ISMIR
OCTOBER 8-12 2012
PORTO
13th
International Society
for Music Information
Retrieval Conference

Leveraging Repetition to Parse the Auditory Scene

Josh McDermott, Bryan Pardo, and Zafar Rafii

interactive audio lab

NORTHWESTERN UNIVERSITY

Outline

I. **Introduction**
II. How humans use repetition to identify sound sources (McDermott)
III. Coffee break
IV. Repetition-based algorithms for source separation (Rafii)
V. Links to other methods for source separation
VI. Conclusions/Questions

Who are we?

Josh McDermott, Assistant Professor
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
jhm@mit.edu
web.mit.edu/jhm/www/

Zafar Rafii, doctoral candidate
Electrical Engineering & Computer Science
Northwestern University
zafarrafii@u.northwestern.edu
www.cs.northwestern.edu/~zra446/

Bryan Pardo, Associate Professor
Electrical Engineering & Computer Science
Music Theory & Cognition
Northwestern University
pardo@northwestern.edu
www.bryanpardo.com

What should you get out of this?

- An understanding of the psychological basis for the application of repetition to audio source separation and identification
- Understanding a new class of practical algorithms that perform repetition-based source separation
- Understanding the relationship of these algorithms to existing work in source separation

The Cocktail Party

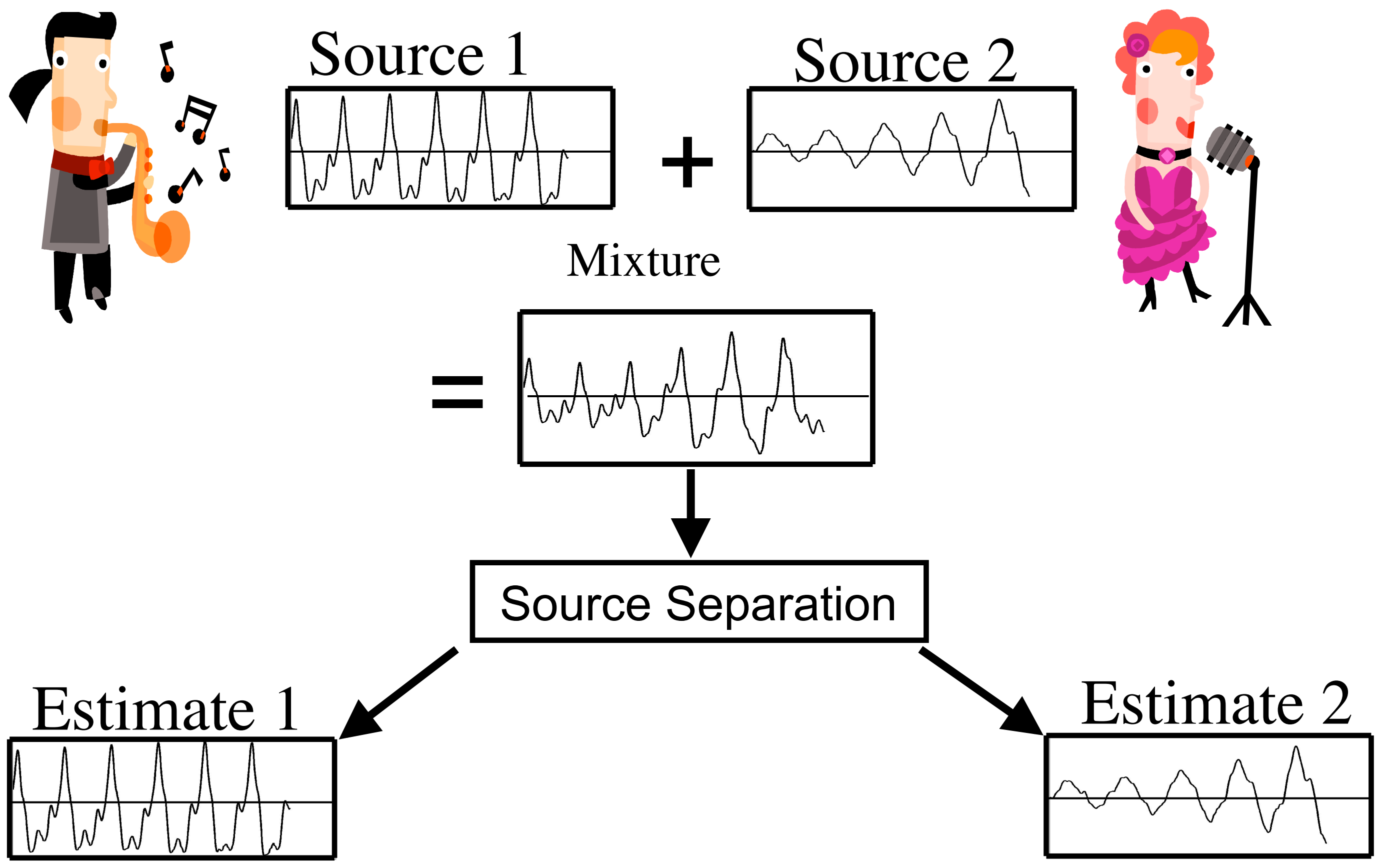
A party, usually in the early evening, at which cocktails are served.

The Cocktail Party Problem

How to listen to a single talker among a mixture of conversations and background noises.

Audio Source Separation

- Separating out the individual sounds in an audio mixture

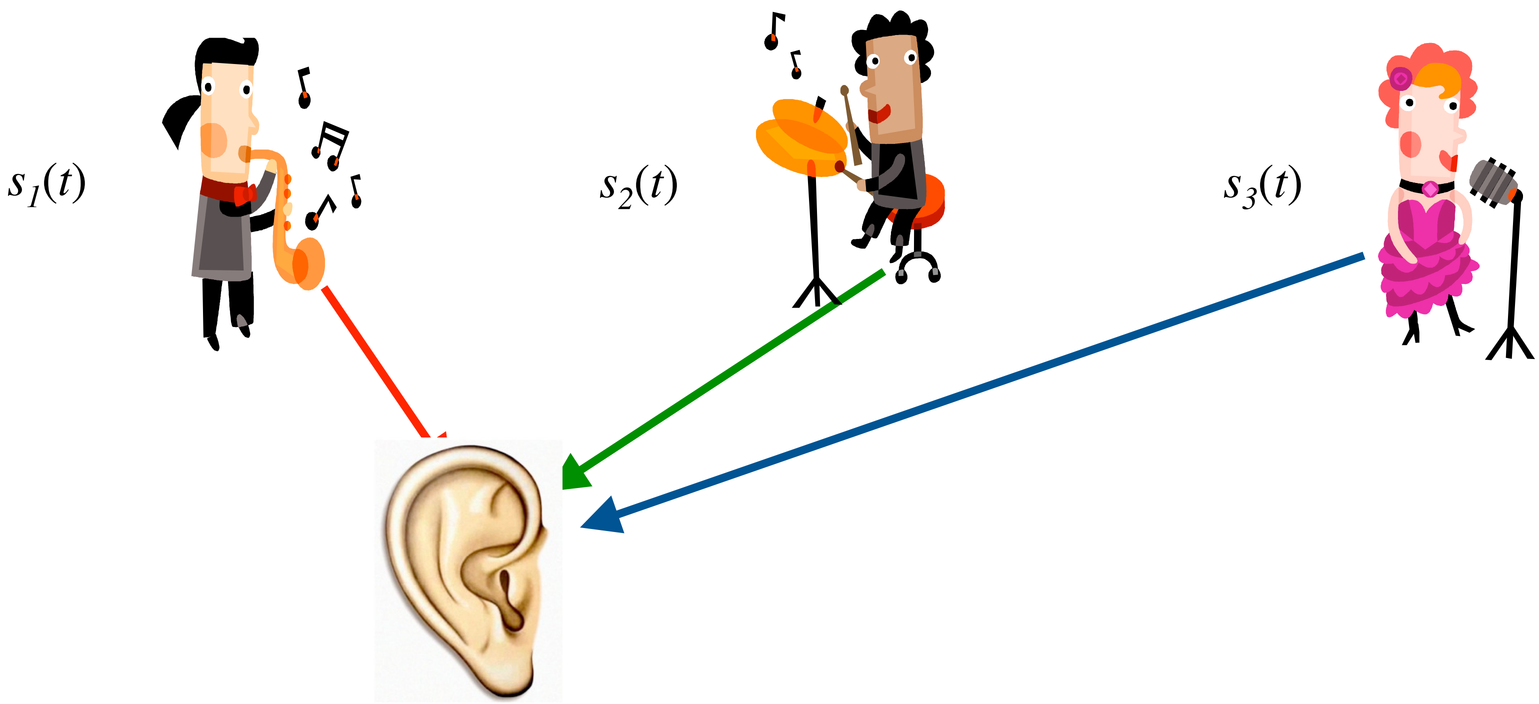


One mixture = underdetermined problem

Mix = Sound1 + Sound2 + Sound 3

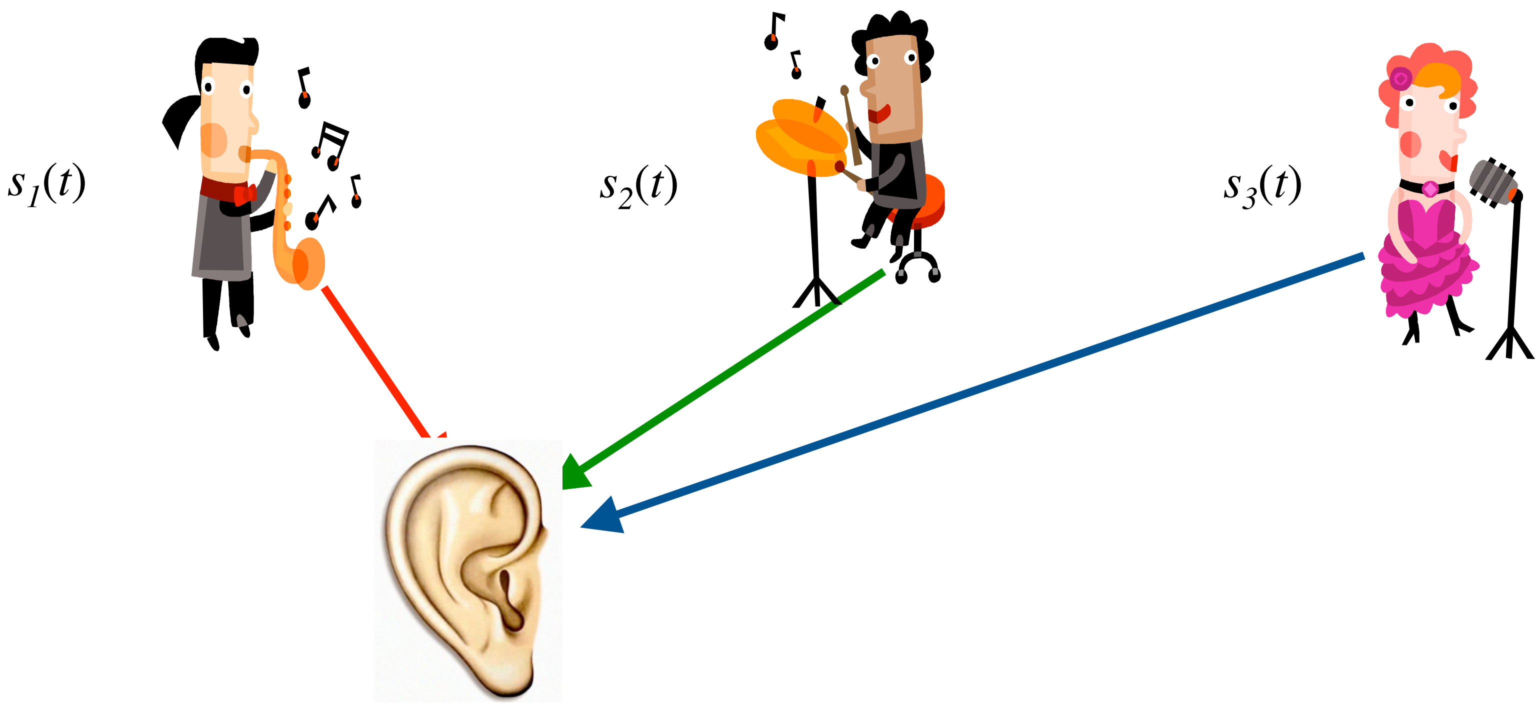
$$x_1 = \sum_{n=1}^{N} s_n(t)$$

Infinite number of solutions!



An underdetermined problem

- Sounds can be segregated only with the aid of prior assumptions about the world.
- We should infer sounds consistent with the acoustic input and our knowledge of real-world sounds.



Assertions

- Repetition is a fundamental element in generating and perceiving structure in music (…and audio in general)

- Repeating acoustic structure provides a cue that can be used to segment audio scenes

Questions

- What evidence is there that humans use repetition to parse an auditory scene?
- Can we build source separation algorithms based only on repetition cues?
- Can we leverage repetition to improve existing approaches to source separation?

Outline

I. Introduction
II. How humans use repetition to identify sound sources (McDermott)
III. Coffee break
IV. Repetition-based algorithms for source separation (Rafii)
V. Links to other methods for source separation
VI. Conclusions/Questions

Recovering Sound Sources From Repetition

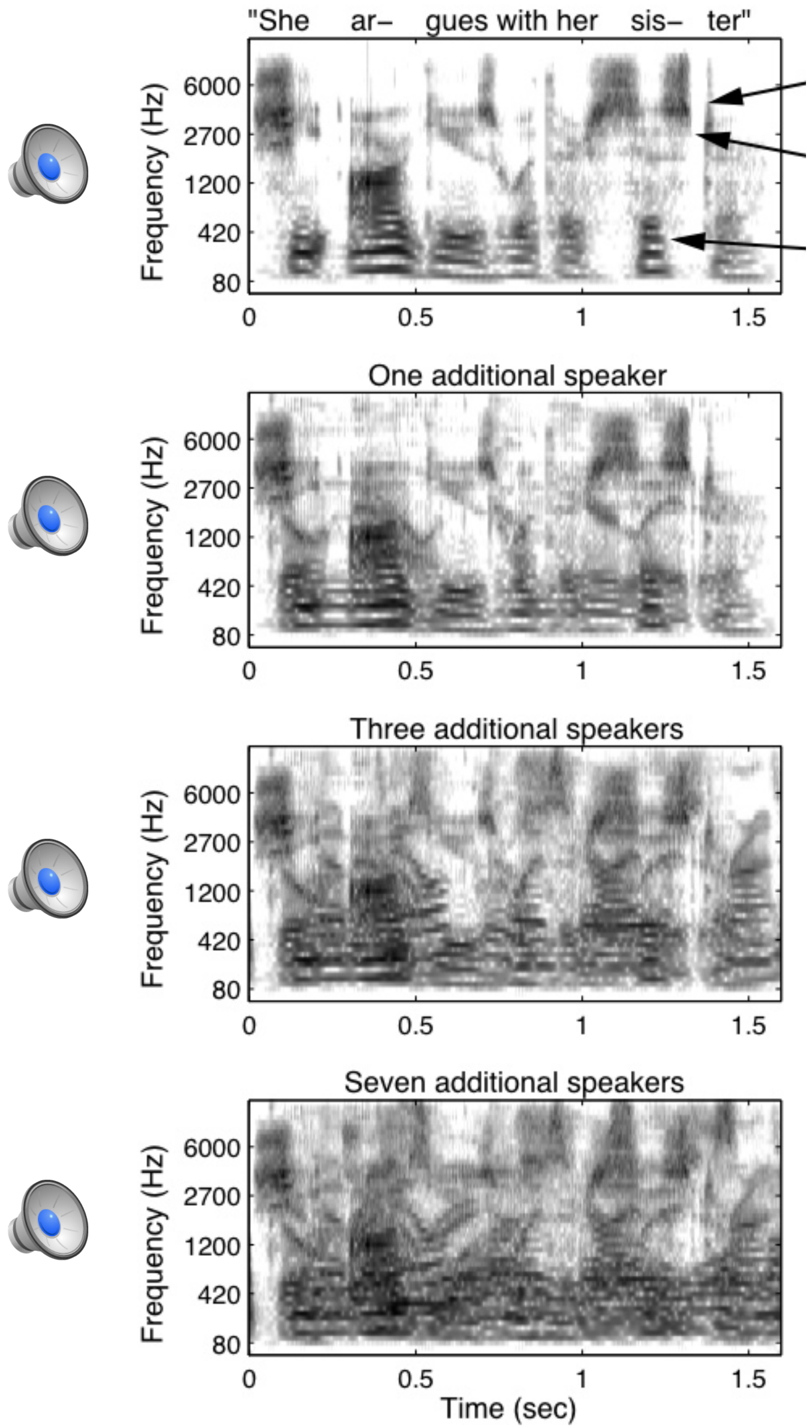
Josh McDermott

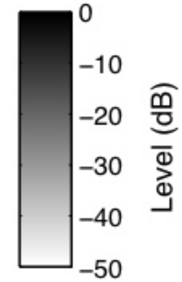
Dept. of Brain and Cognitive Sciences

MIT

THE COCKTAIL PARTY PROBLEM

Natural auditory environments have many sound sources:

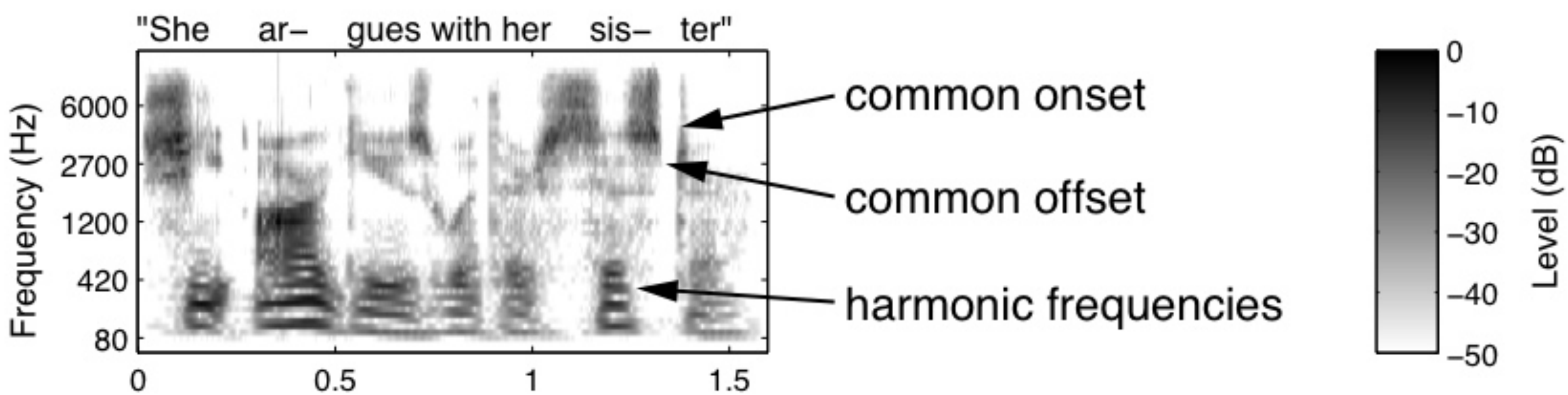




McDermott 2009, Current Biology

Sound Segregation

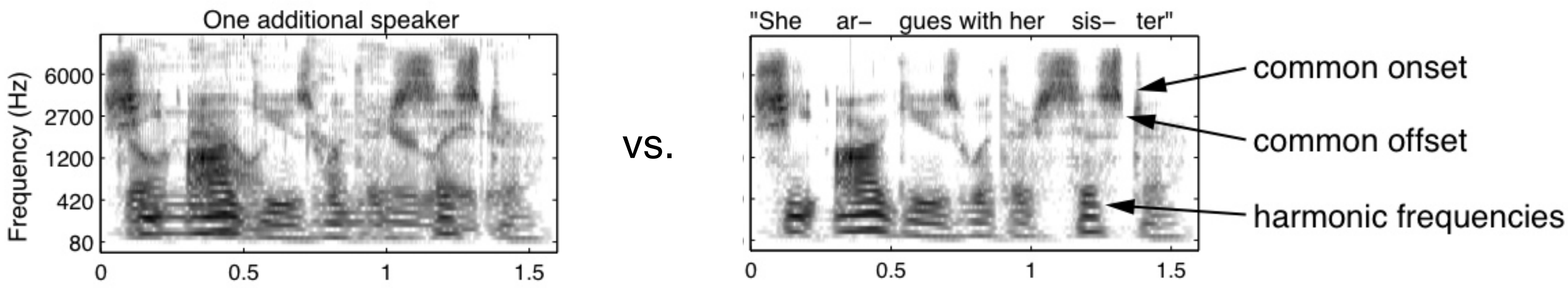
- Classic ill-posed problem in perception.
- To estimate sources, we need prior knowledge:



Humans use both generic (bottom-up) and specific (top-down) cues.

But… How do we acquire prior knowledge of sources?

If most of our auditory input is mixtures, how do we get started?



Need to know properties of individual sources to segregate them, but need to have segregated them to learn their properties…

Spatial cues are not of great help.

Idea: Perhaps if same source repeats, auditory system can detect repeating structure, infer presence of sound source.

Mixtures are accidental, don't occur repeatedly

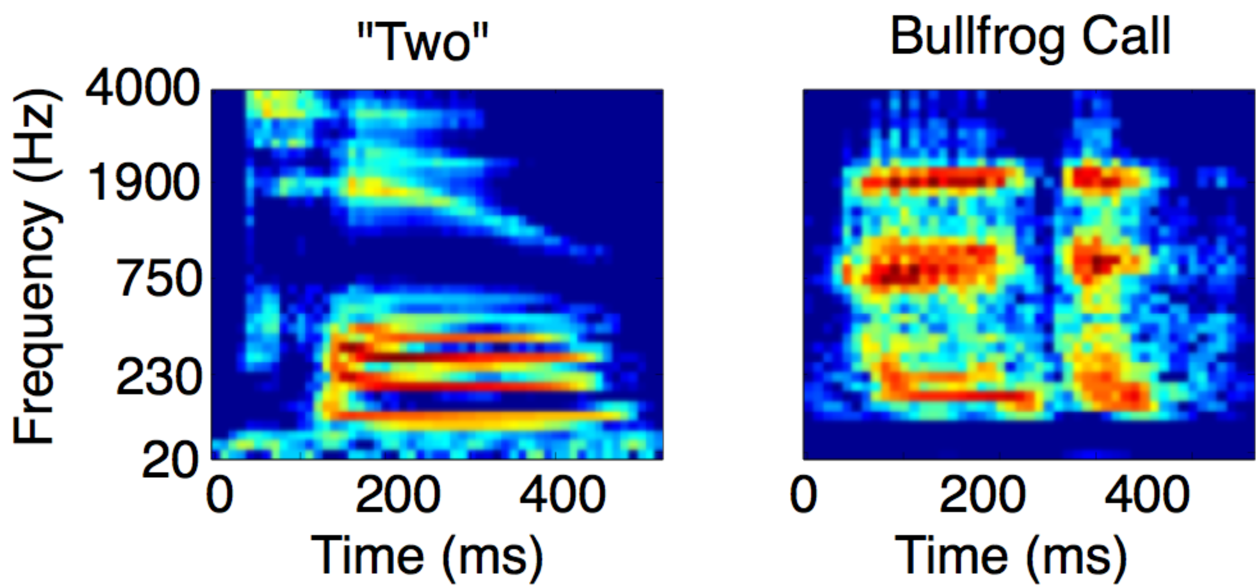
→ Repeating structure is likely to be a single source

To test, need a way to generate novel sound sources…

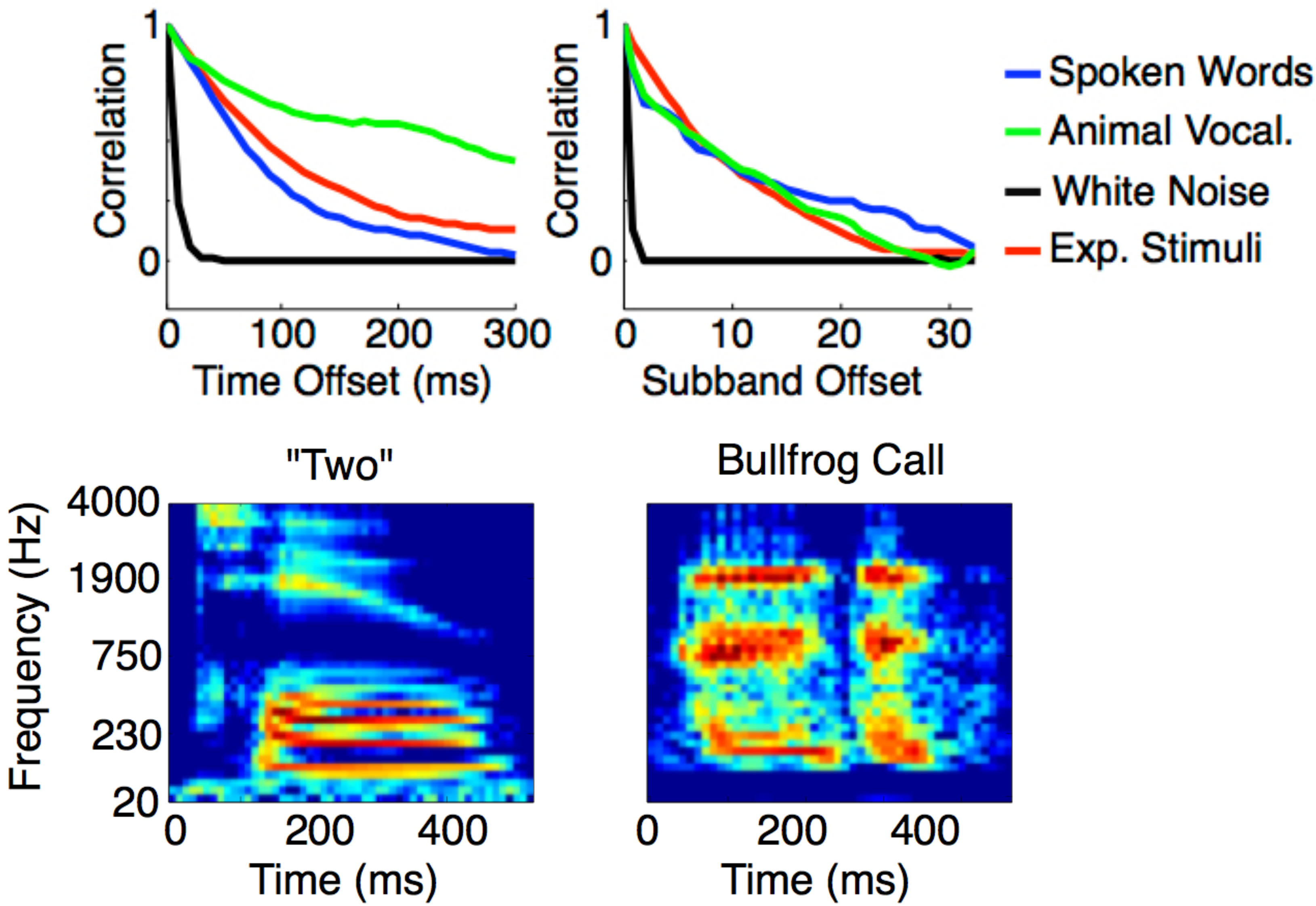
White noise is no good - all samples sound the same:

Want stimuli to have some properties of natural sounds, so that they don't all sound the same (cf. white noise).

