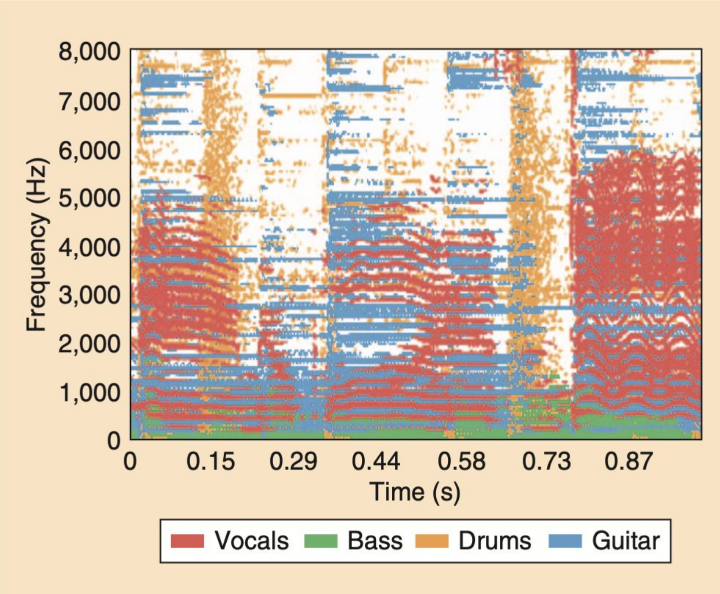
Musical Source Separation

Mahshid Alinoori

Table of Contents

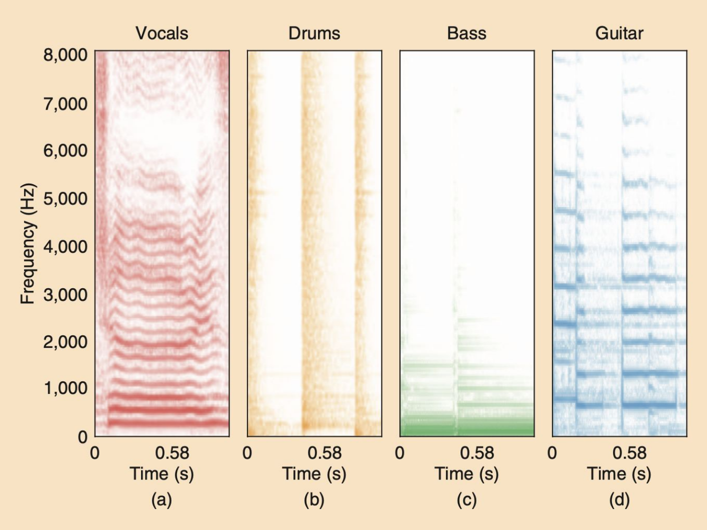
What is Musical Source Separation?

- To recover one or more of the source signals that are present in a mixture
- Applications are better remixing, upmixing, rebalancing, simpler transcription
- We can leverage the characteristics of musical sources to perform source separation



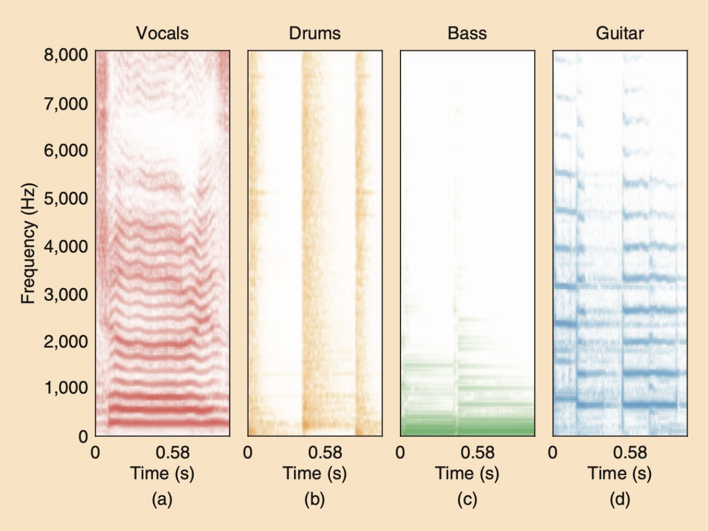
Musical Characteristics

- **Domain:** Voice Separation, harmonic-percussive separation, or instrument-based separation
- **Harmonic sources:** presented as horizontal components, characterized as f0 & harmonic series, contribute to forming timbre
- **Percussive sources:** have vertical structures, hold rhythmic information



