation script results and empirical perception.

Thank you!