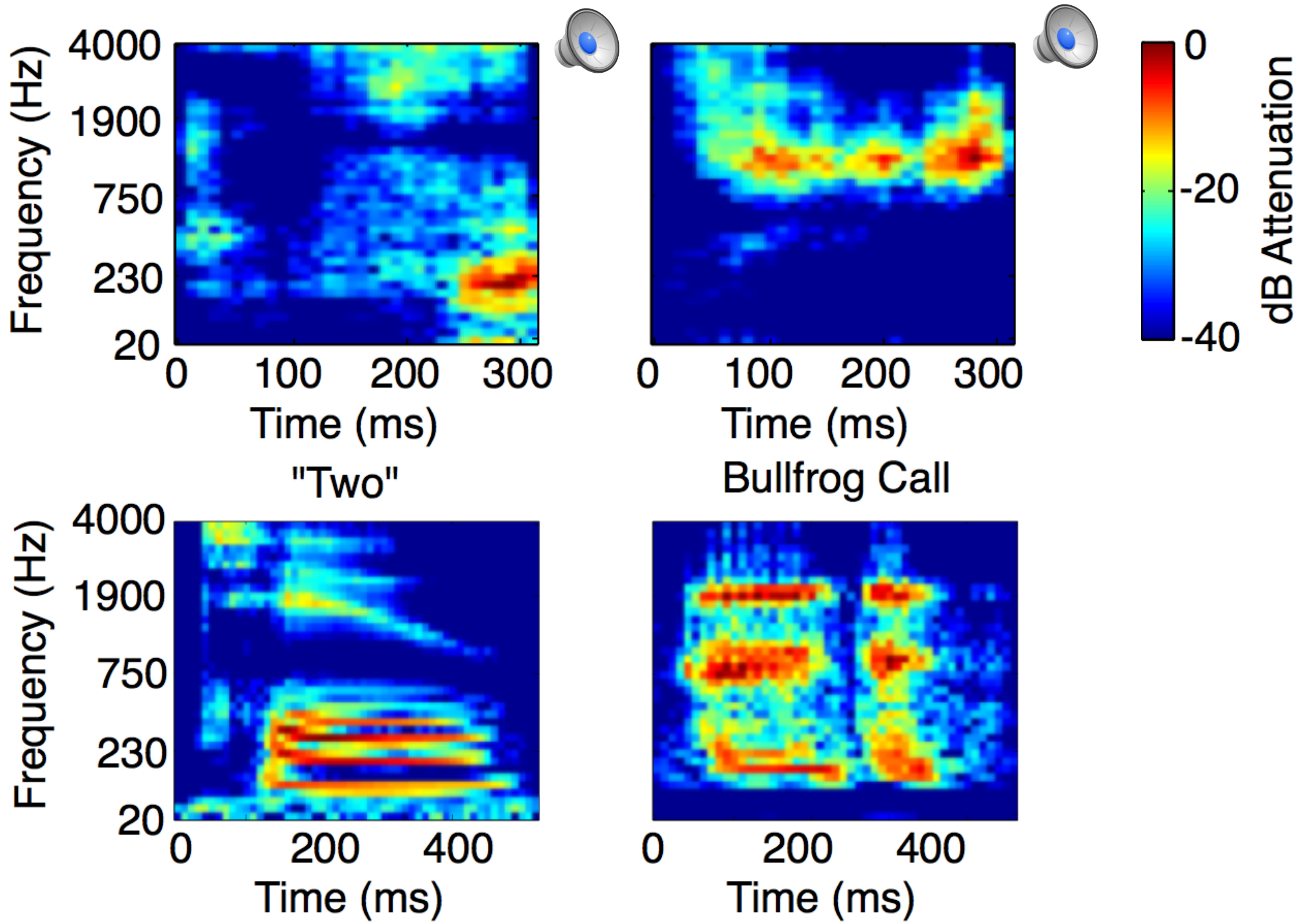
But want them NOT to have strong bottom-up grouping cues, so that we can examine how sounds might be recovered from mixtures BEFORE other grouping cues have been learned.



Time-frequency decompositions of real-world sounds exhibit correlations in both time and frequency:

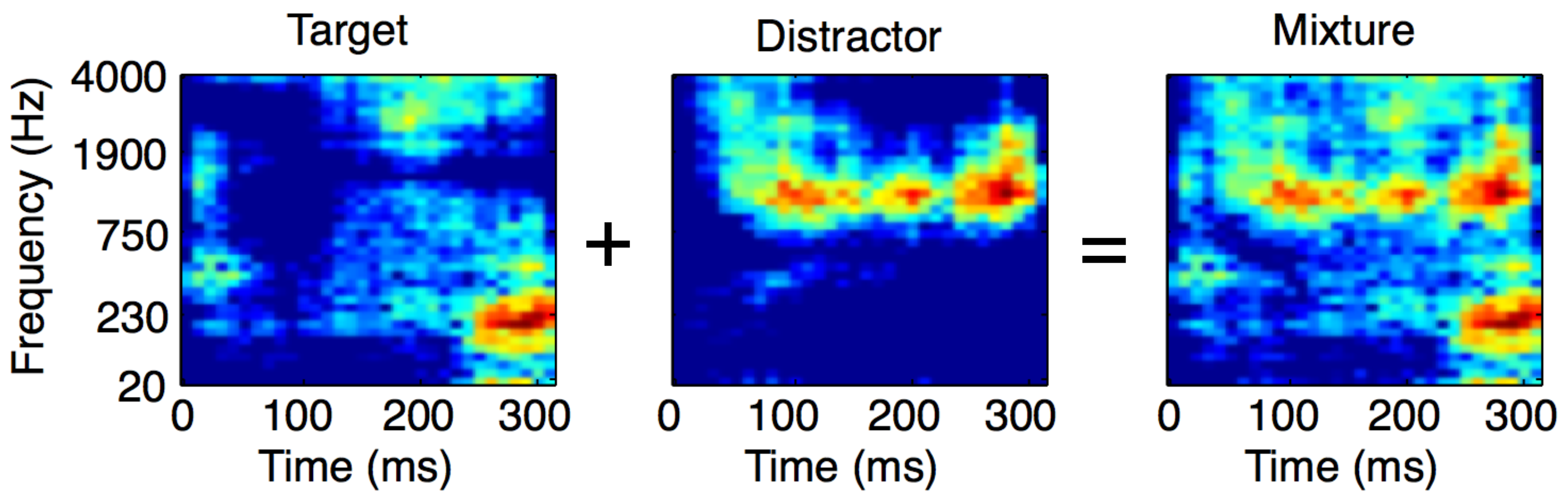


We captured these correlations by modeling log-energy spectrograms as a multivariate Gaussian random variable:

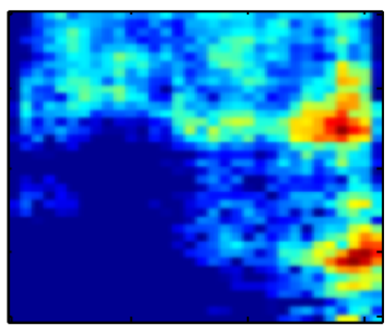


McDermott, Wrobleski & Oxenham, PNAS 2011

Synthetic sources can be combined into mixtures:



Present mixture, then probe sound:

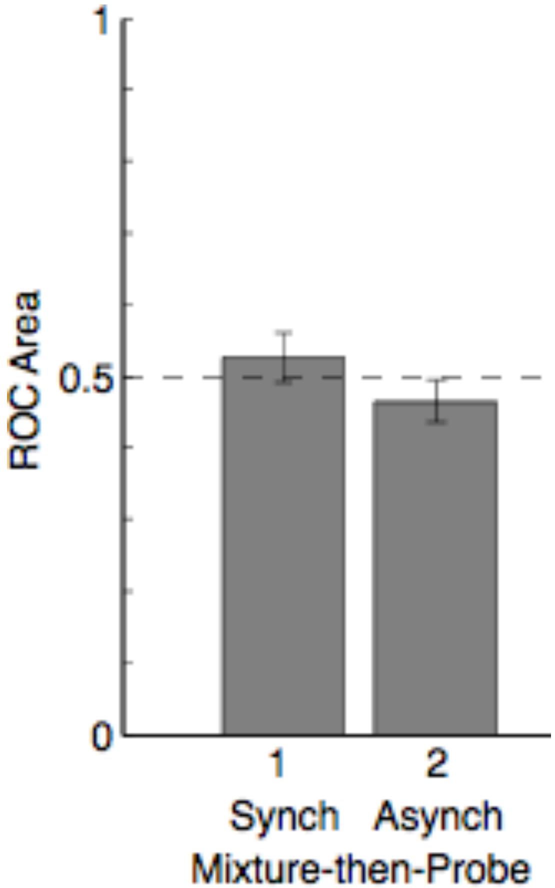


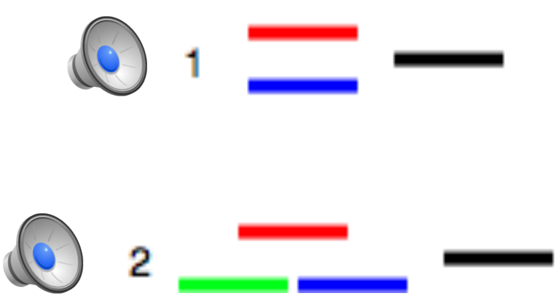
Was the probe one of the sounds in the mixture?

Sounds have structure, but not enough to allow segregation.

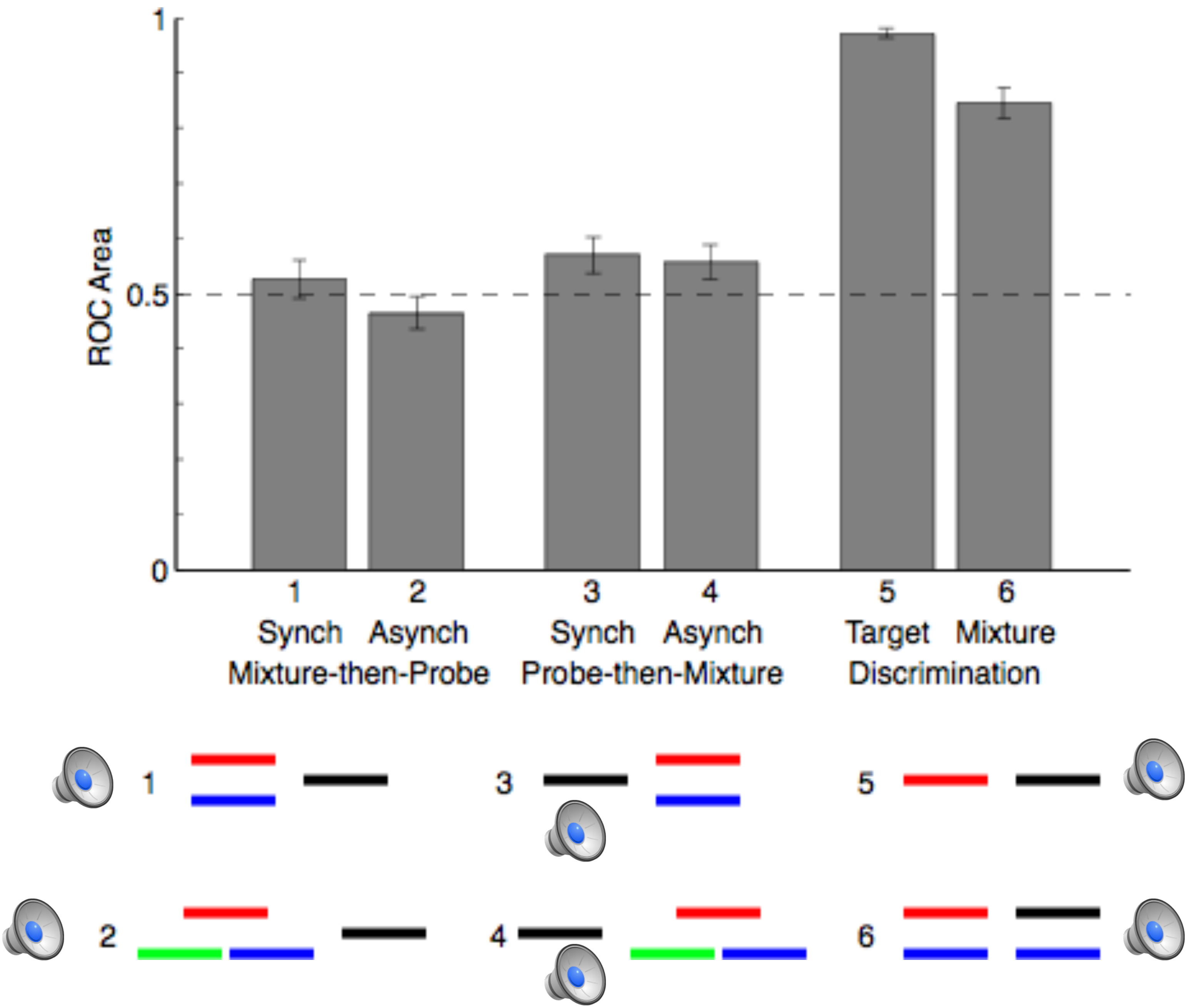
McDermott, Wrobleski & Oxenham, PNAS 2011

Single mixtures are hard to segment:





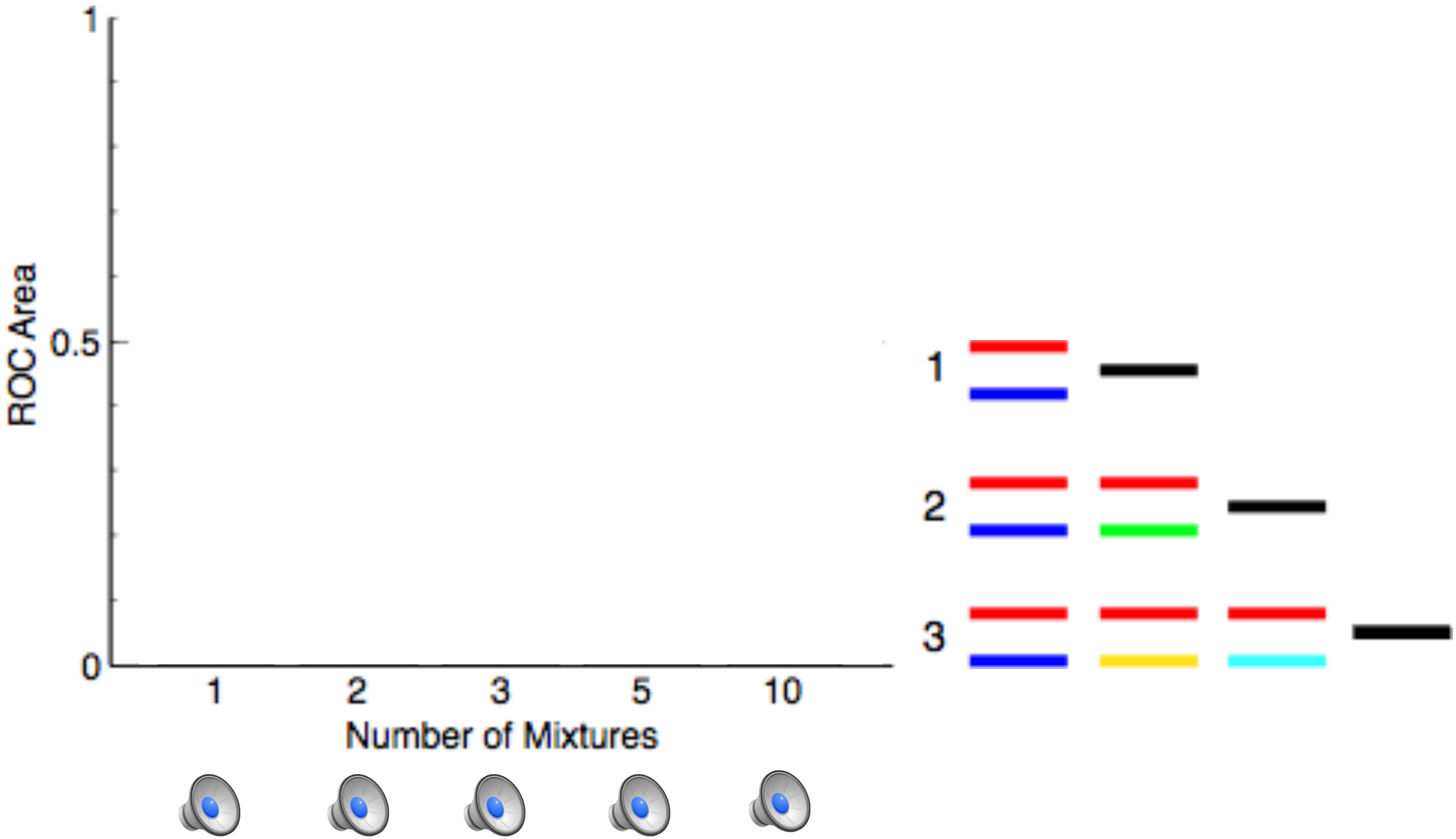
Performance not limited by discriminability:



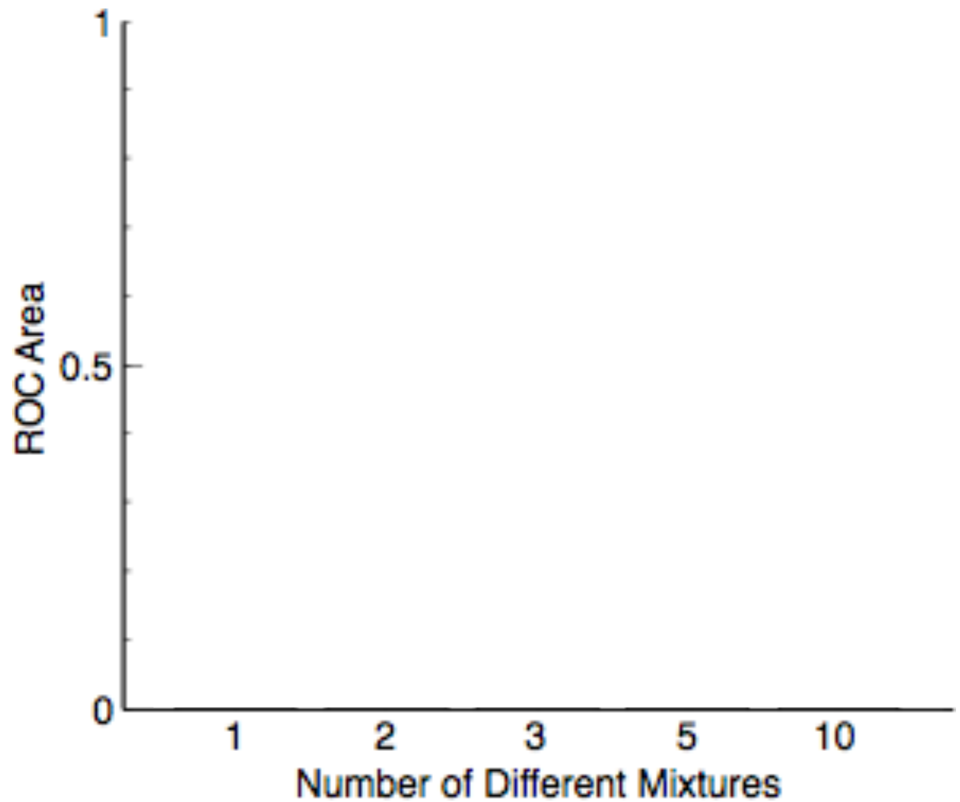
- Performance seems to be limited by ability to segregate sounds.
- Stimuli evidently contain few bottom-up segregation cues.

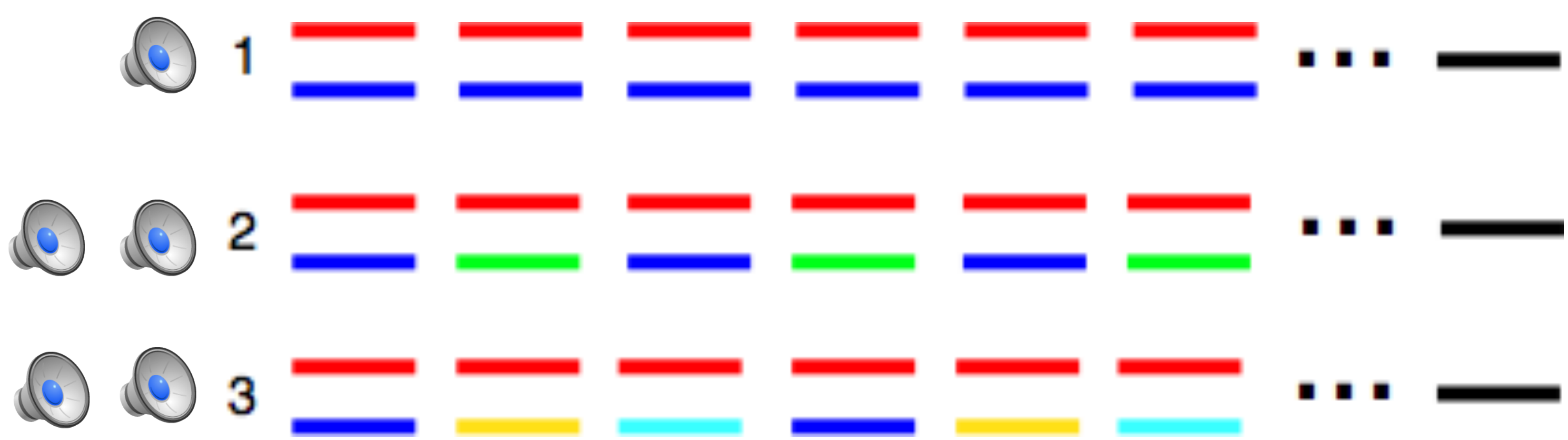
Can people recover these sources if they are repeated?