Musical Characteristics

- **Number of channels:** possibility of spatial positioning in multichannel mixtures
- **Repeating structures:** used in kernel models
- **Manipulation of sources:** applied by hardware devices and digital audio workstations while recording and mixing



MSS Steps

TF Transformation

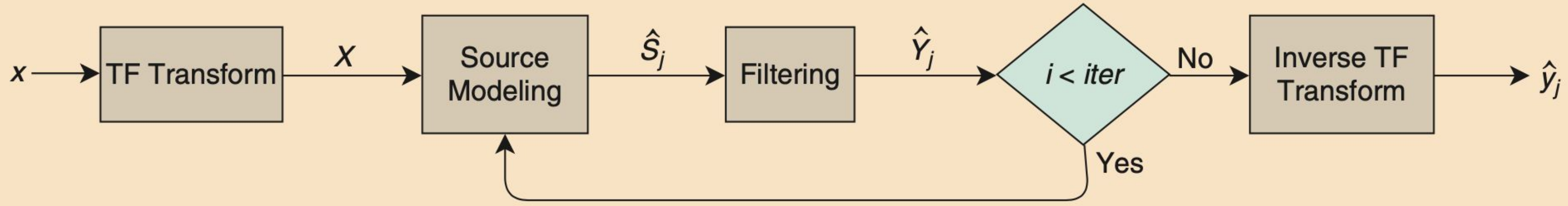
Mostly using STFT.
The mixture equals the sum of the sources in the transformed domain:
$X = \Sigma Y_j$

Source Modeling

Estimating the model of the spectrogram or the location of the target source

Filtering

Estimating the separated music source given the source modeling by applying masks.



More on Filtering...

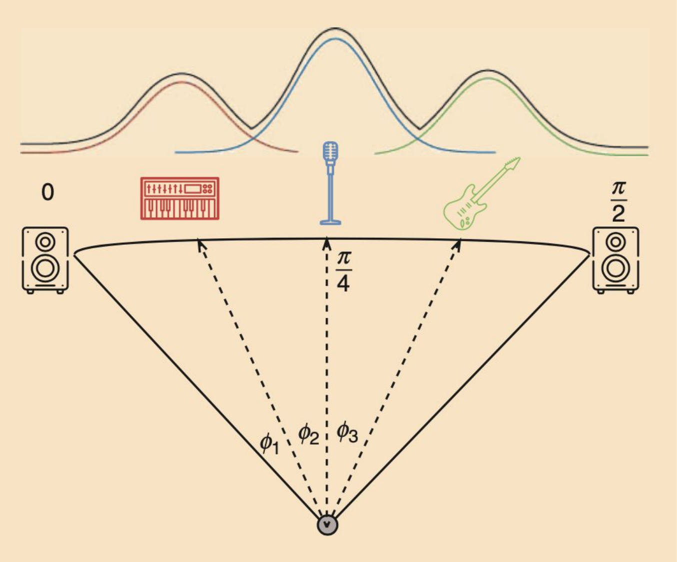
- **Mask:** expresses the extent to which each of TF bins belongs to the target source and it is applied to the original spectrogram

- **Binary vs Soft:** whether the mask contains 0/1 values or contains ratio of the magnitude of the source to the sum of all source magnitudes

- **Wiener Filtering:** a common softmask filtering with the output: $Y_1(k,n) = X(k,n)S_1(k,n)/\Sigma S_j(k,n)$