Effect of presenting target multiple times, each time with different distractor:

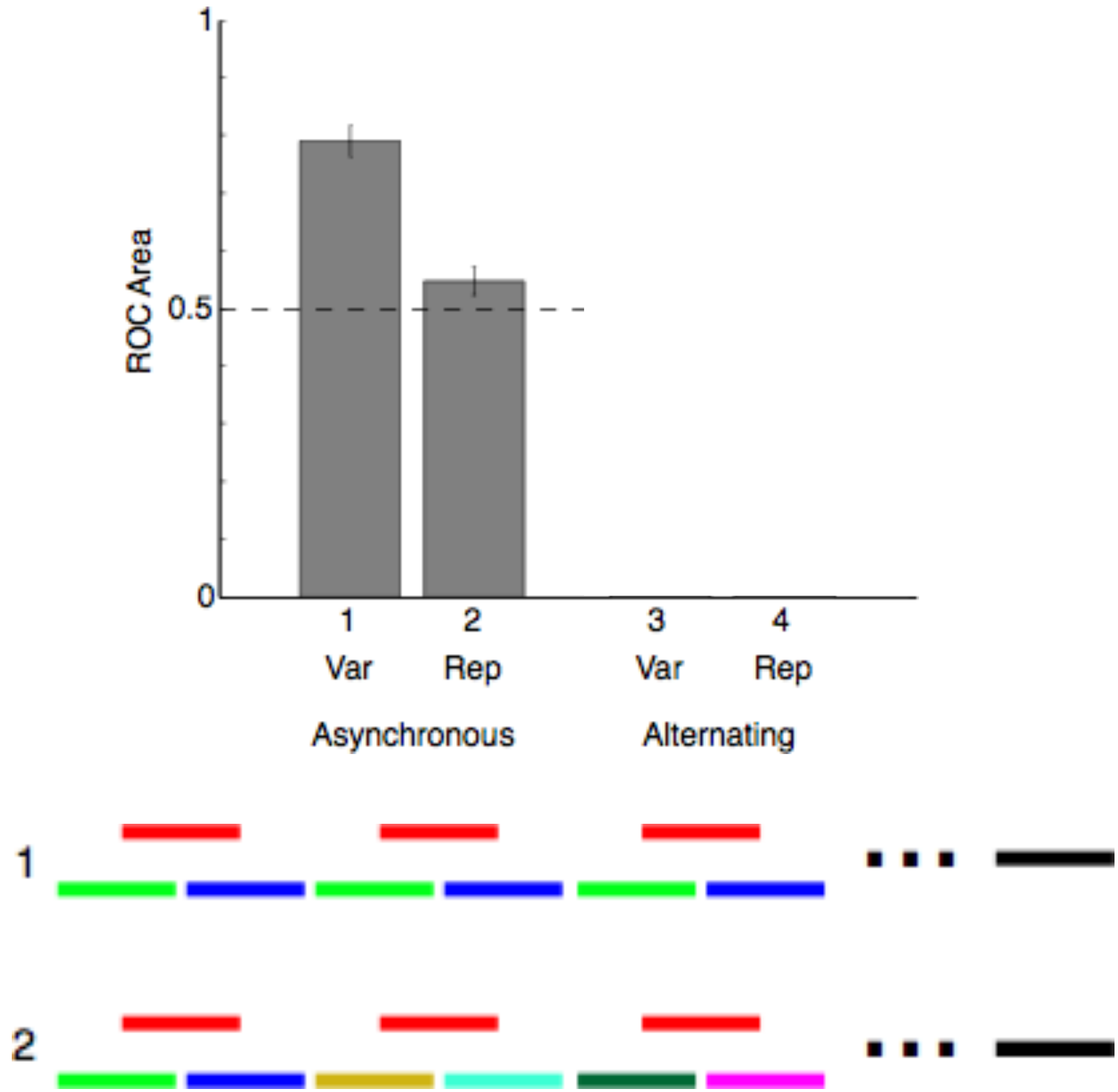


Performance depends on number of *different* mixtures:

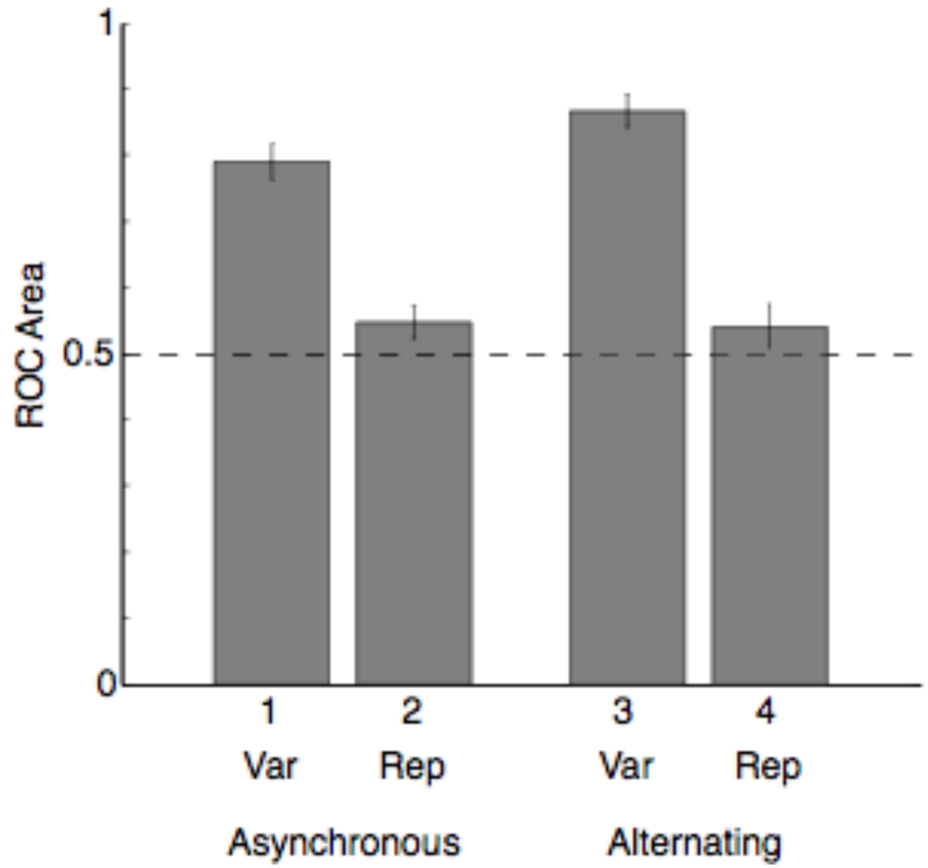




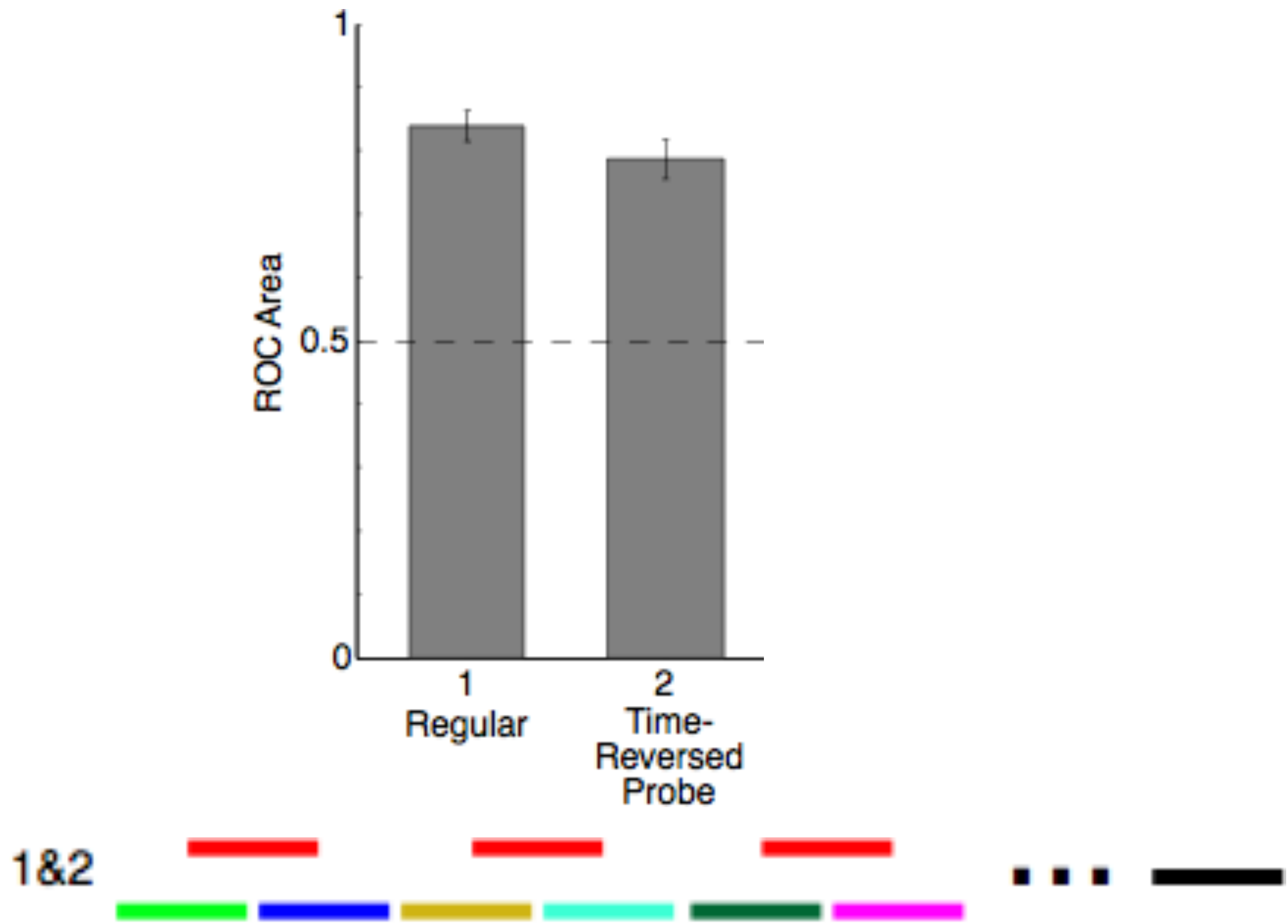
Effect of multiple mixtures swamps that of asynchrony:



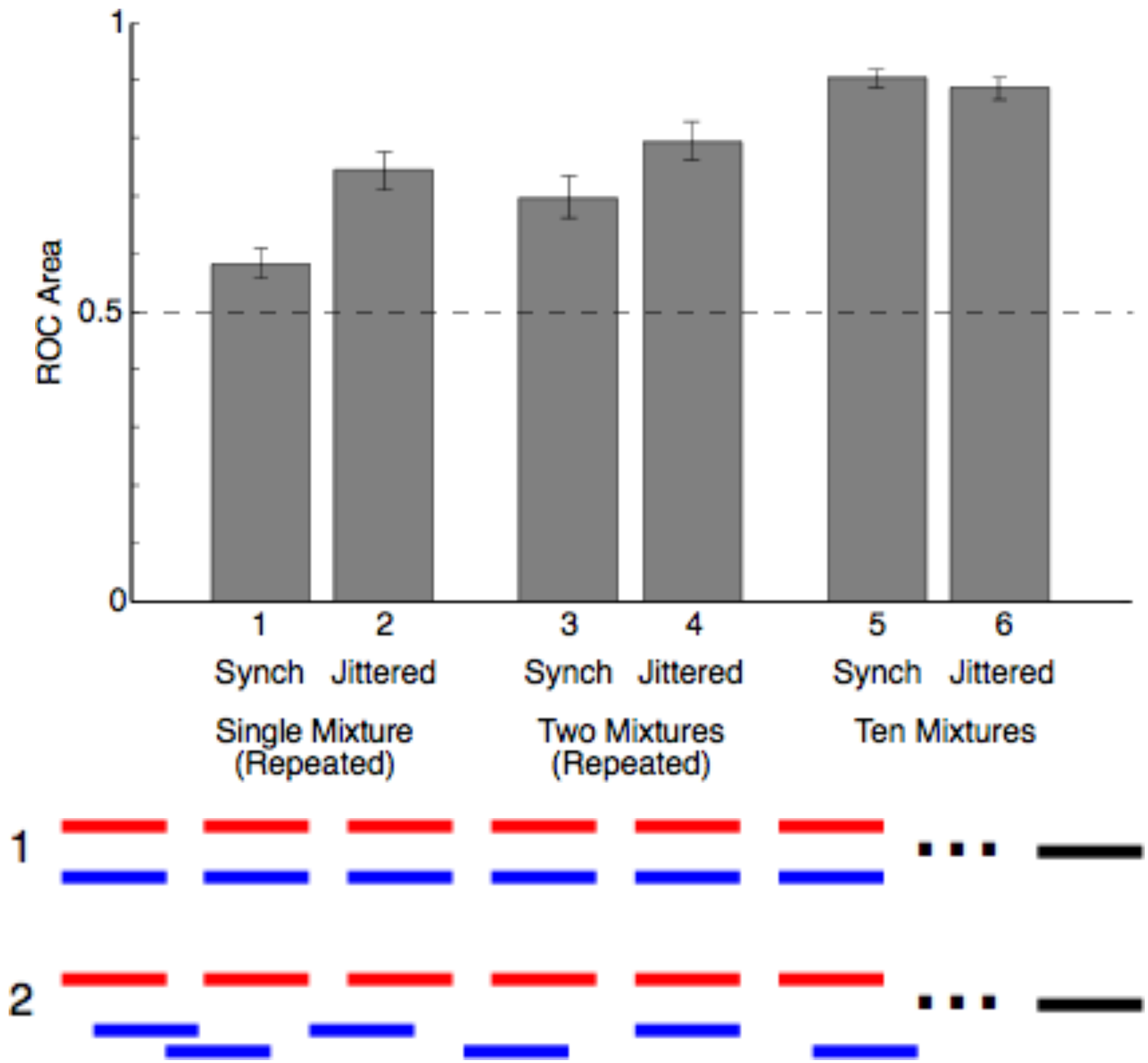
Only variability of distractors mixed with targets matters:



Listeners are not simply using average spectrum:



Jittering onset of distractors has similar effect to varying them:

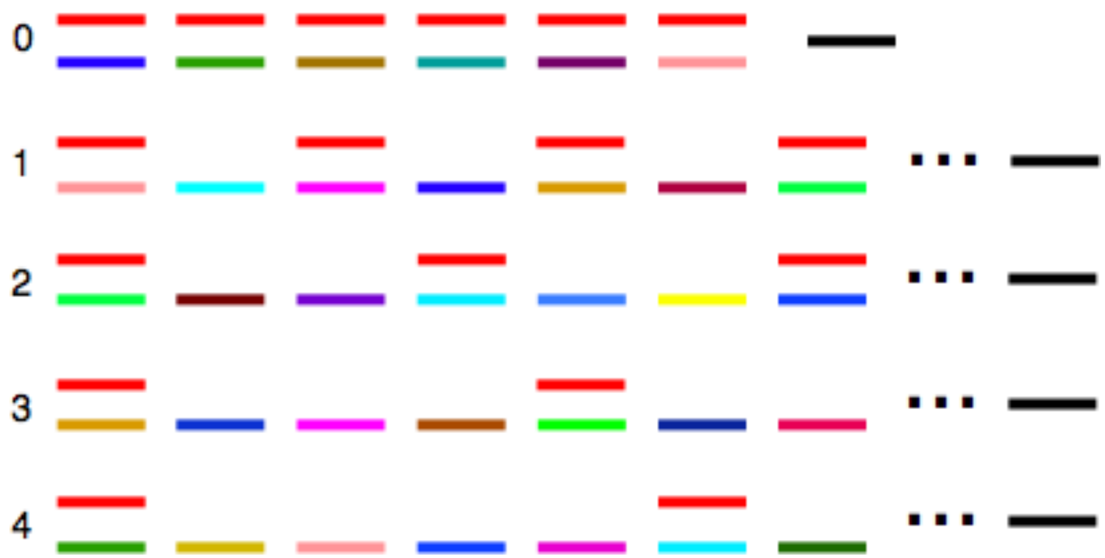


Auditory system seems to be tracking repeating structure.

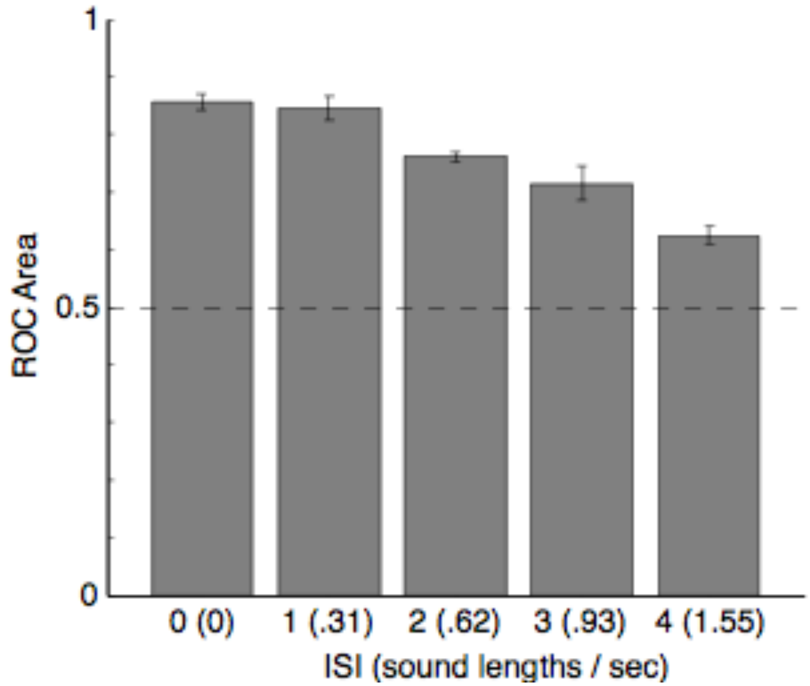
Listeners can recover source when it occurs in multiple distinct mixtures.

Performance should be constrained by storage capacity: recognizing repeating structure requires comparison of input at different time points.

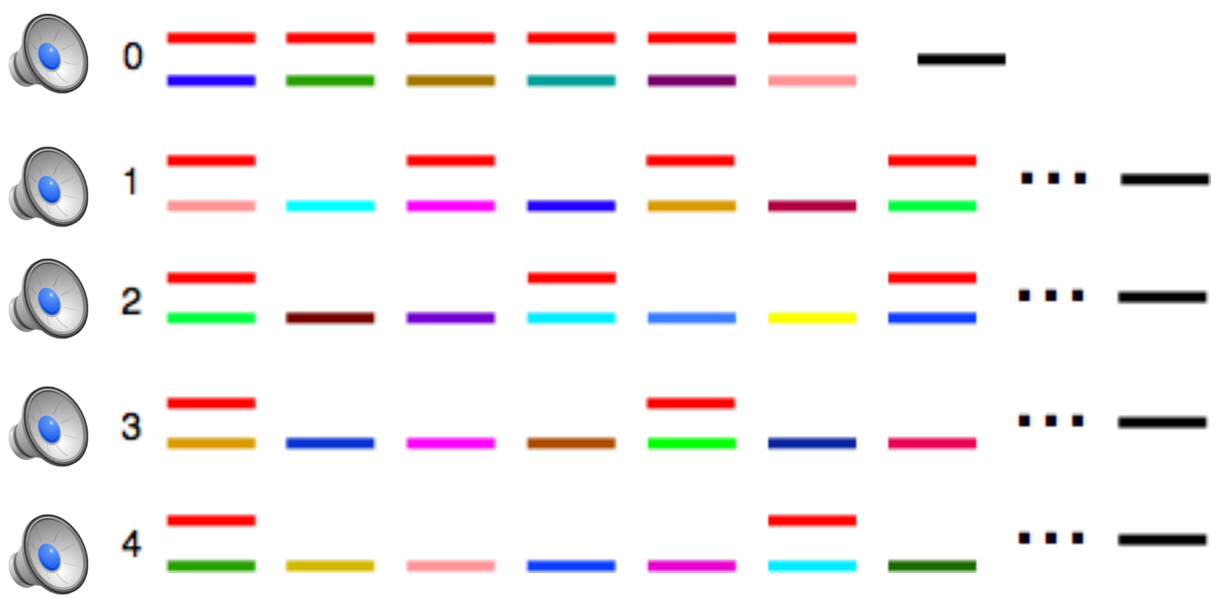
Can test by varying ISI:



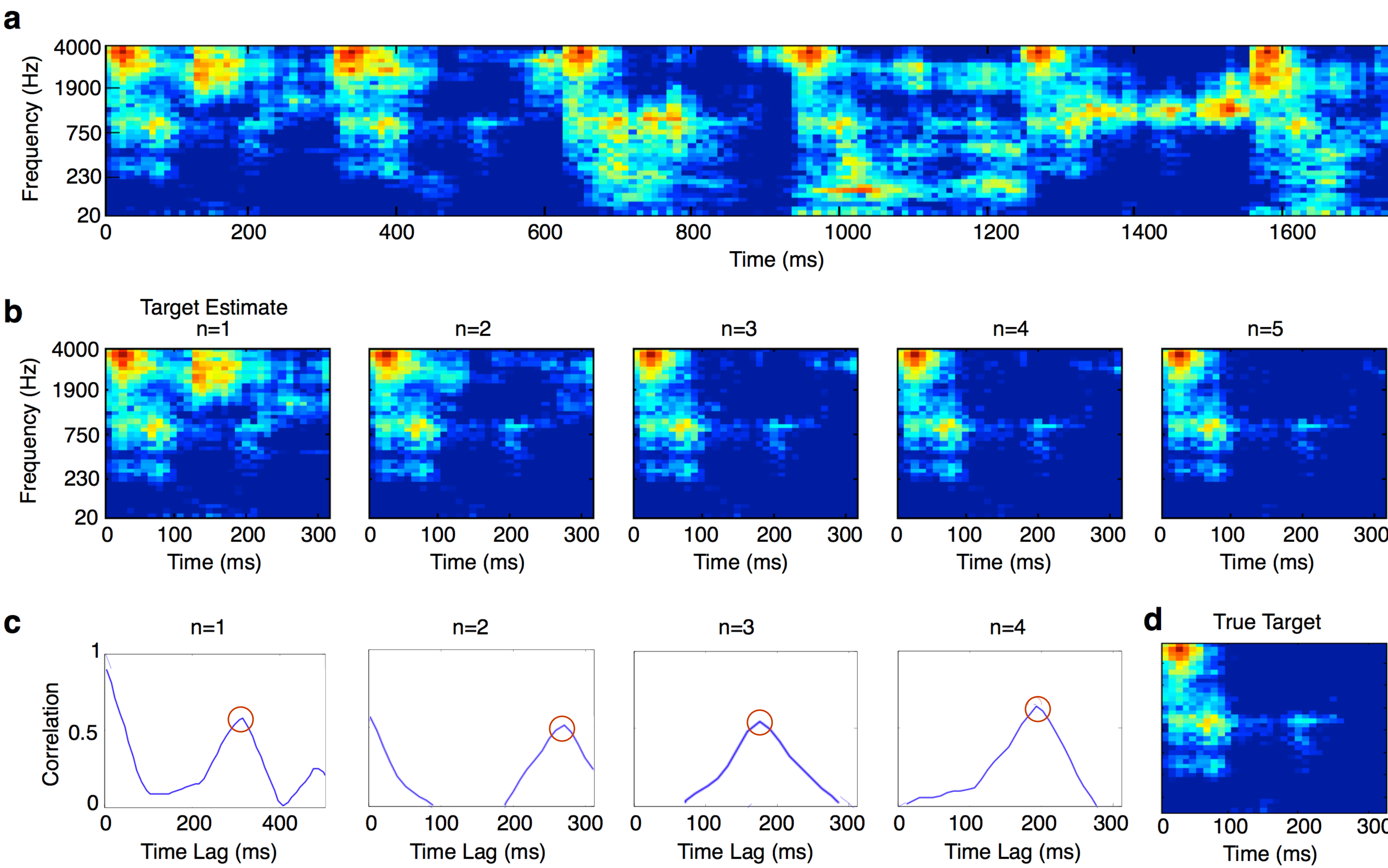
Performance declines once targets are spaced by > ~400 ms:



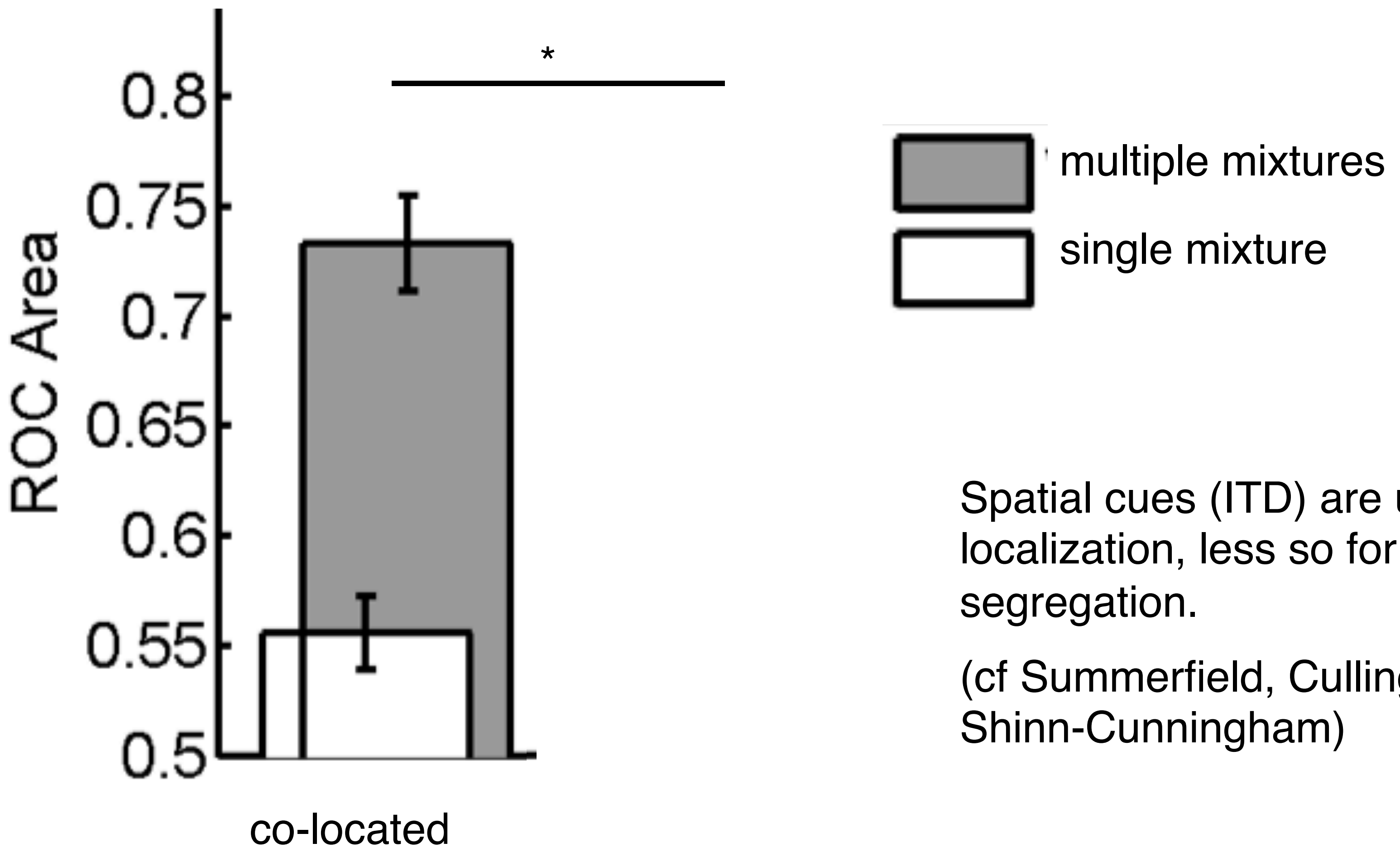
Suggests listeners combine information across presentations, using short-term buffer.



Proof of concept: target can be extracted via cross-correlation.



How does repetition compare to spatial separation?



Spatial cues (ITD) are useful for localization, less so for segregation.

(cf Summerfield, Culling, Darwin, Shinn-Cunningham)

Schwartz, McDermott, Shinn-Cunningham, JASA 2012

•Listeners can recognize sound sources from mixtures, if presented more than once across different mixtures.

•Repetition is not explicit in the auditory input, but inc
auditory system detects, uses to infer sources.

•Repetition can bootstrap sound segregation in the absence of bottom-up grouping cues, knowledge of sounds.

There are lots of repeating sounds in natural auditory environments for which this could be relevant, e.g. animal vocalizations.

Music perception may co-opt this mechanism.

Outline

I. Introduction
II. How humans use repetition to identify sound sources (McDermott)
III. Coffee break
IV. Repetition-based algorithms for source separation (Rafii)
V. Links to other methods for source separation
VI. Conclusions/Questions

Coffee Break

http://coffee-urn-info.blogspot.com/2011/08/clean-his-coffee-cup-was.html

Outline

I. Introduction
II. How humans use repetition to identify sound sources (McDermott)
III. Coffee break
IV. Repetition-based algorithms for source separation (Rafii)
V. Links to other methods for source separation
VI. Conclusions/Questions

REpeating Pattern Extraction Technique (REPET)

Zafar Rafii

NORTHWESTERN UNIVERSITY

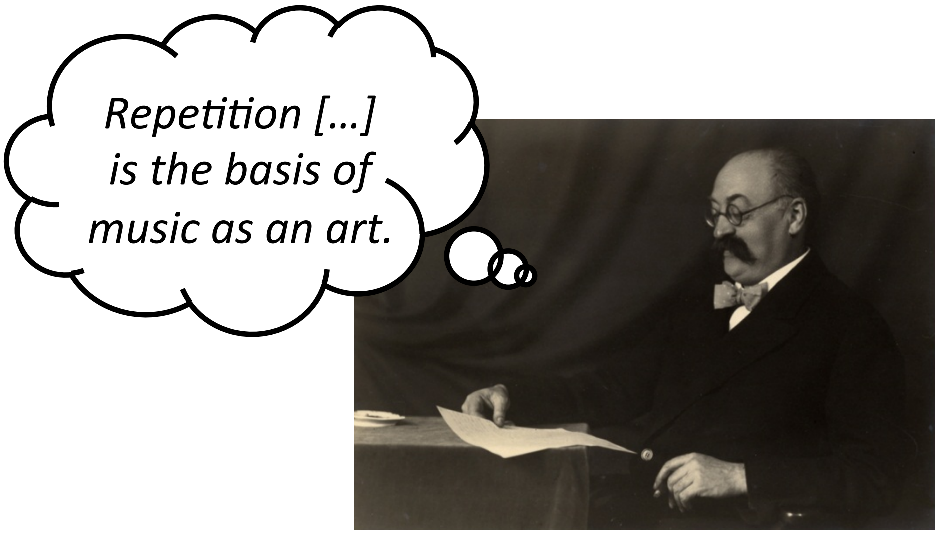
interactive audio lab

Outline

I. **Introduction**
II. REPET
III. REPET-SIM
IV. Conclusion

Introduction

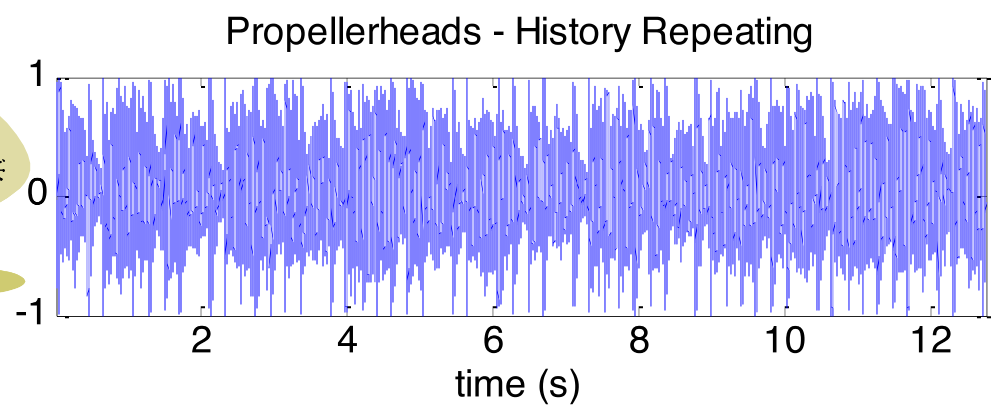
- **Repetition** is a fundamental element in generating and perceiving structure



Heinrich Schenker (1868-1935)

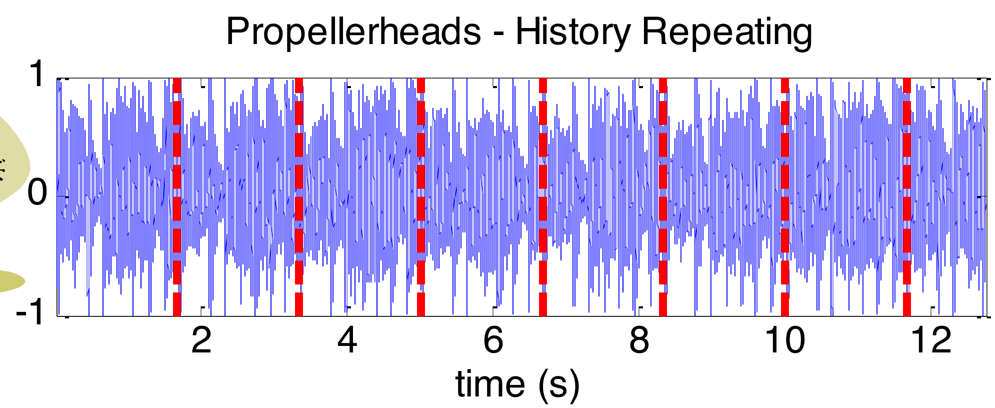
Introduction

- In music, pieces are often characterized by an underlying **repeating structure** over which varying elements are superimposed



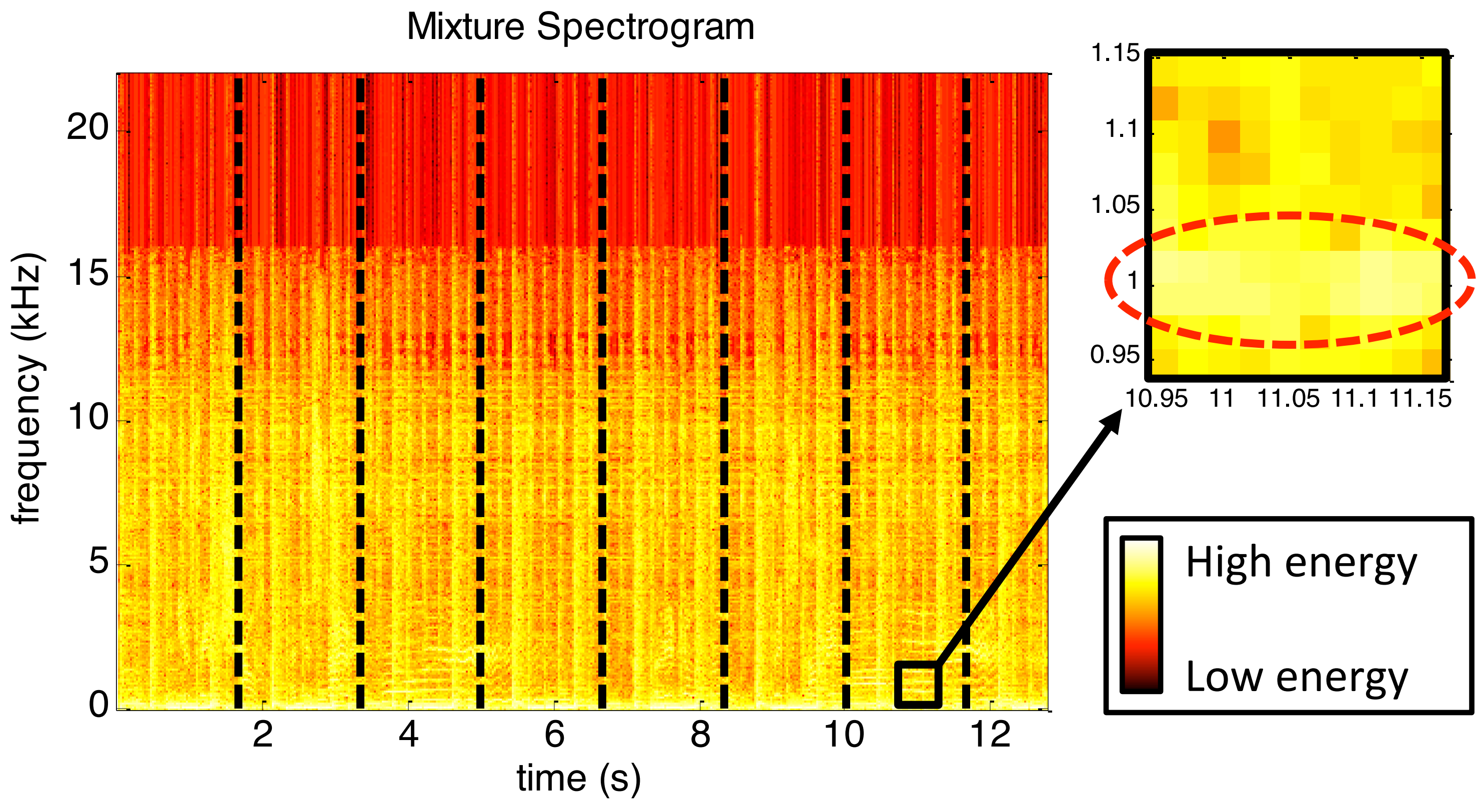
Introduction

- In music, pieces are often characterized by an underlying **repeating structure** over which varying elements are superimposed



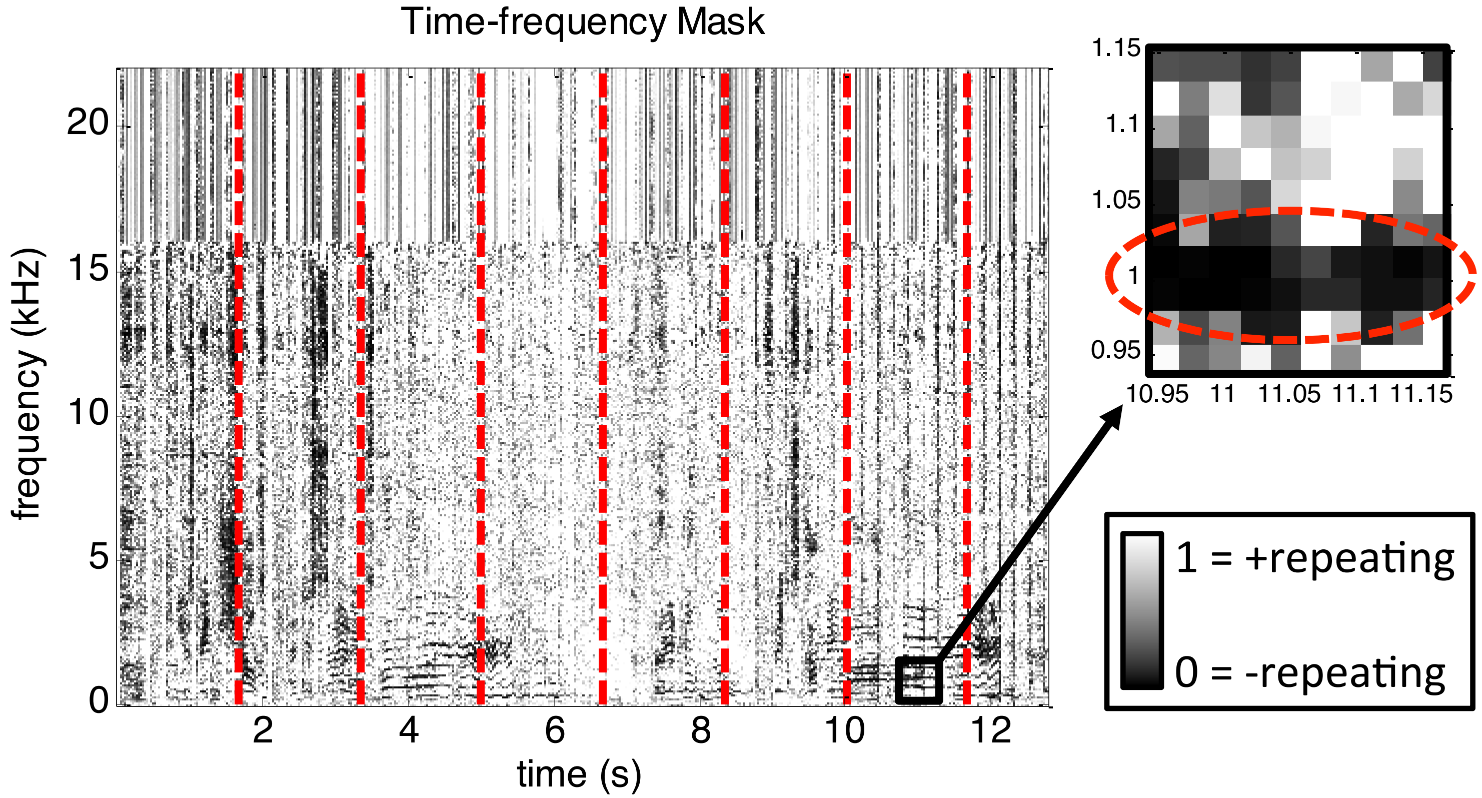
Introduction

- This means there should be patterns that are more or less **repeating in time and frequency**



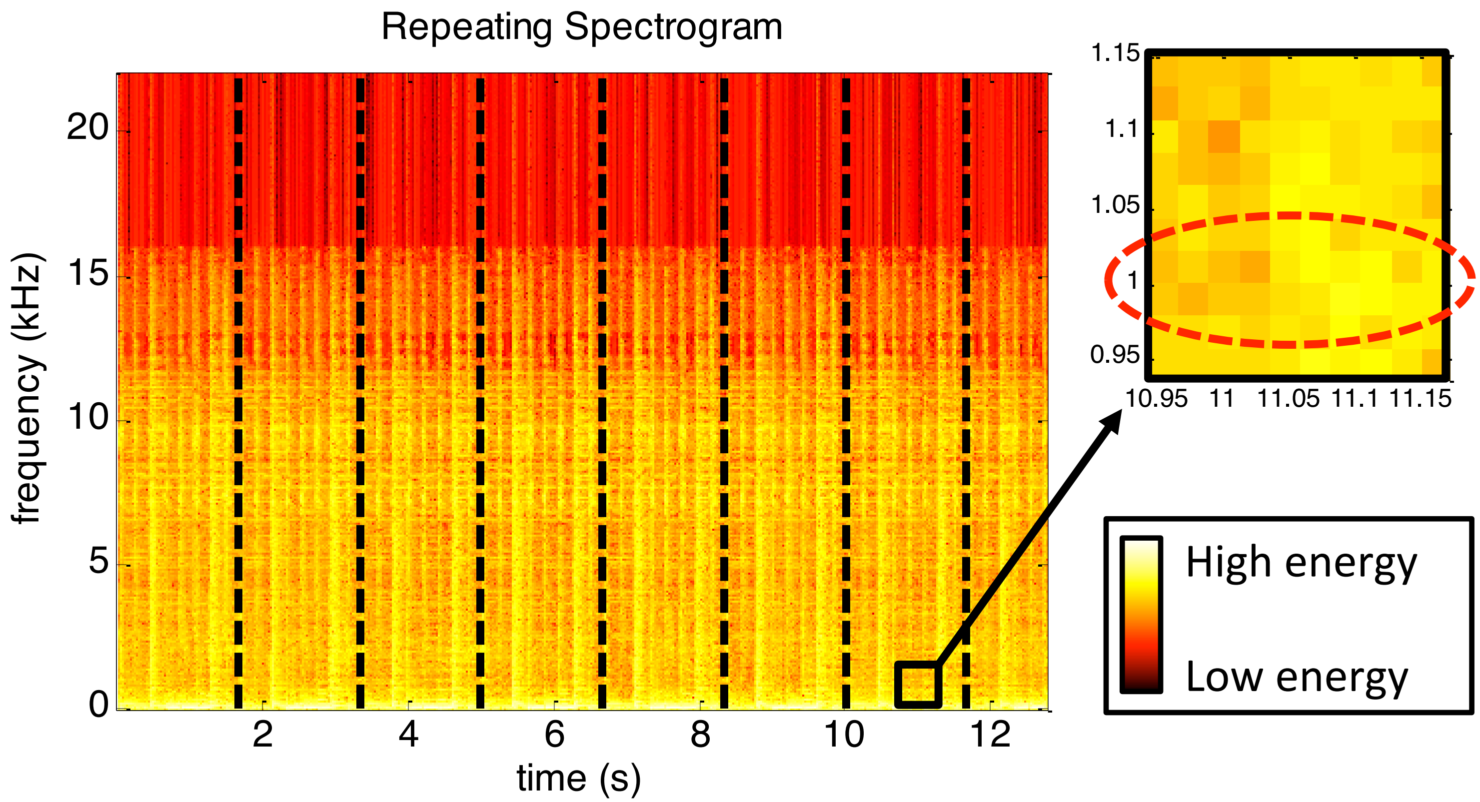
Introduction

- The (more or less) repeating patterns could be identified using a **time-frequency mask**



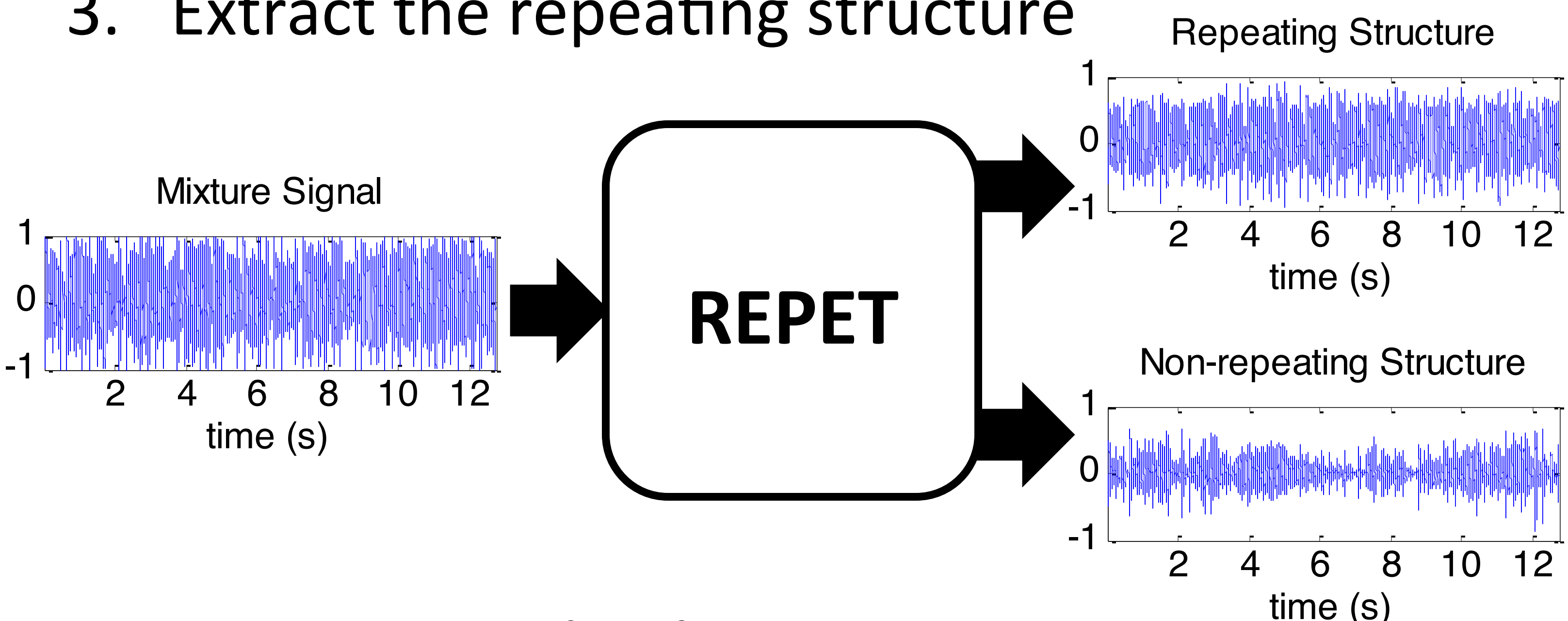
Introduction

- The t-f mask could then be applied on the mixture to extract the **repeating patterns**



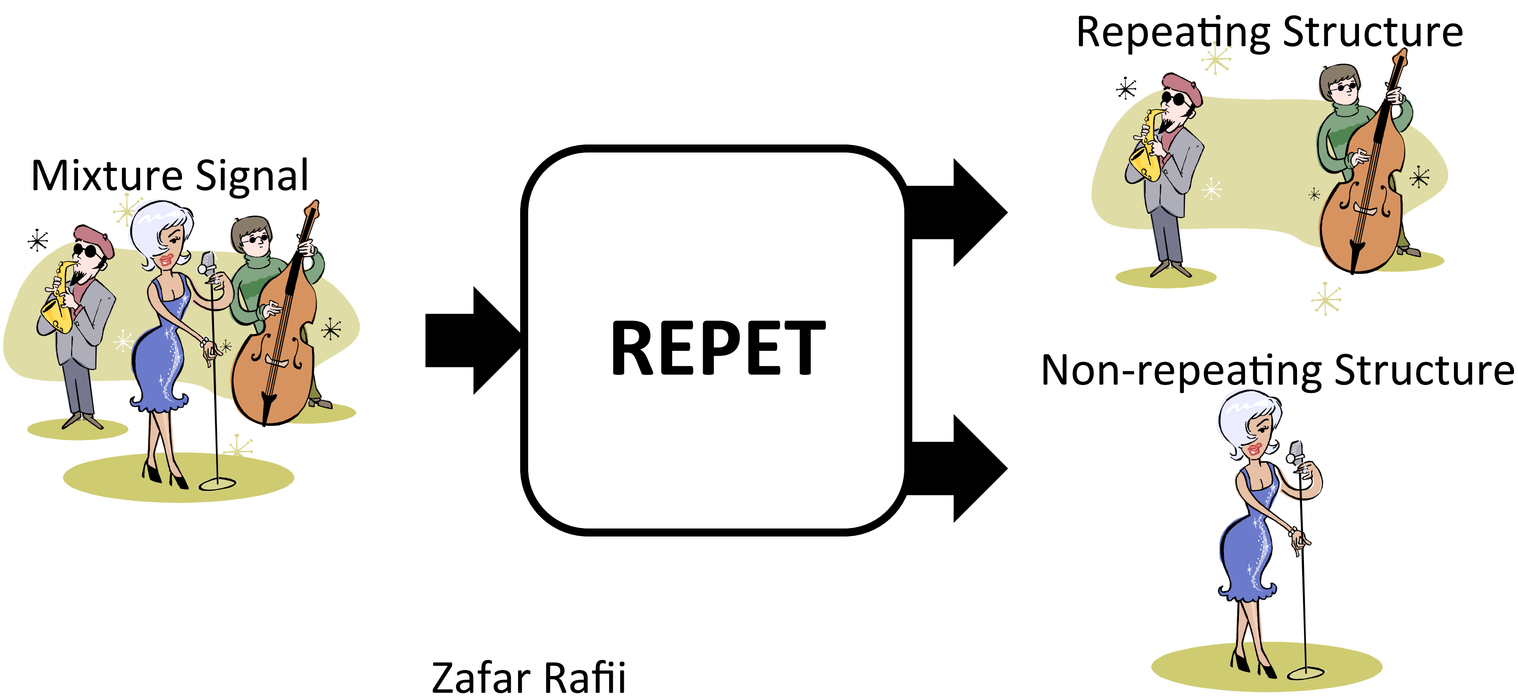
Introduction

- **REpeating Pattern Extraction Technique!**
  1. Identify the repeating elements
  2. Derive a repeating model
  3. Extract the repeating structure



Introduction

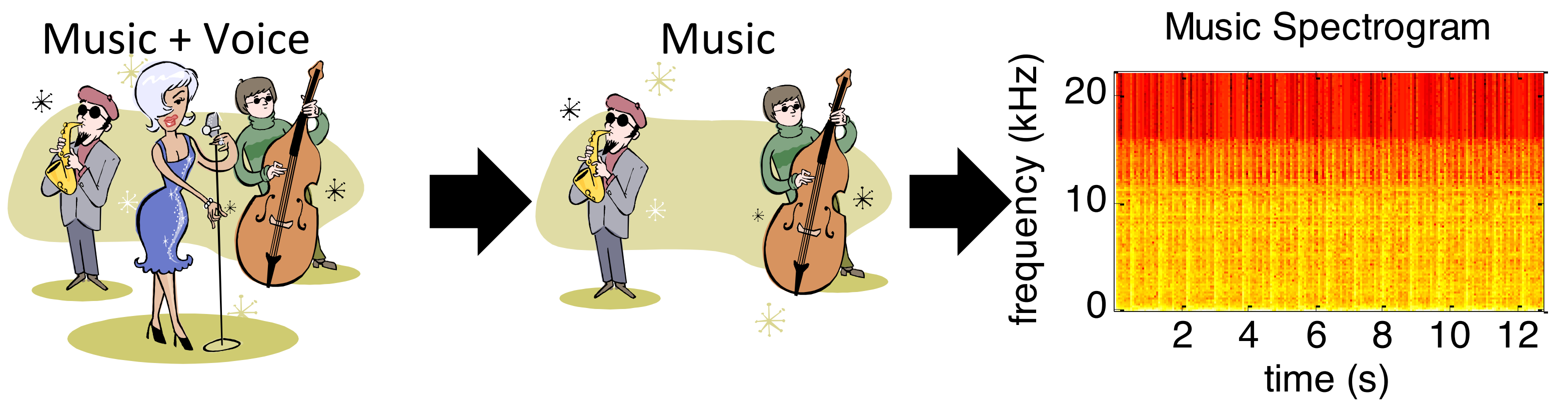
- Simple **music/voice separation** method!
  - Repeating structure ≈ musical background
  - Non-repeating structure ≈ vocal foreground