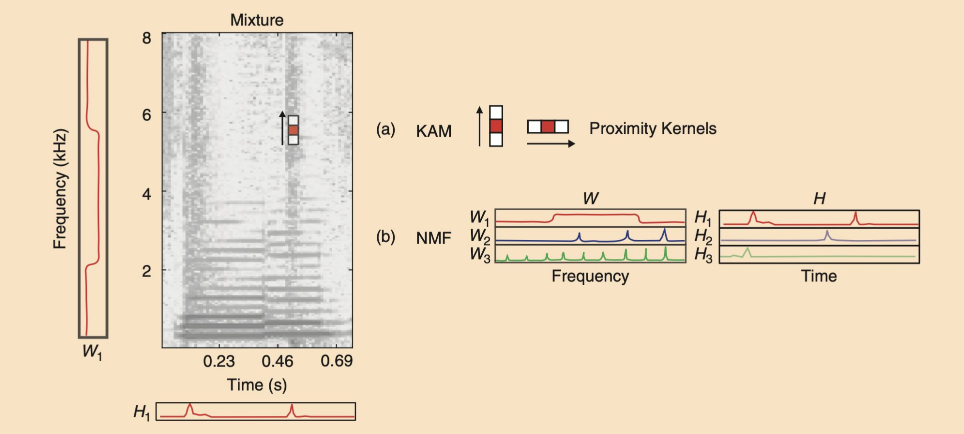
Musical Source Position Models

- Spatial position of sources is used
- The mixture is a stereo signal
- Constant power panning law: The overall volume is perceived regardless of the panning
- Assumption: very little overlap in the TF representation of the sources → only one source contribute to a single point in the TF representation defined by ϕ_i
- DUET, ADRess, PROJET: histogram of angle estimates and masking



Musical Source Models: Kernel Models

Kernel additive models exploit repetitions, continuity, and common fate.



Selecting the TF bins with similar values as the proximity kernel.

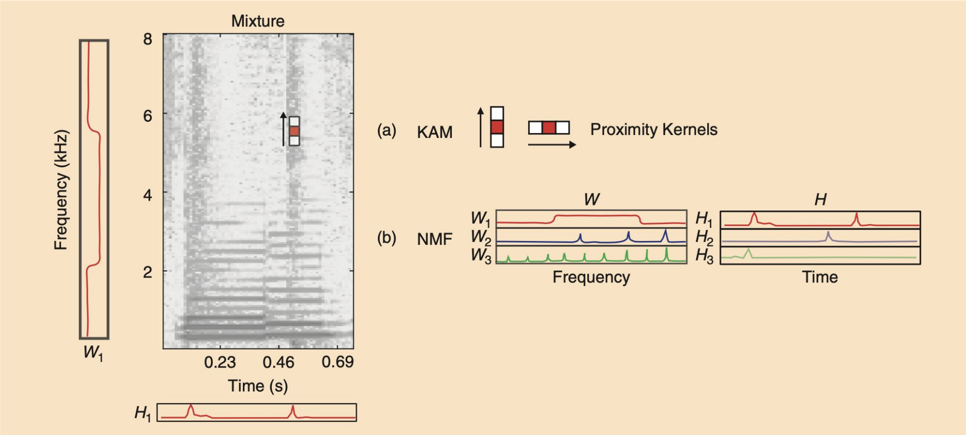
Assume the interference from other sources as outlier and removes them using median filtering.

Musical Source Models:
Spectrogram Factorization Models

Nonnegative matrix factorization: M ≈ WH applied to the magnitude spectrogram.

Solved as an optimization problem to minimize the difference between M and WH or by iterative learning.

Not very good at singing voice separation.



Musical Source Models:

Sinusoidal Models

Modeling the music signal with multiple sinusoids.

Mostly effective for harmonic sound separation and harmonic-percussive separation

DNN Models

No explicit source modeling is required.

Targets can be either separated sources or the masks.

We still need to fix the number and nature of the sources.

Looking for better cost function than MSE

MSS Evaluation

Subjective

Seems necessary but time-consuming and costly

Perceptual

Mapping results from listening tests to create metrics which has not been successful

Blind (BSS)

Objective and non perceptual based on energy ratios like SDR, SIR, SAR

Research directions

- Reducing artifacts
- Assumptions on the number of sources
- Separation of similar sources
- Unified and robust evaluation metrics
- Availability of larger datasets

THANKS!