Introduction

- **Assumptions:**
  - The repeating background is **dense & low-ranked**

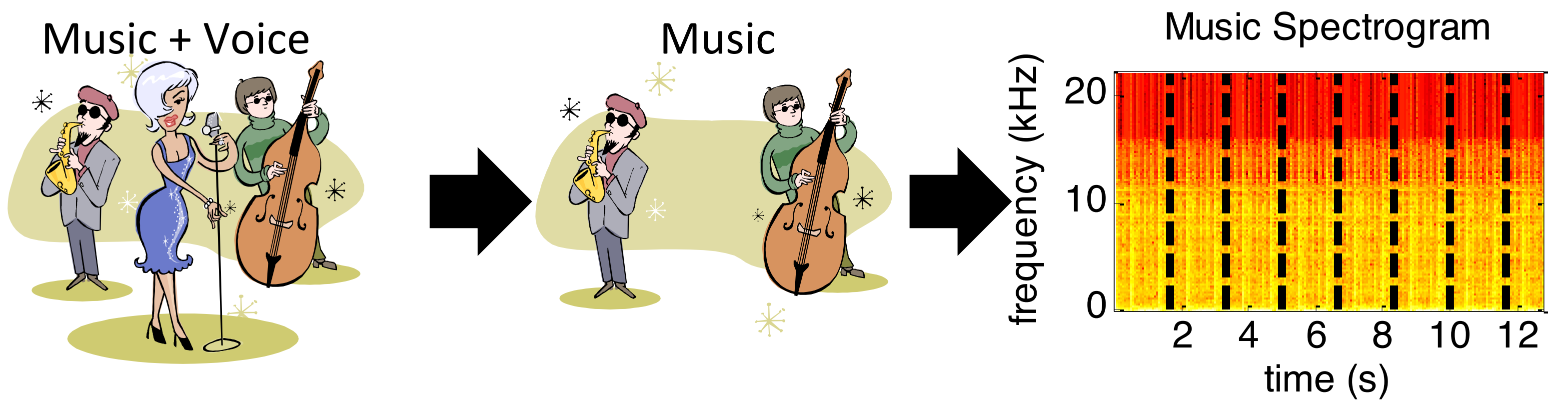
  → often true for **music** in a mixture of music + voice



Introduction

- **Assumptions:**
  - The repeating background is **dense & low-ranked**

  → low-ranked = repetitions at some **period rate**



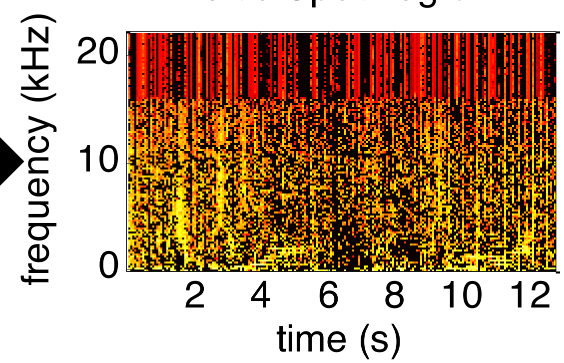
Introduction

- **Assumptions:**
  – The non-repeating foreground is **sparse & varied**
  → often true for **voice** in a mixture of music + voice

Music + Voice

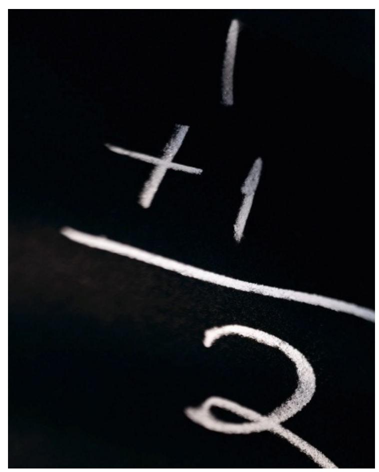
Voice

Voice Spectrogram



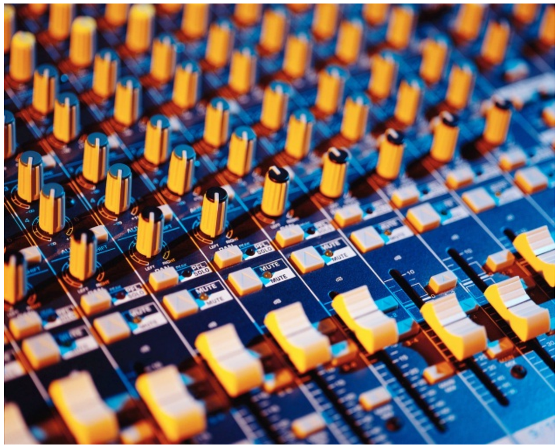
Introduction

- **Practical advantages:**
  - Does not depend on special parameterizations
  - Does not rely on complex frameworks
  - Does not require prior training



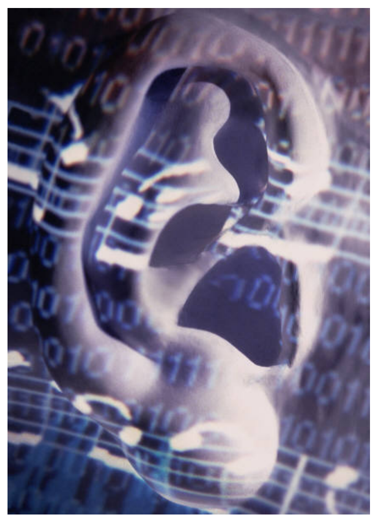
Introduction

- **Practical interests:**
  - Audio post processing
  - Melody extraction
  - Karaoke gaming



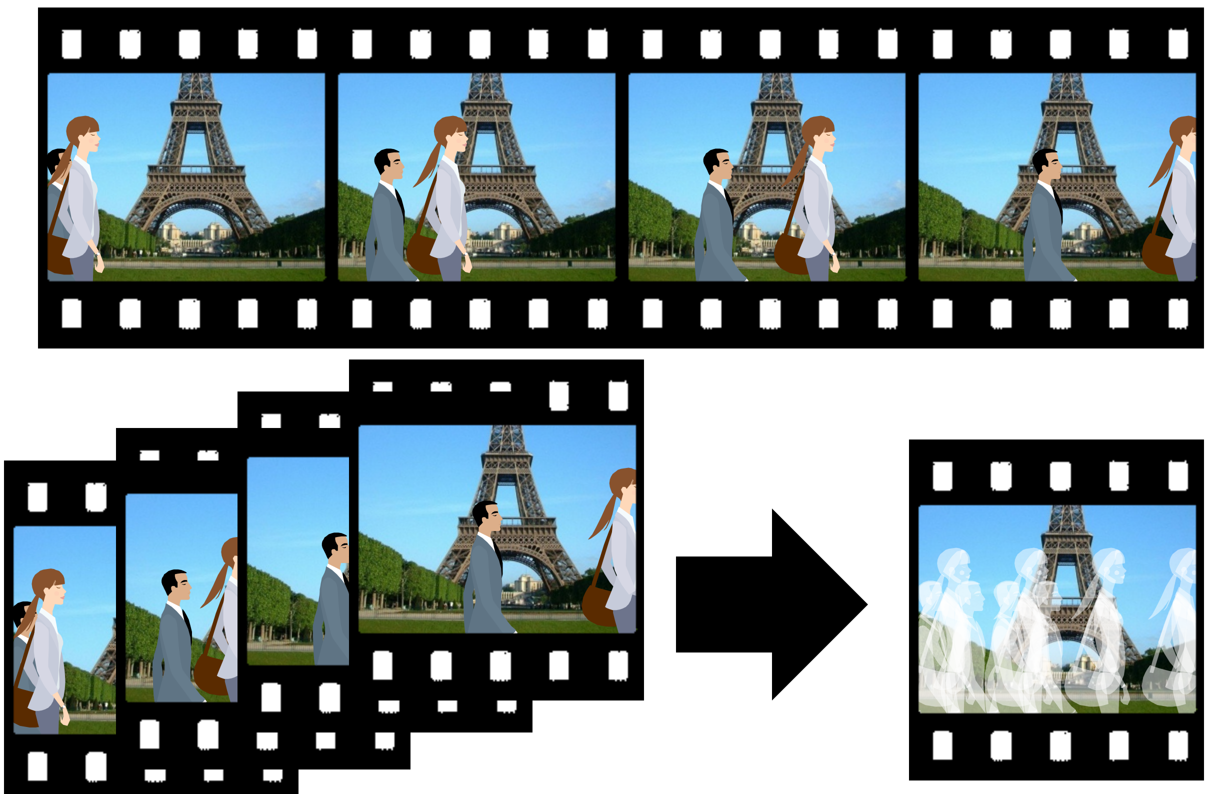
Introduction

- **Intellectual interests:**
  - Music perception
  - Music understanding
  - Simply based on repetition!



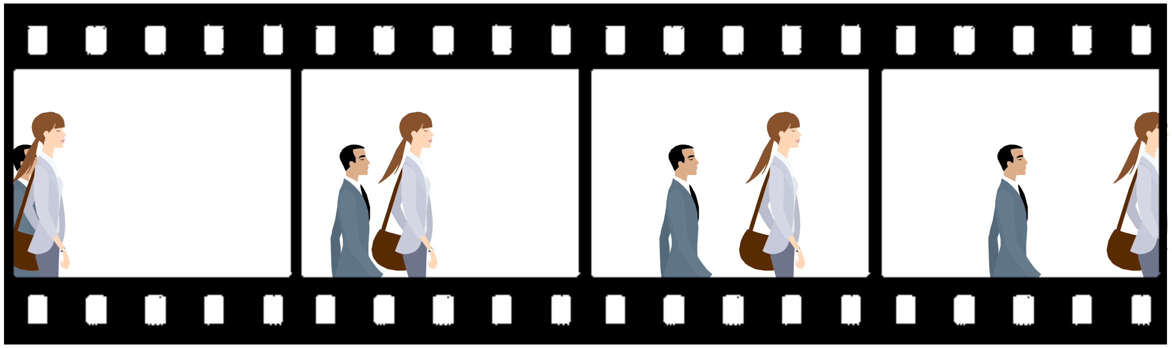
Introduction

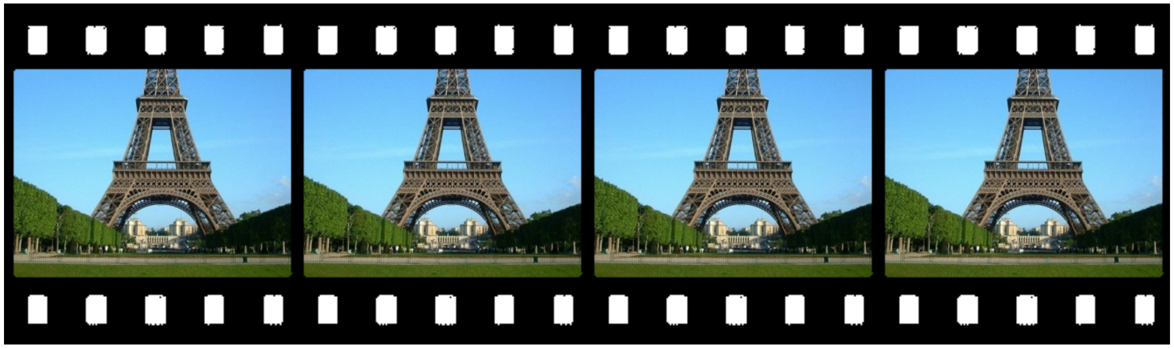
- Parallel with **background subtraction** in vision
  - Compare frames to estimate a background model



Introduction

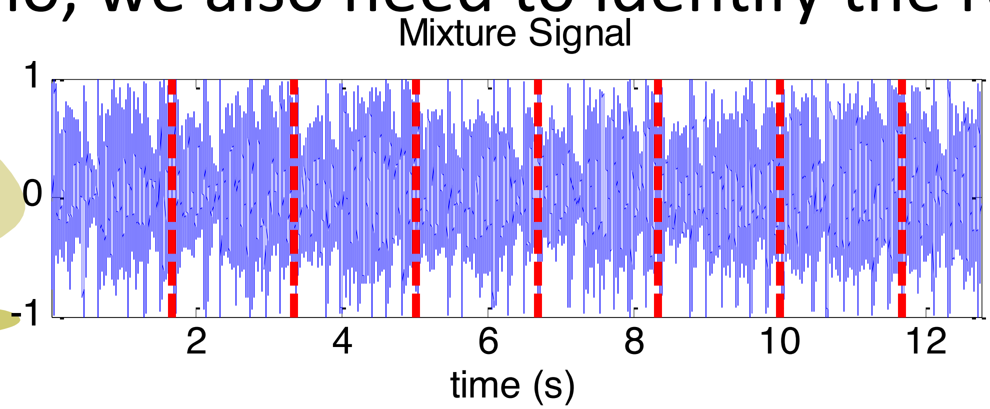
- Parallel with **background subtraction** in vision
  - Extract the background from the foreground





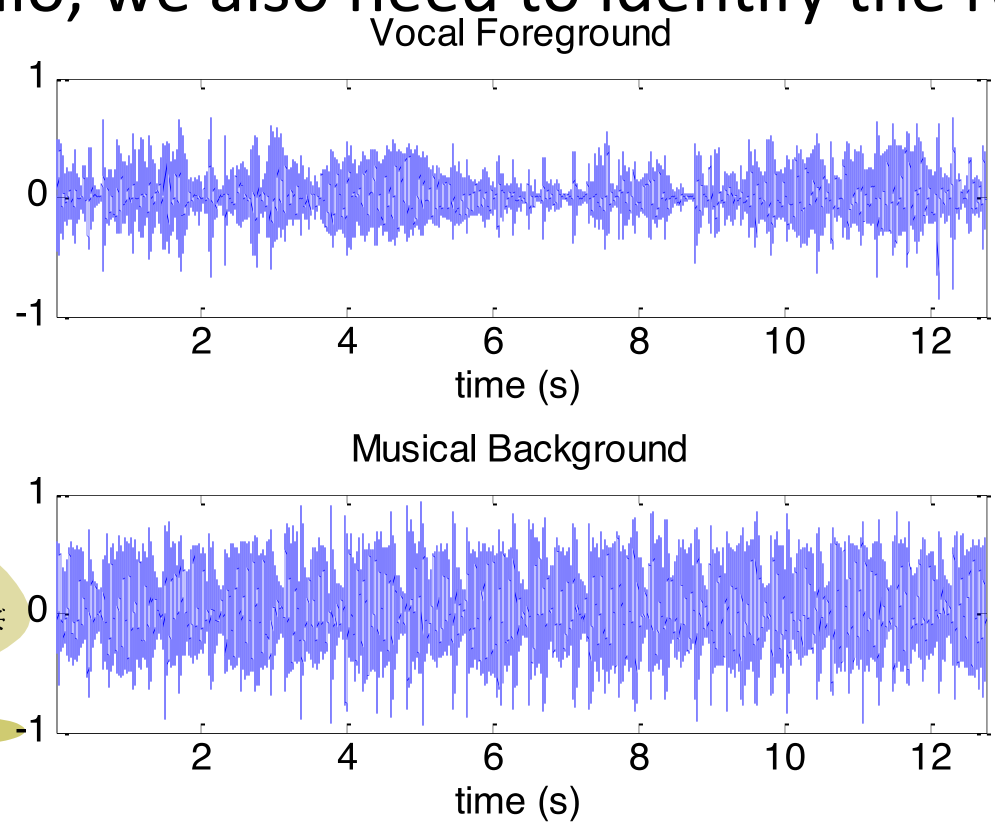
Introduction

- Parallel with **background subtraction** in vision
  - In audio, we also need to identify the repetitions!



Introduction

- Parallel with **background subtraction** in vision
  - In audio, we also need to identify the repetitions!



Outline

I. Introduction

II. REPET

  1. Method

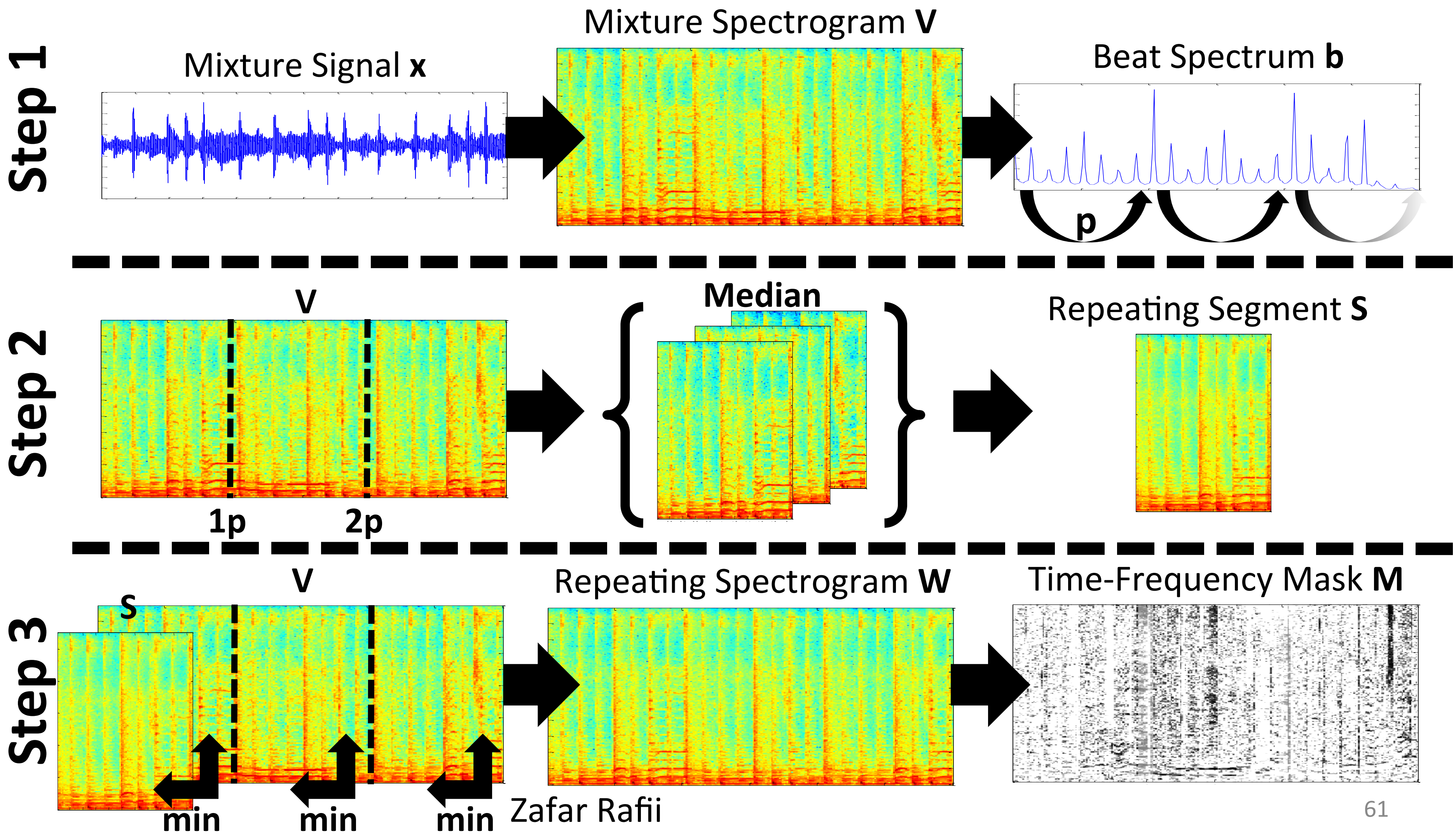
  2. Extensions

  3. Evaluation

III. REPET-SIM

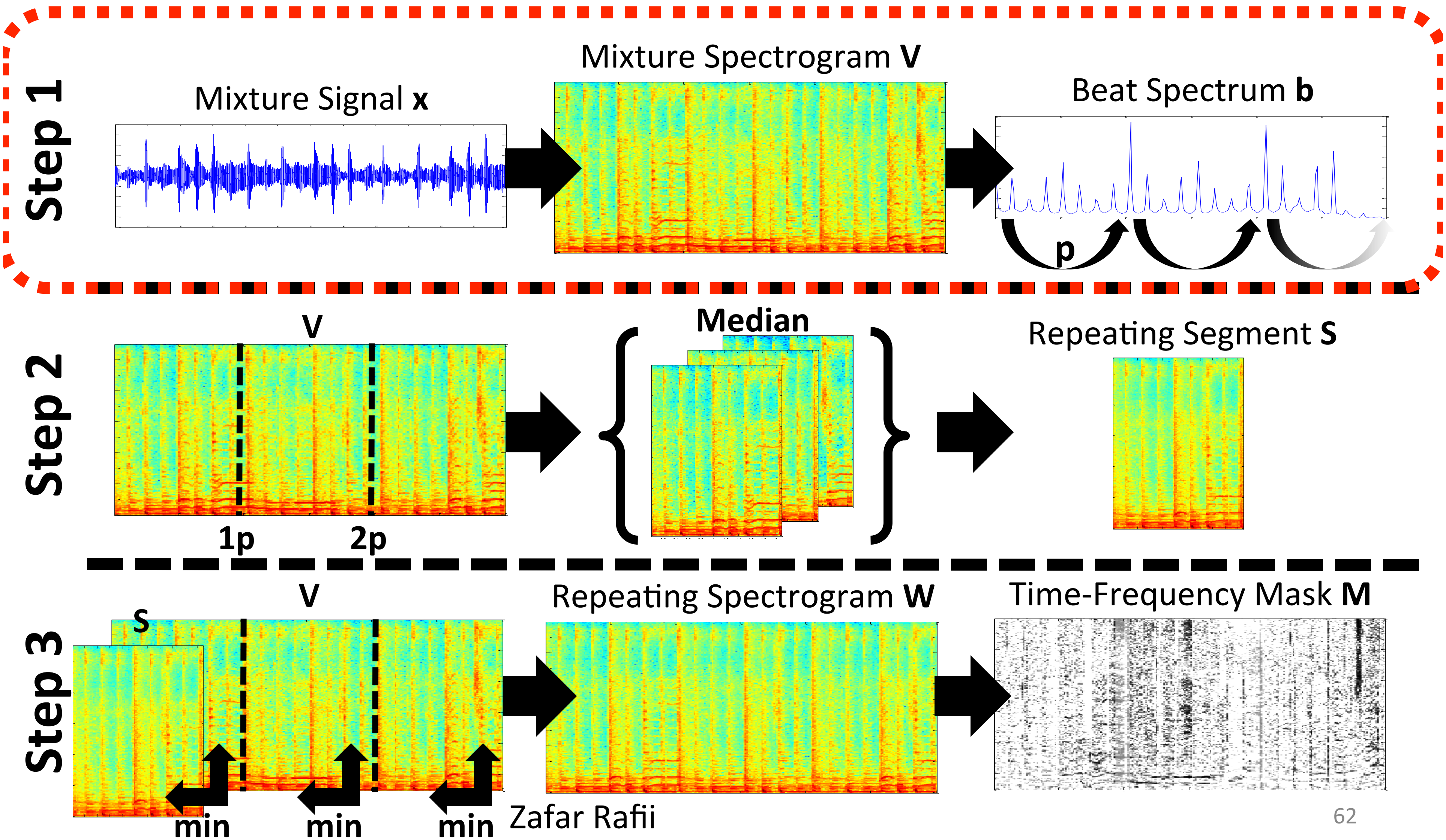
IV. Conclusion

Method



Zafar Rafii

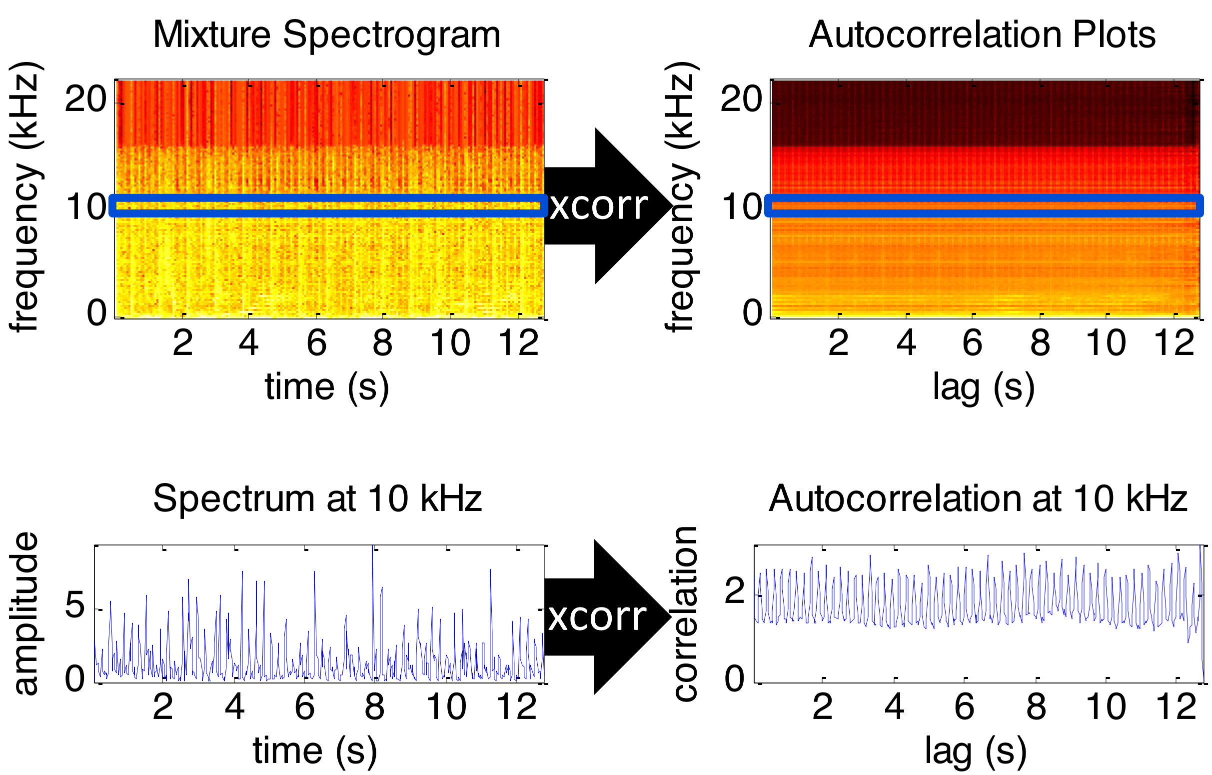
1. Repeating Period



Zafar Rafii

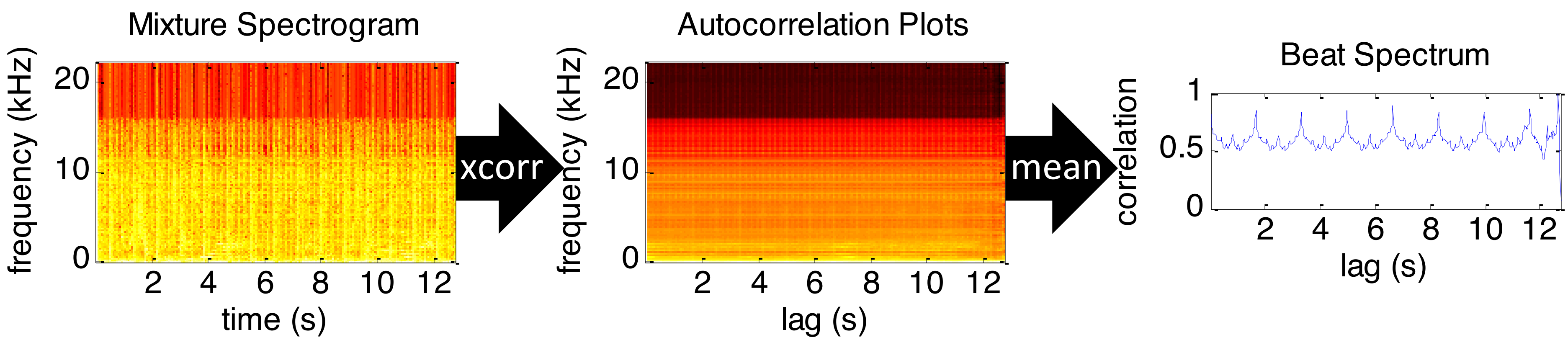
1. Repeating Period

- We compute the **autocorrelations** of the rows of the spectrogram to find periodicities



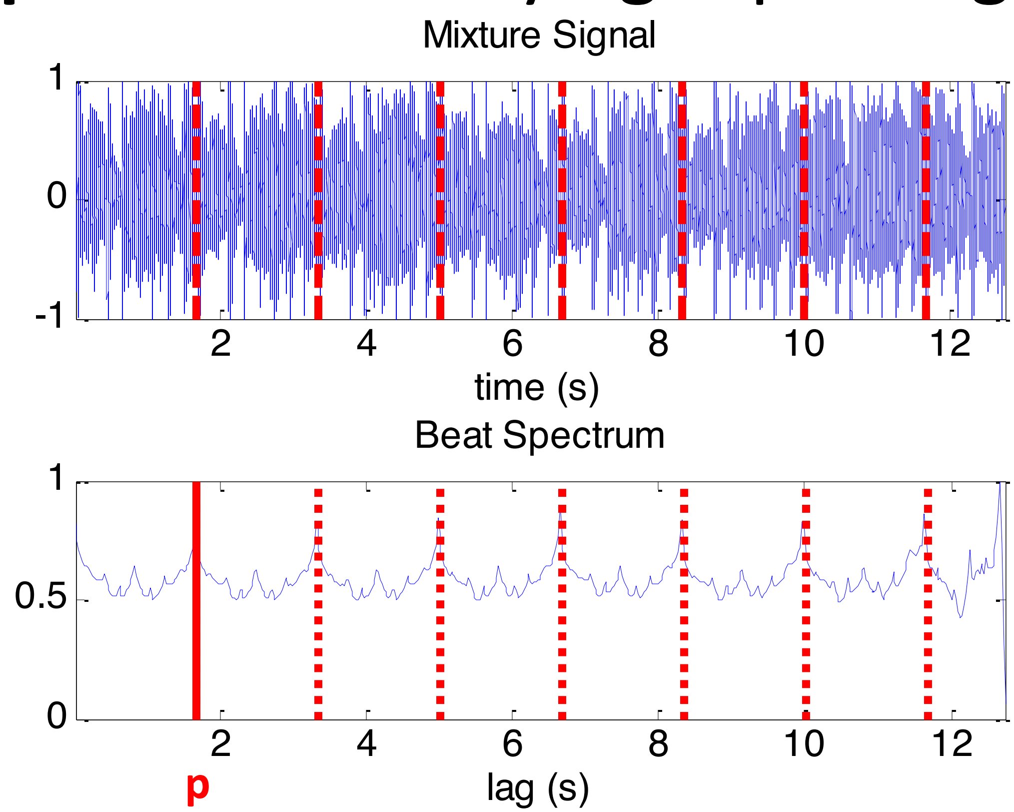
1. Repeating Period

- We take the mean of the autocorrelations (rows) and obtain the **beat spectrum**



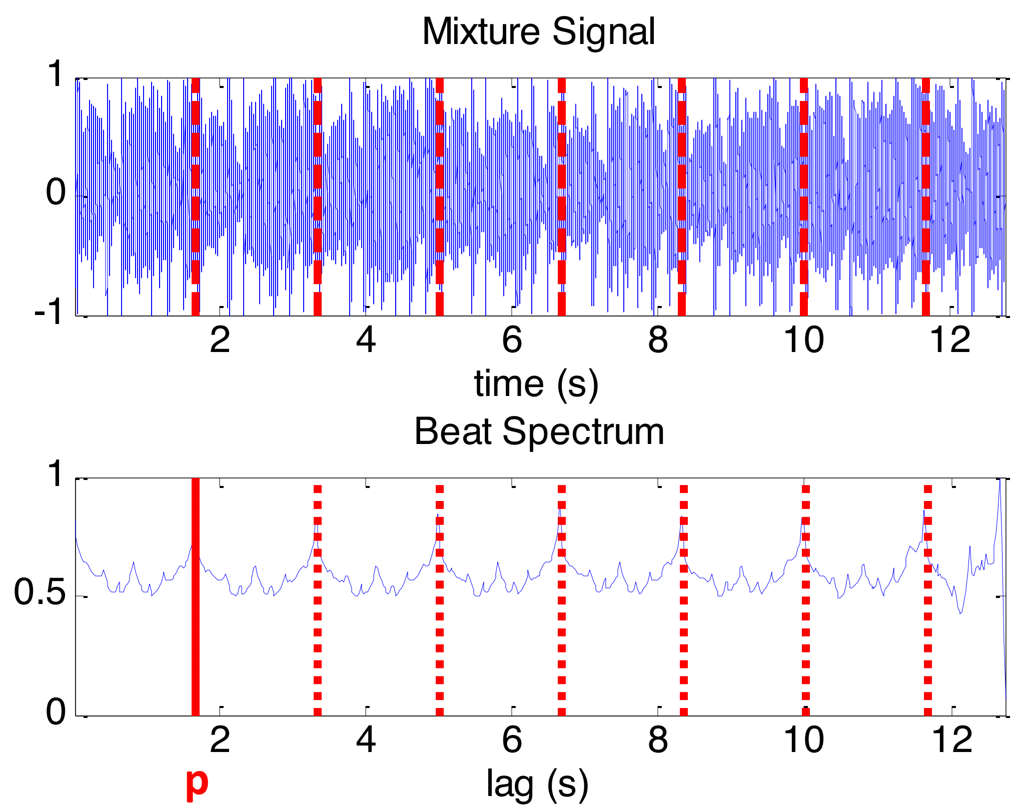
1. Repeating Period

- The beat spectrum reveals the **repeating period p** of the underlying repeating structure

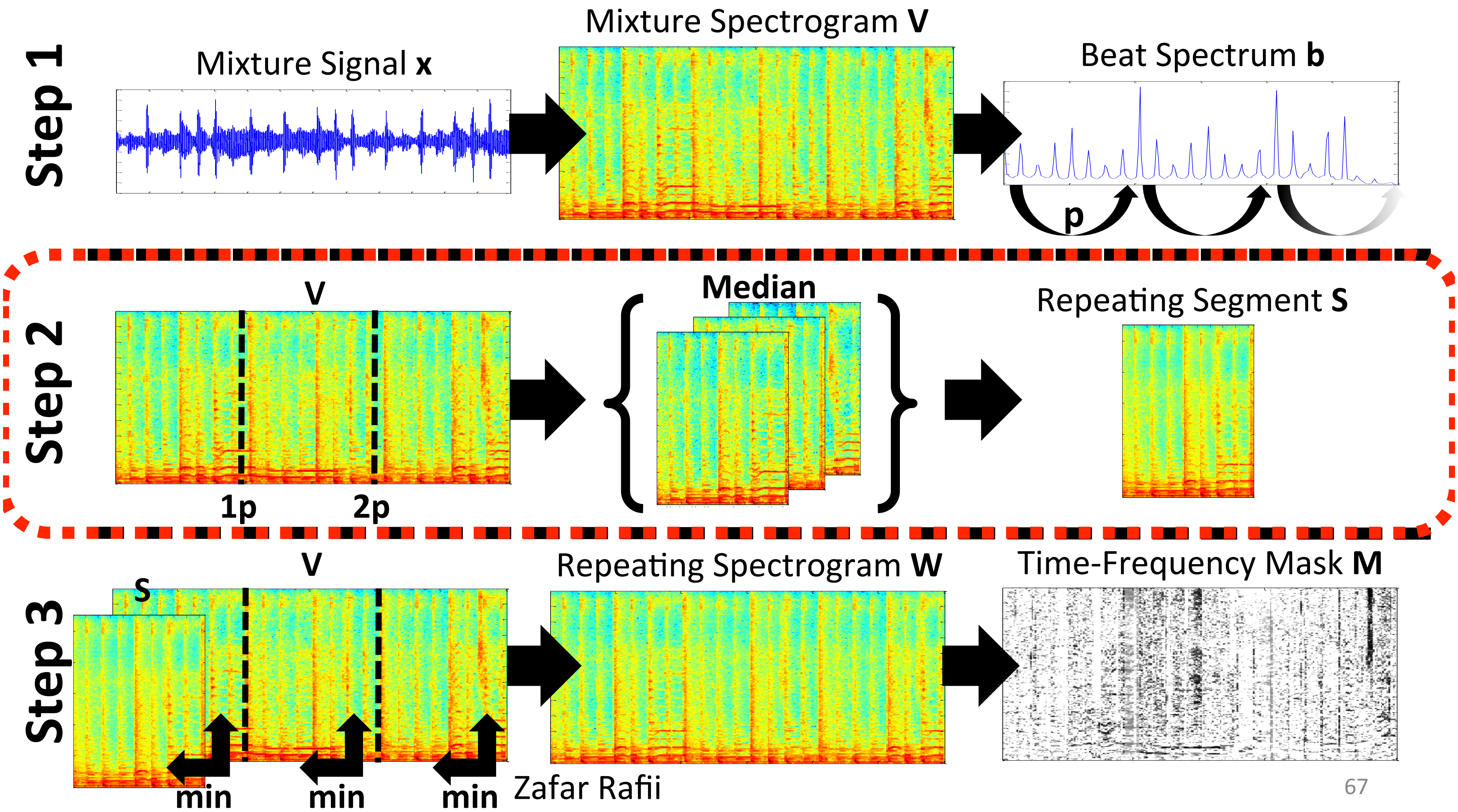


1. Repeating Period

- We assume here that the background is **more dense and low-ranked** than the foreground



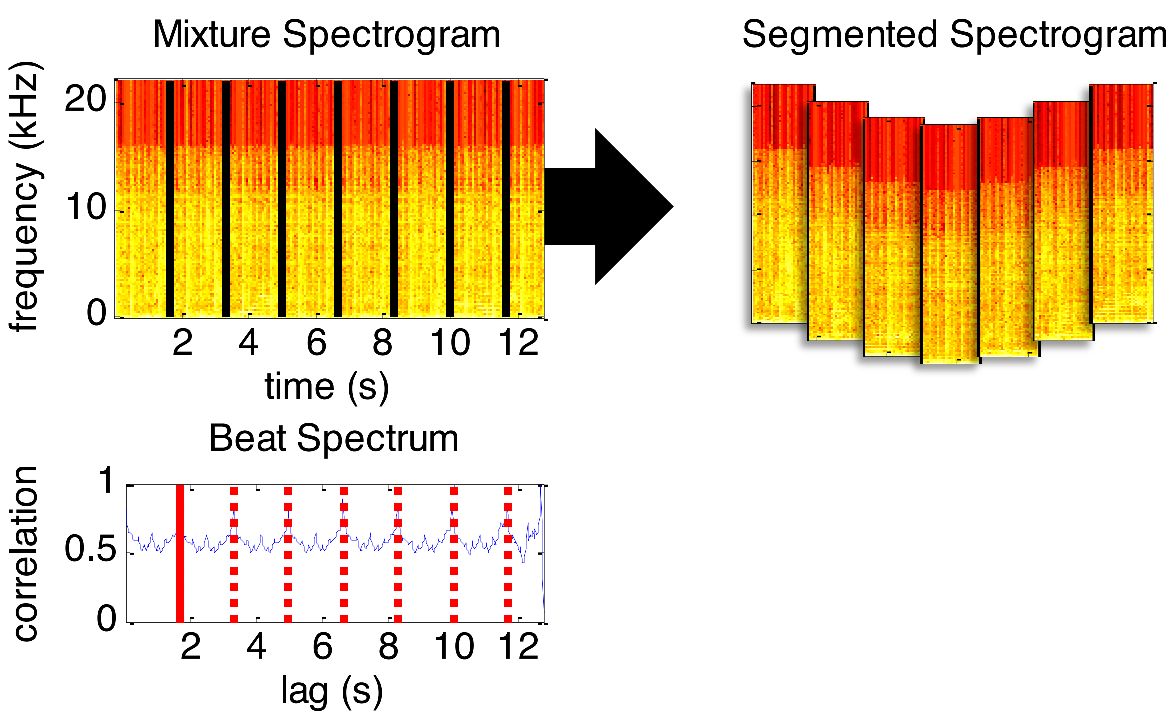
2. Repeating Segment



Zafar Rafii

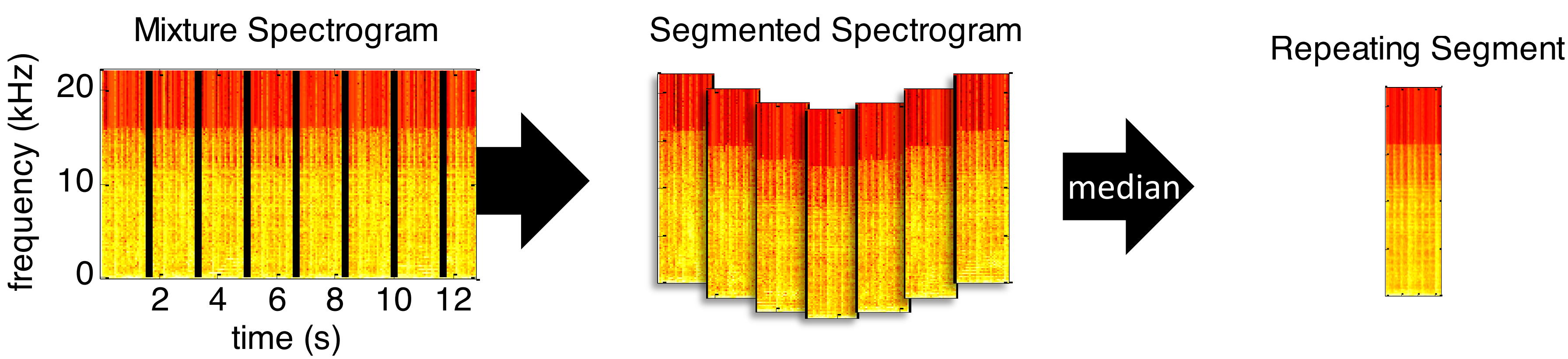
2. Repeating Segment

- The repeating period is then used to **segment** the mixture spectrogram at period rate



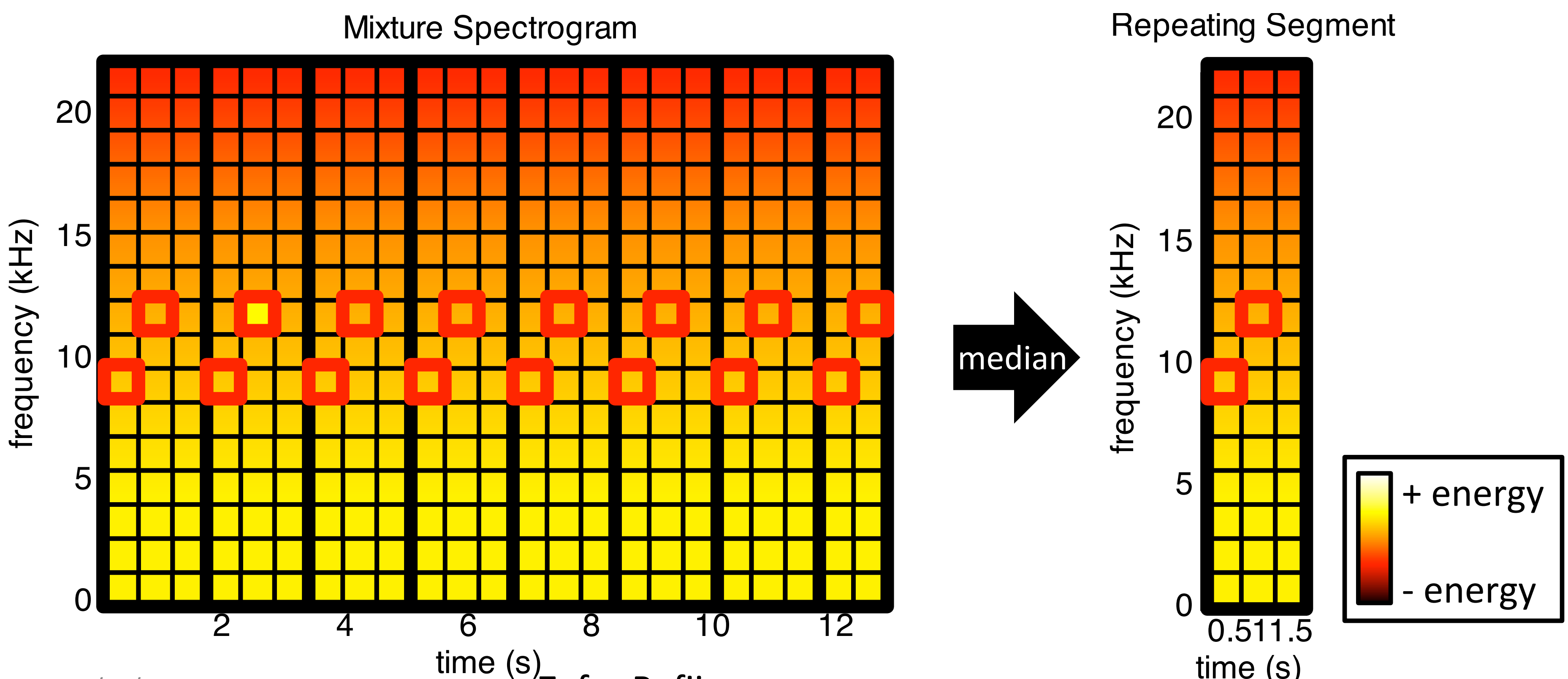
2. Repeating Segment

- The **repeating segment model** is calculated as the element-wise median of the segments



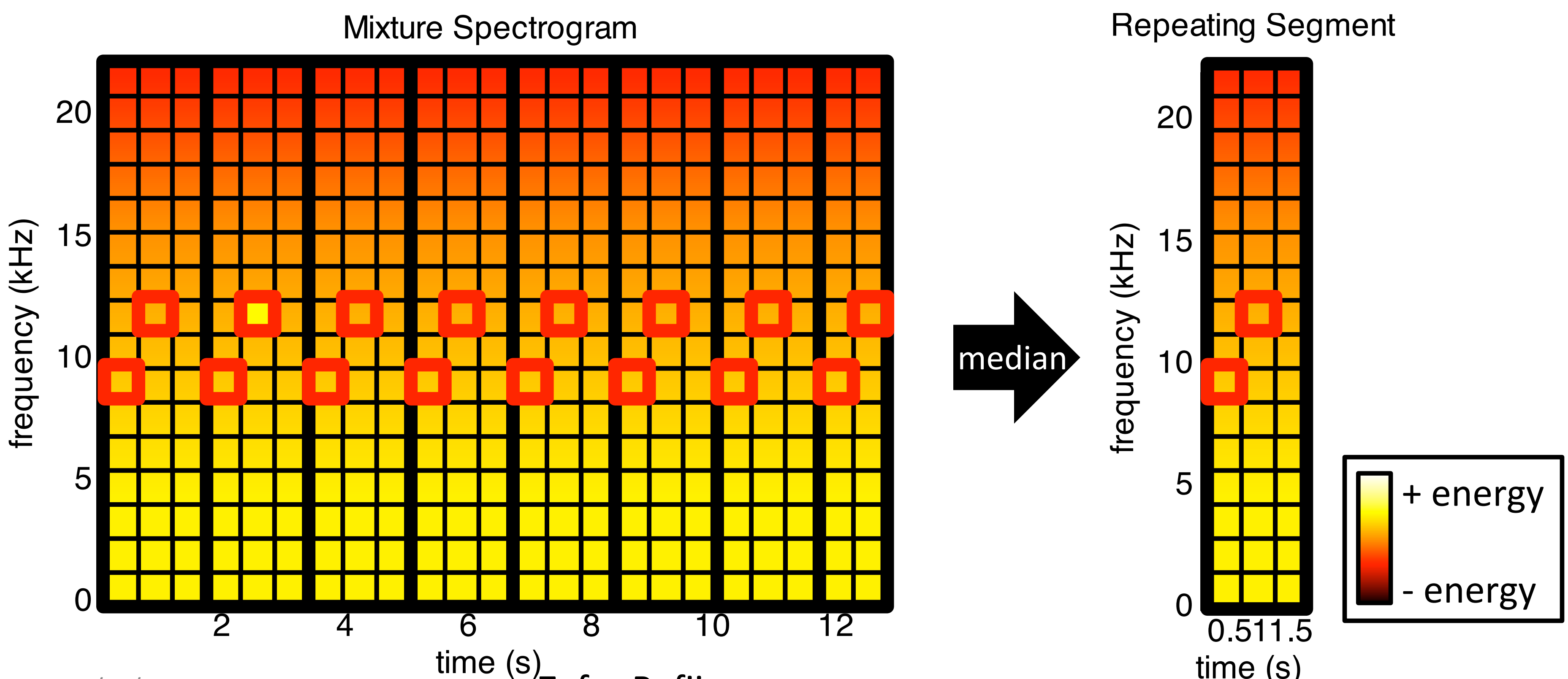
2. Repeating Segment

- The **median** helps to derive a clean repeating segment, removing the non-repeating outliers

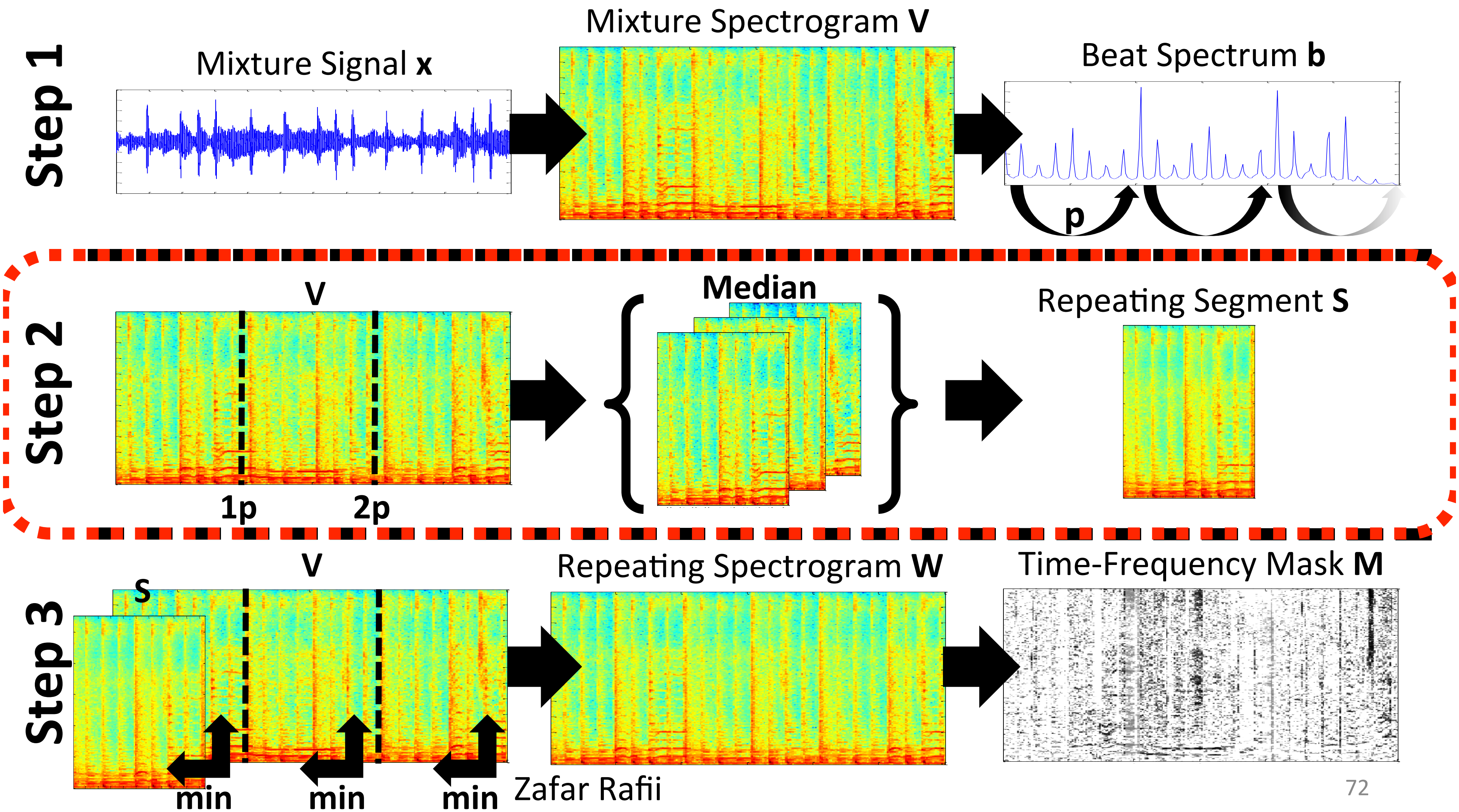


2. Repeating Segment

- We assume here that the foreground is **more sparse and varied** than the background



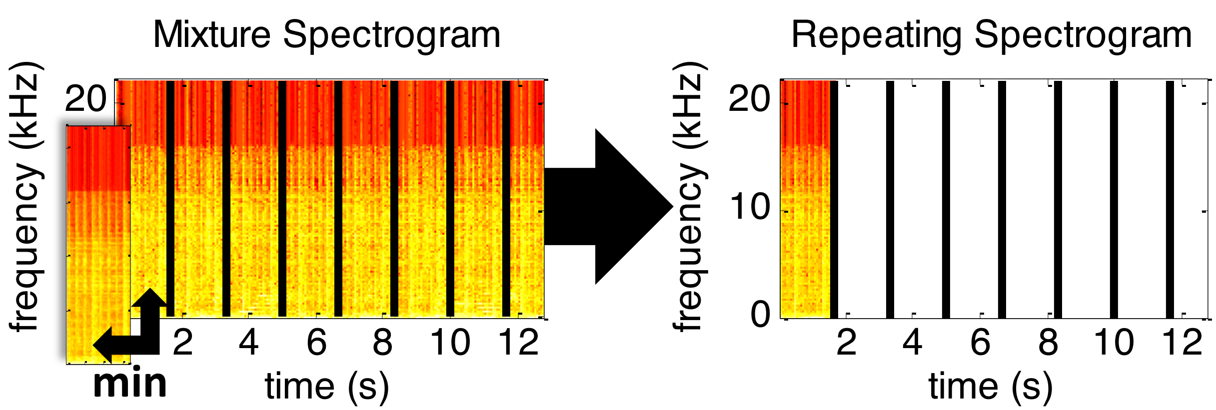
2. Repeating Segment



Zafar Rafii

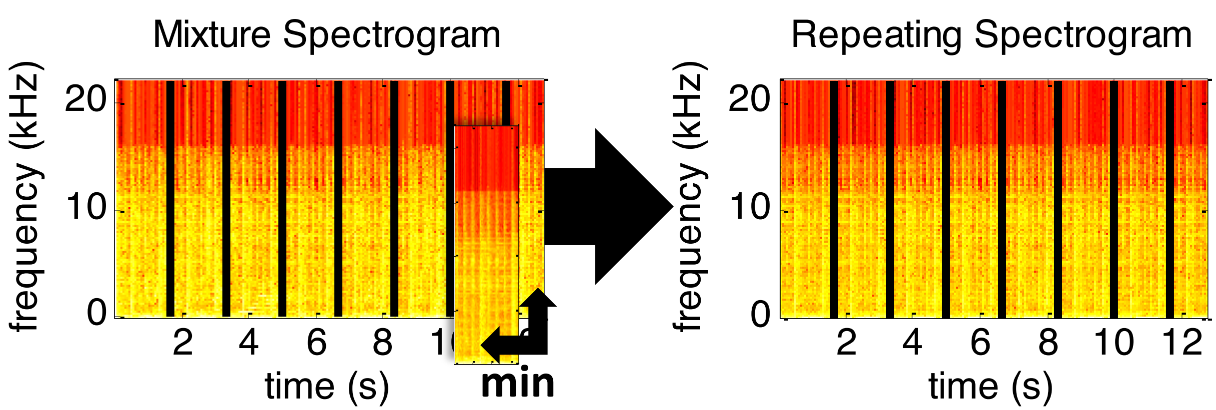
3. Repeating Structure

- We take the element-wise **minimum** between the repeating segment and the segments



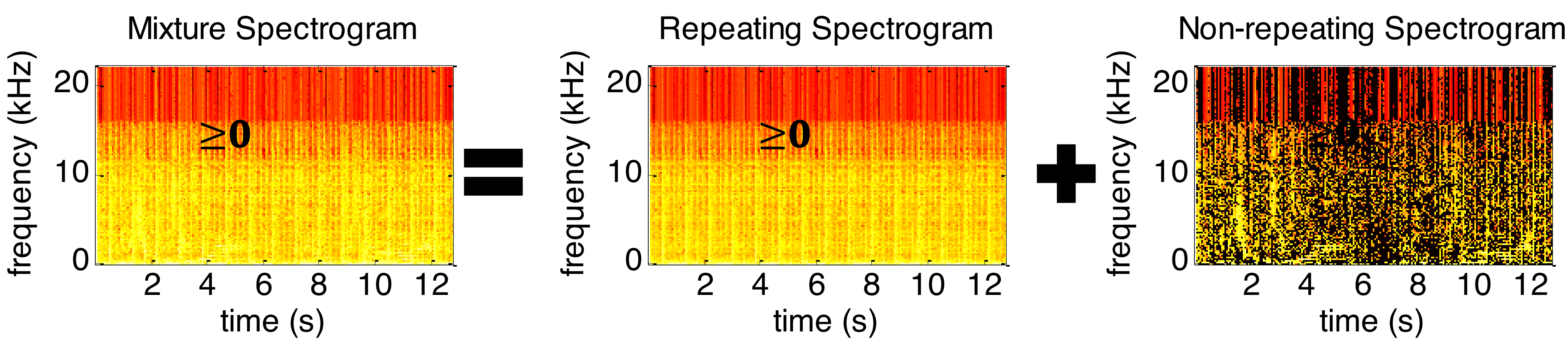
3. Repeating Structure

- We obtain a **repeating spectrogram model** for the repeating background



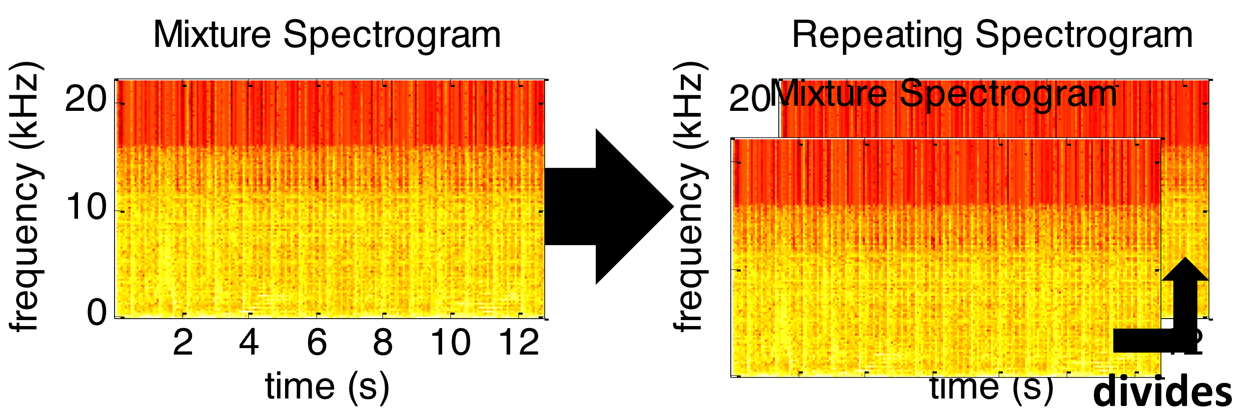
3. Repeating Structure

- The repeating spectrogram **cannot have values higher than** the mixture spectrogram



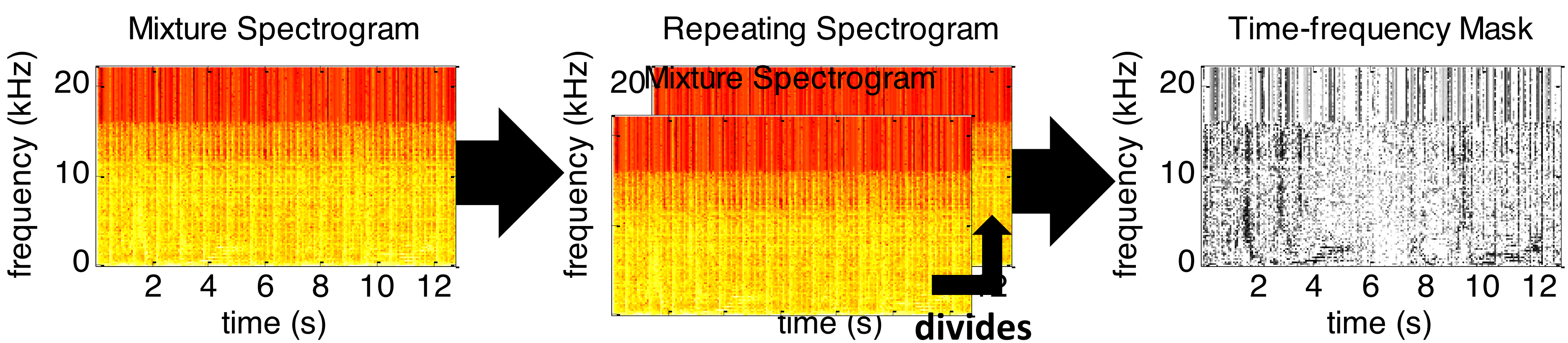
3. Repeating Structure

- We **divide** the repeating spectrogram by the mixture spectrogram, element-wise



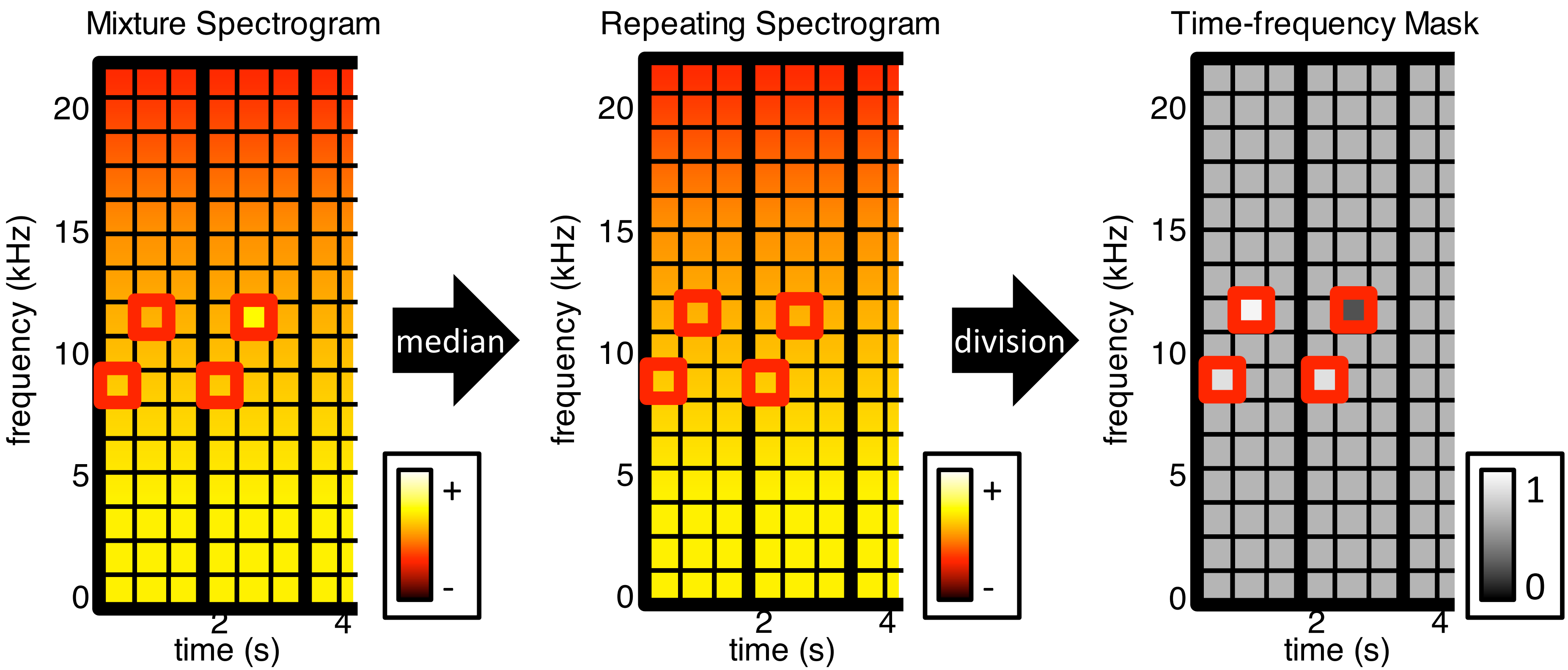
3. Repeating Structure

- We obtain a **soft time-frequency mask** (with values in [0,1])



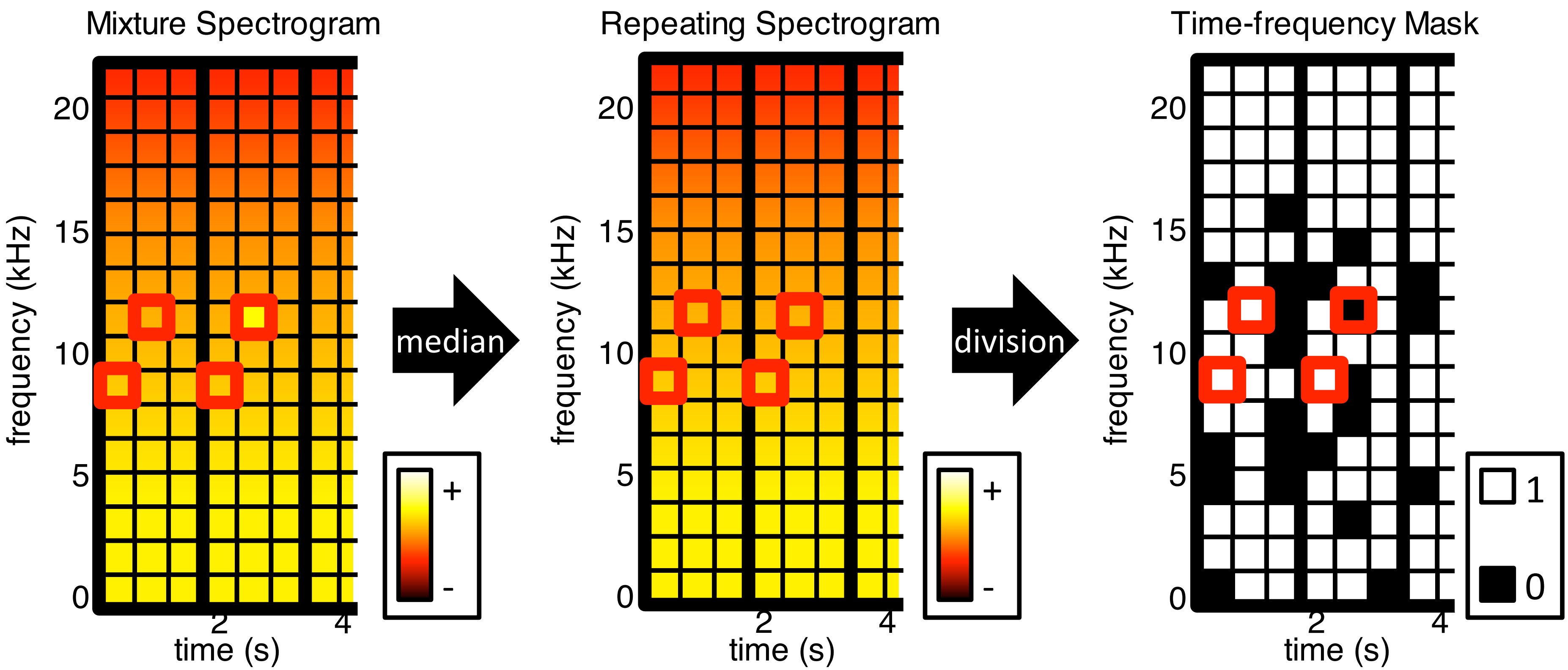
3. Repeating Structure

- In the soft t-f mask, the **less/more a t-f bin is repeating**, the more it is **weighted toward 0/1**



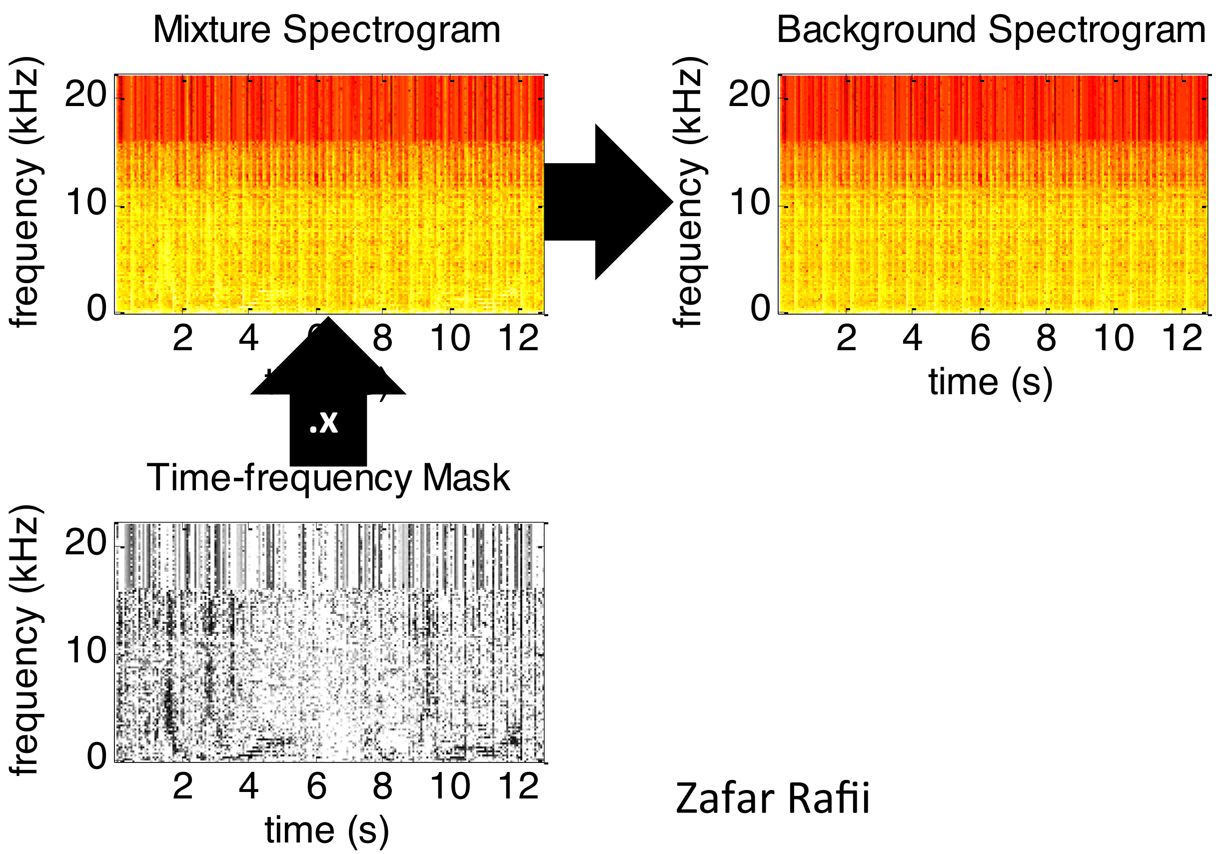
3. Repeating Structure

- A **binary t-f mask** can be further derived by choosing a threshold between 0 and 1



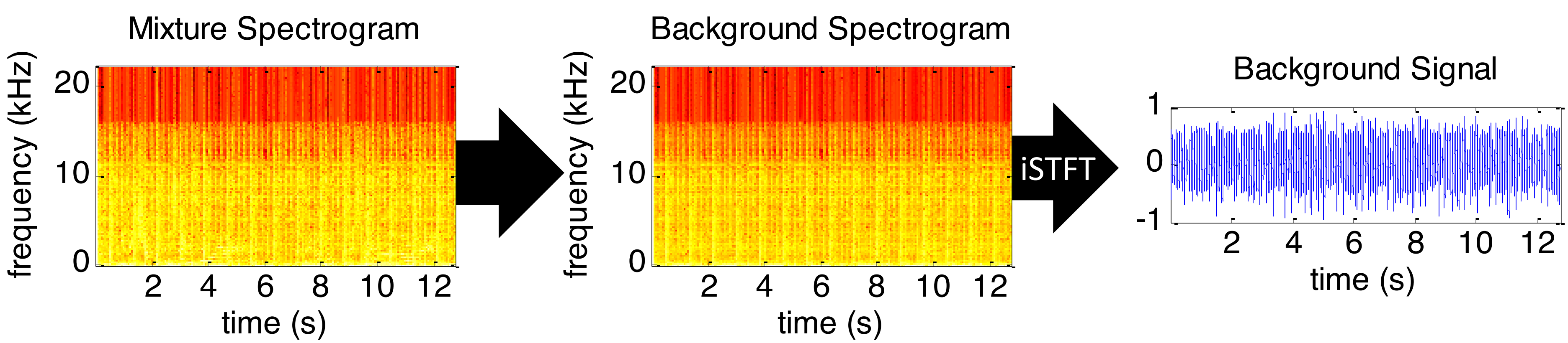
3. Repeating Structure

- We **multiplied** the t-f mask with the mixture STFT to extract the repeating background STFT



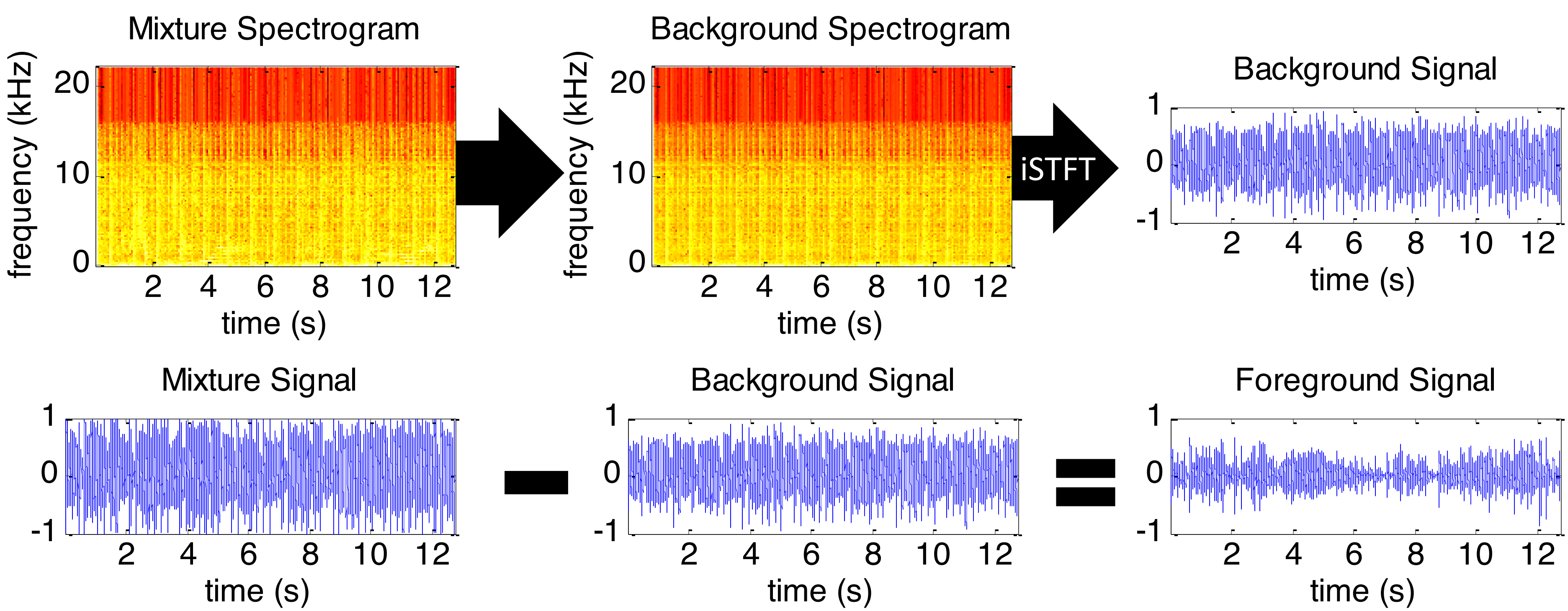
3. Repeating Structure

- The **repeating background** is obtained by inverting its STFT into the time domain



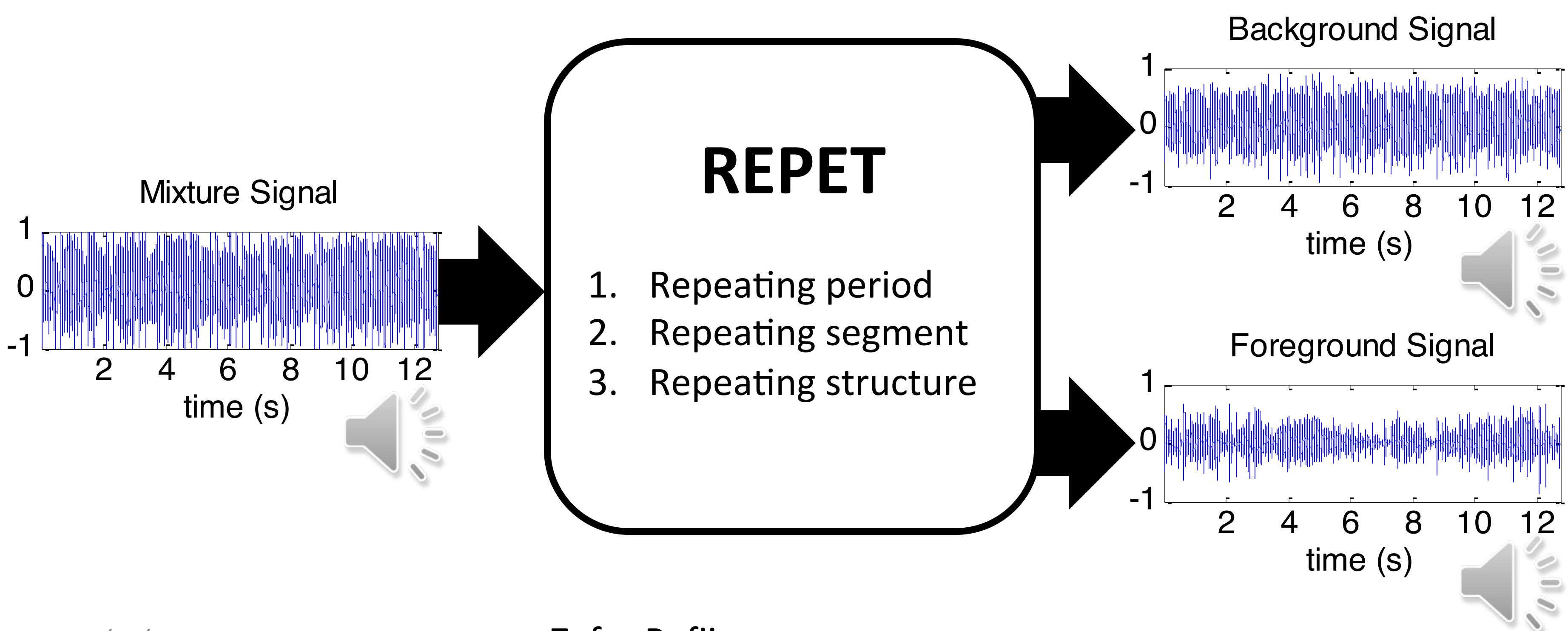
3. Repeating Structure

- The **non-repeating foreground** is obtained by subtracting the background from the mixture



Method

- Repeating background ≈ **music component**
- Non-repeating foreground ≈ **voice component**



Outline

I. Introduction

II. REPET

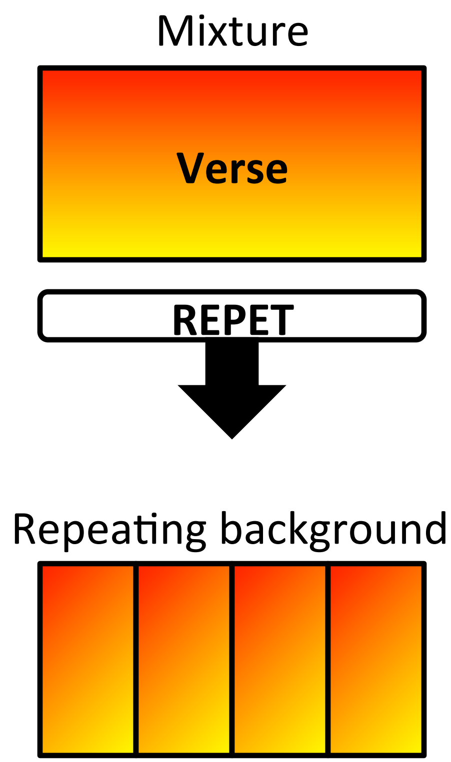
1. Method
2. **Extensions**
3. Evaluation

III. REPET-SIM

IV. Conclusion

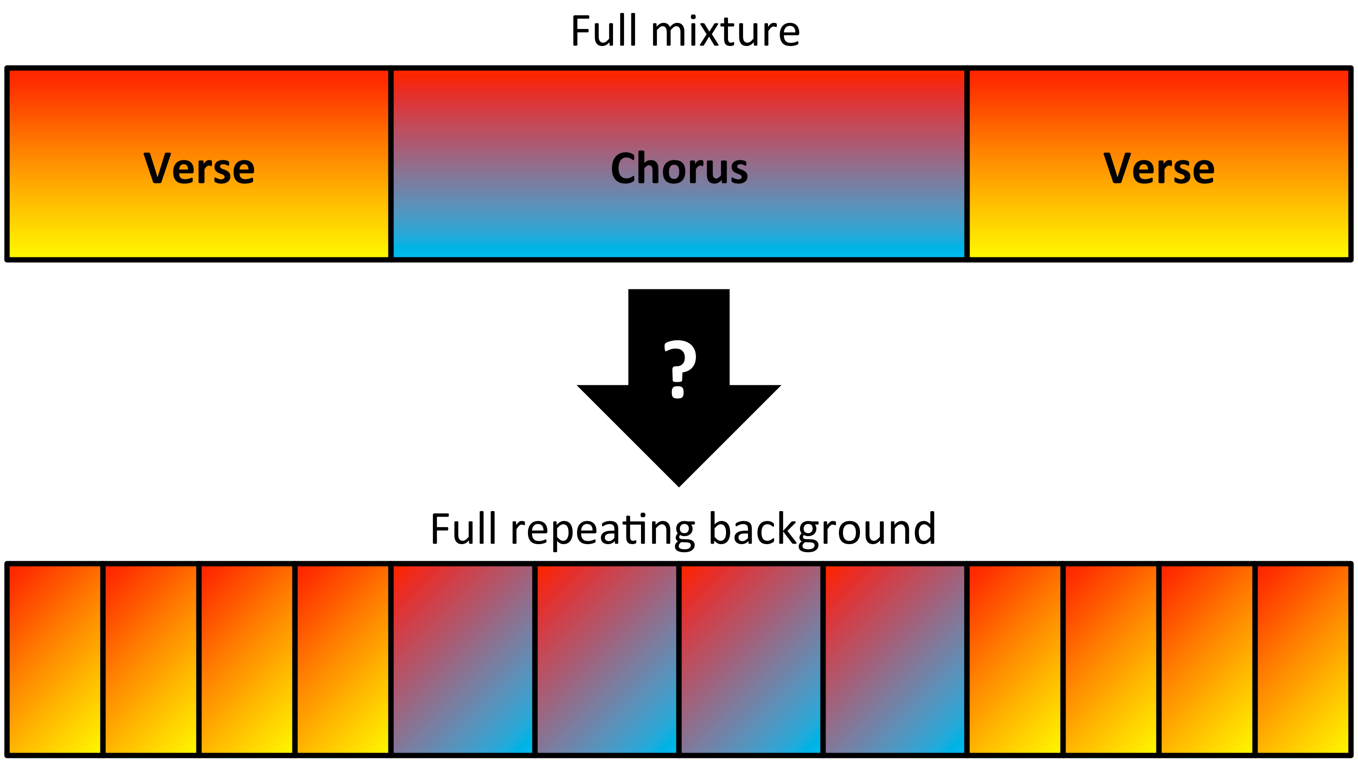
Extensions

- REPET works well on excerpts with a relatively **stable repeating background** (e.g., 10 s verse)



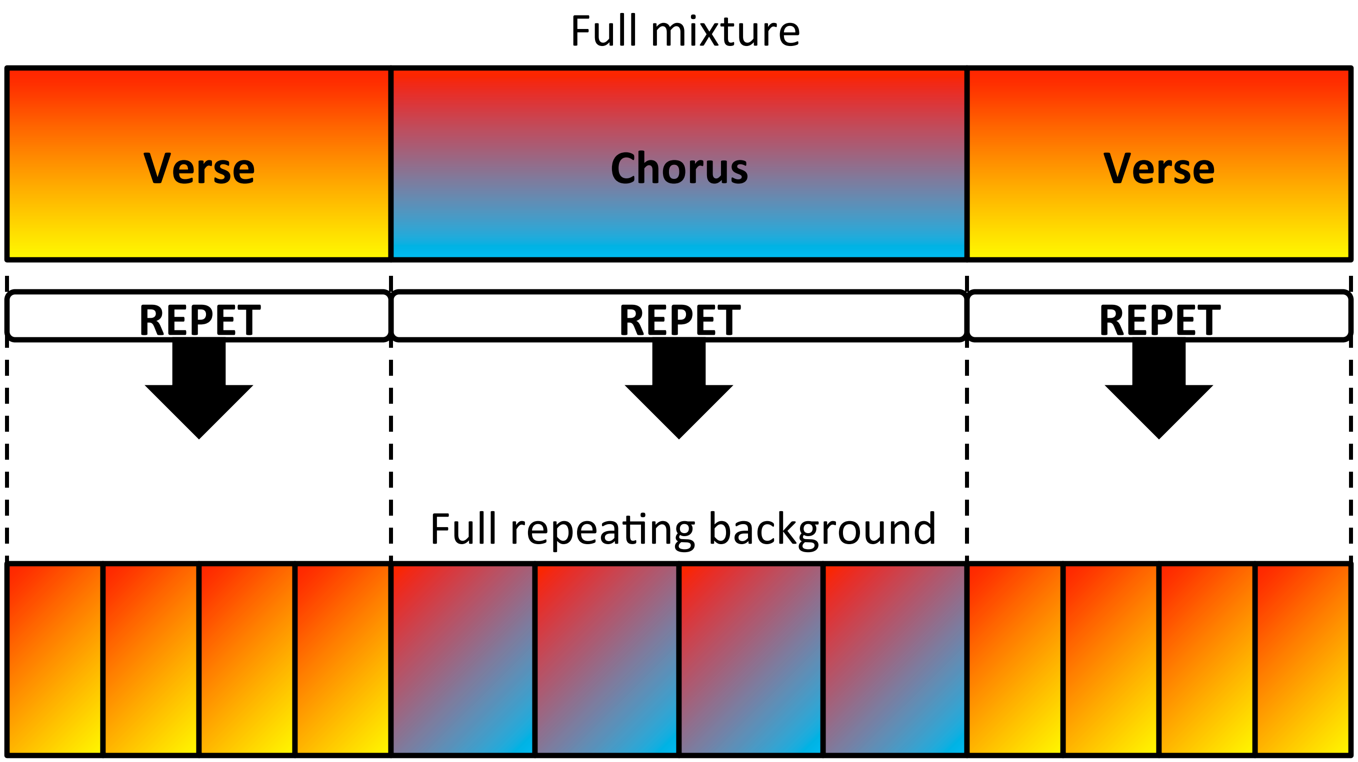
Extensions

- For full-track songs, the repeating background is likely to **vary over time** (e.g., verse/chorus)



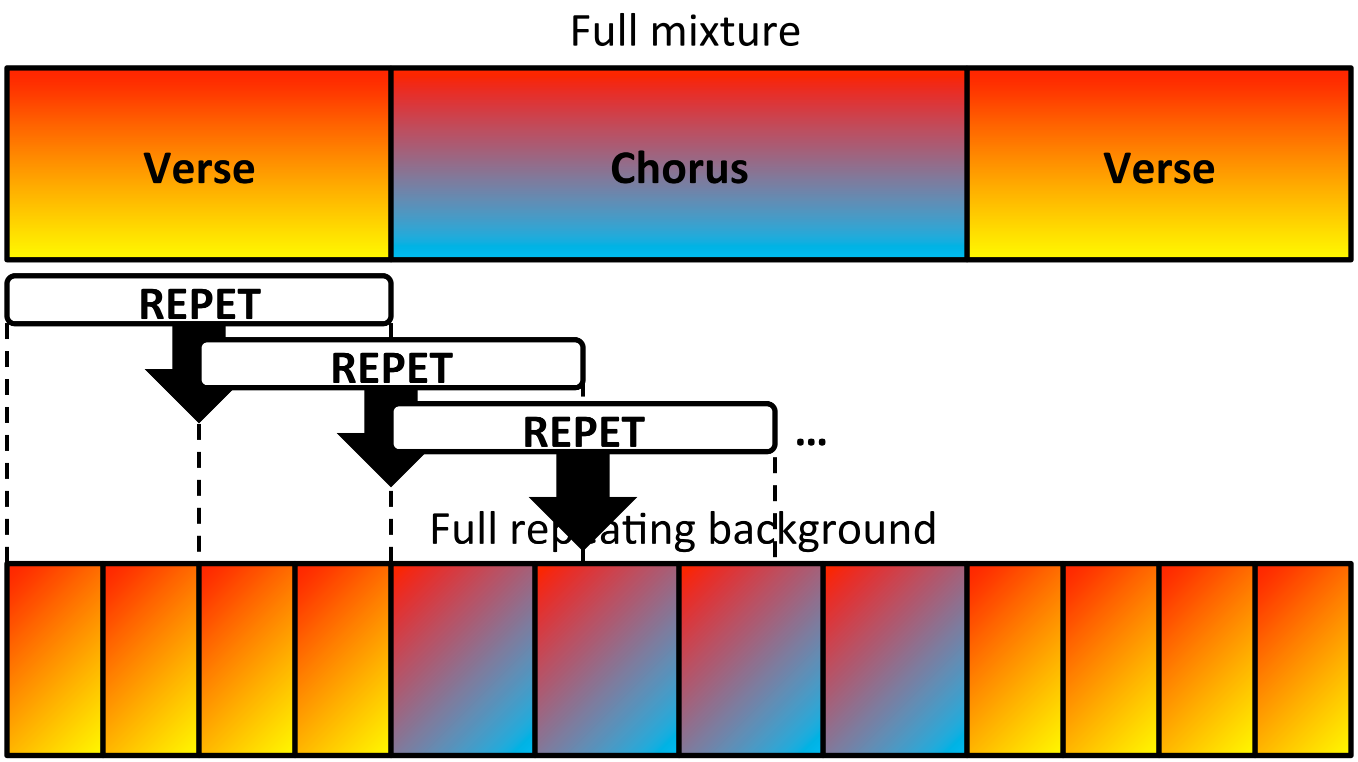
1. Prior Segmentation

- We could do a **prior segmentation** of the song and apply REPET to the individual sections



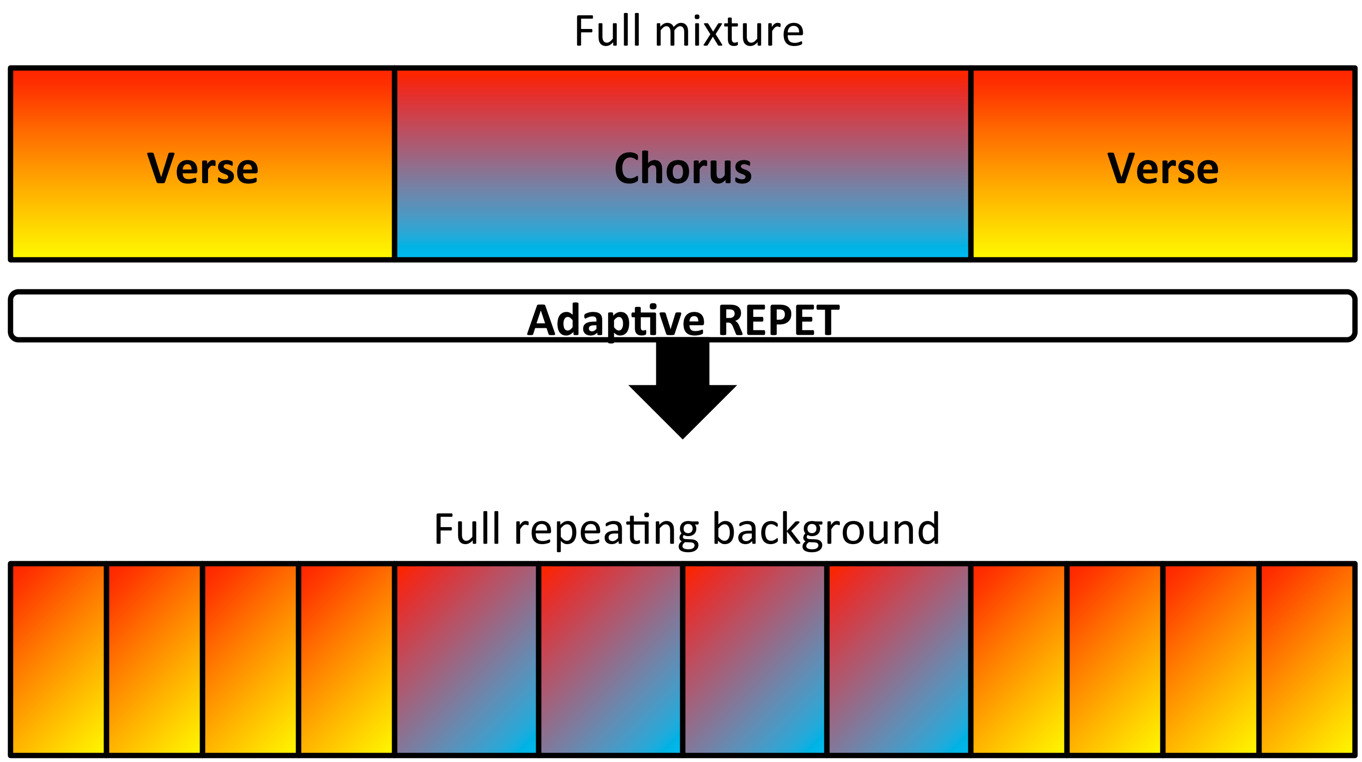
2. Sliding Window

- We could apply REPET to local sections of the song over time via a fixed **sliding window**

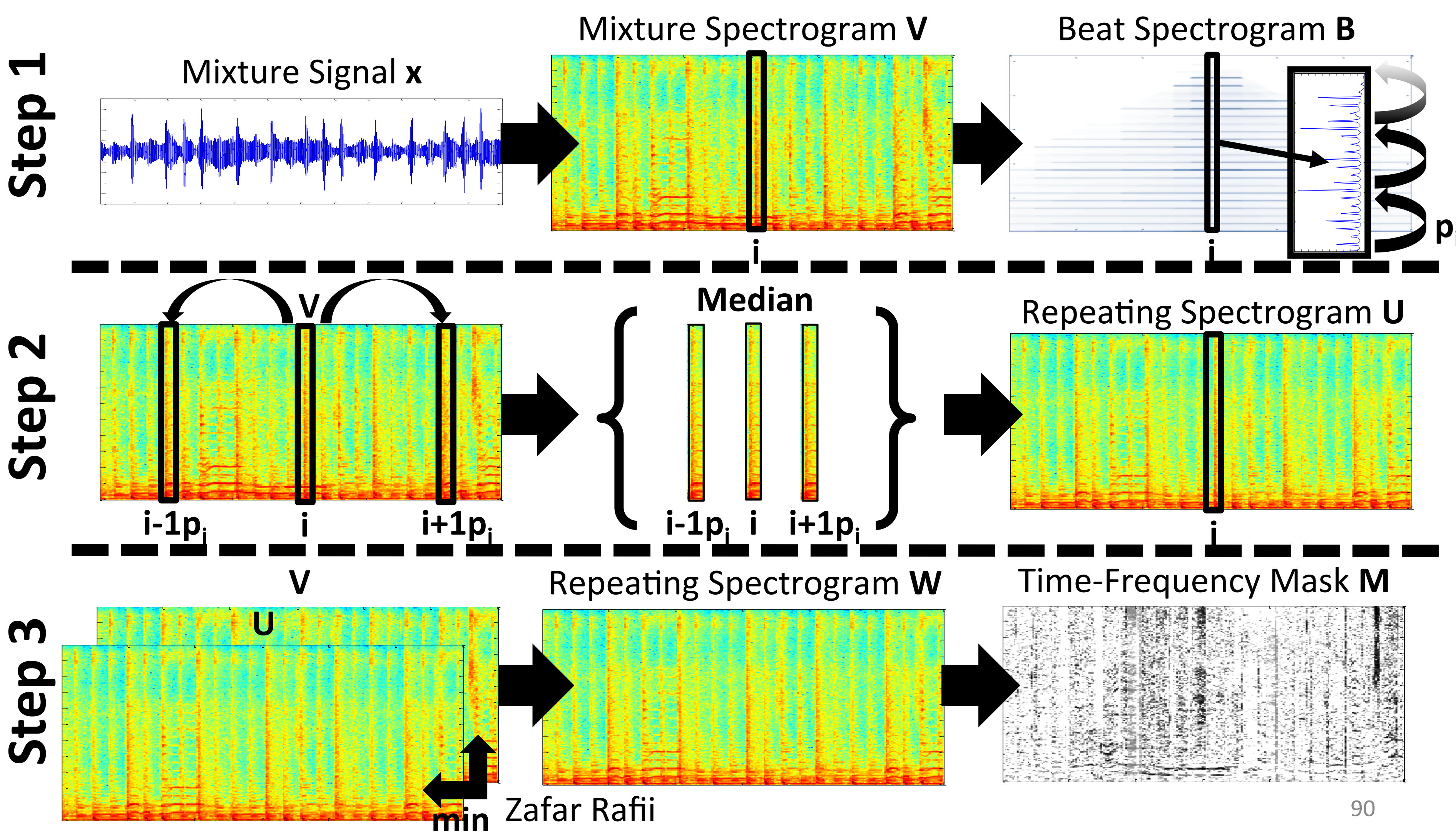


3. Adaptive REPET

- We could **adapt REPET** along time by locally modeling the repeating background

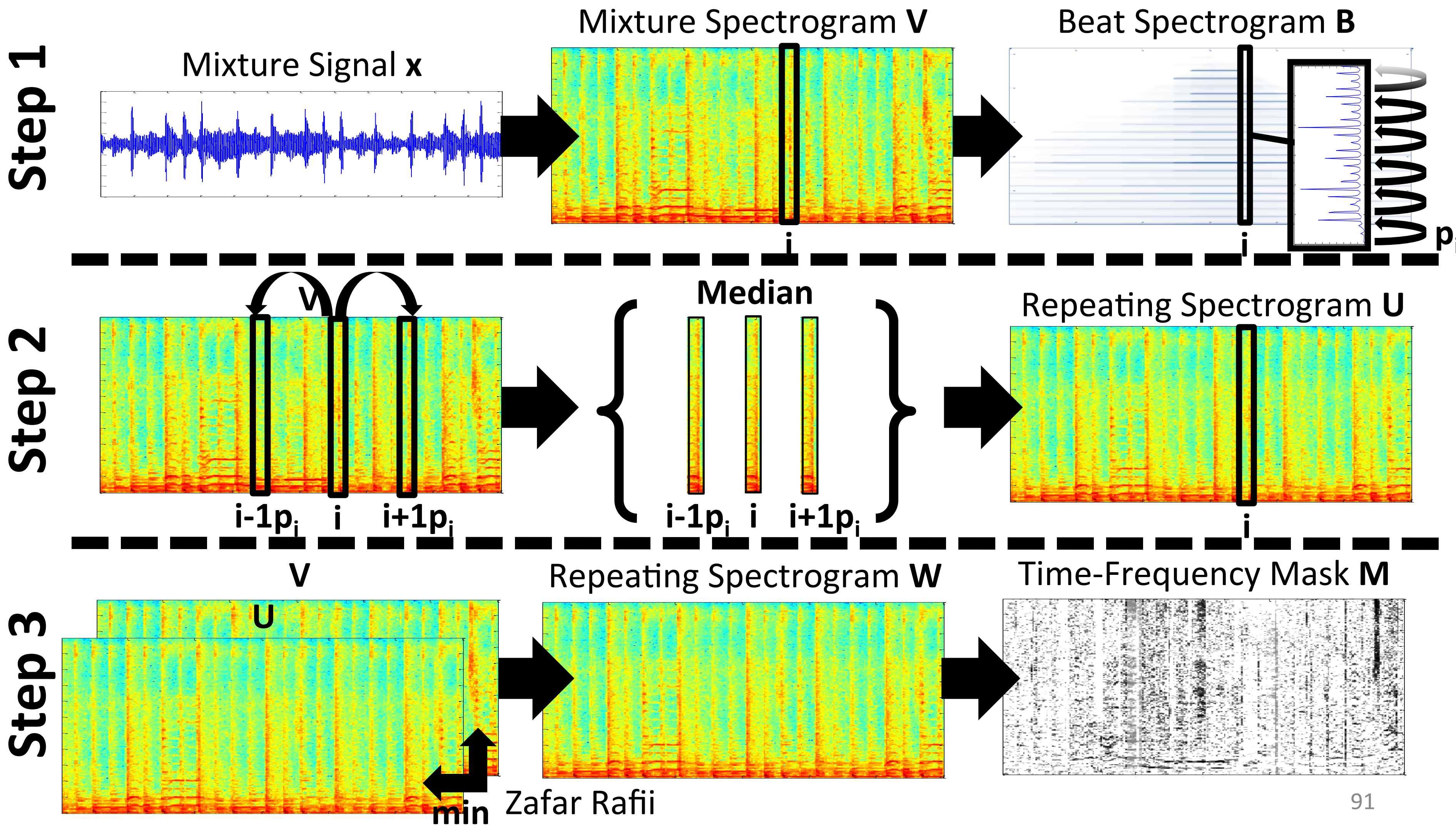


Adaptive REPET



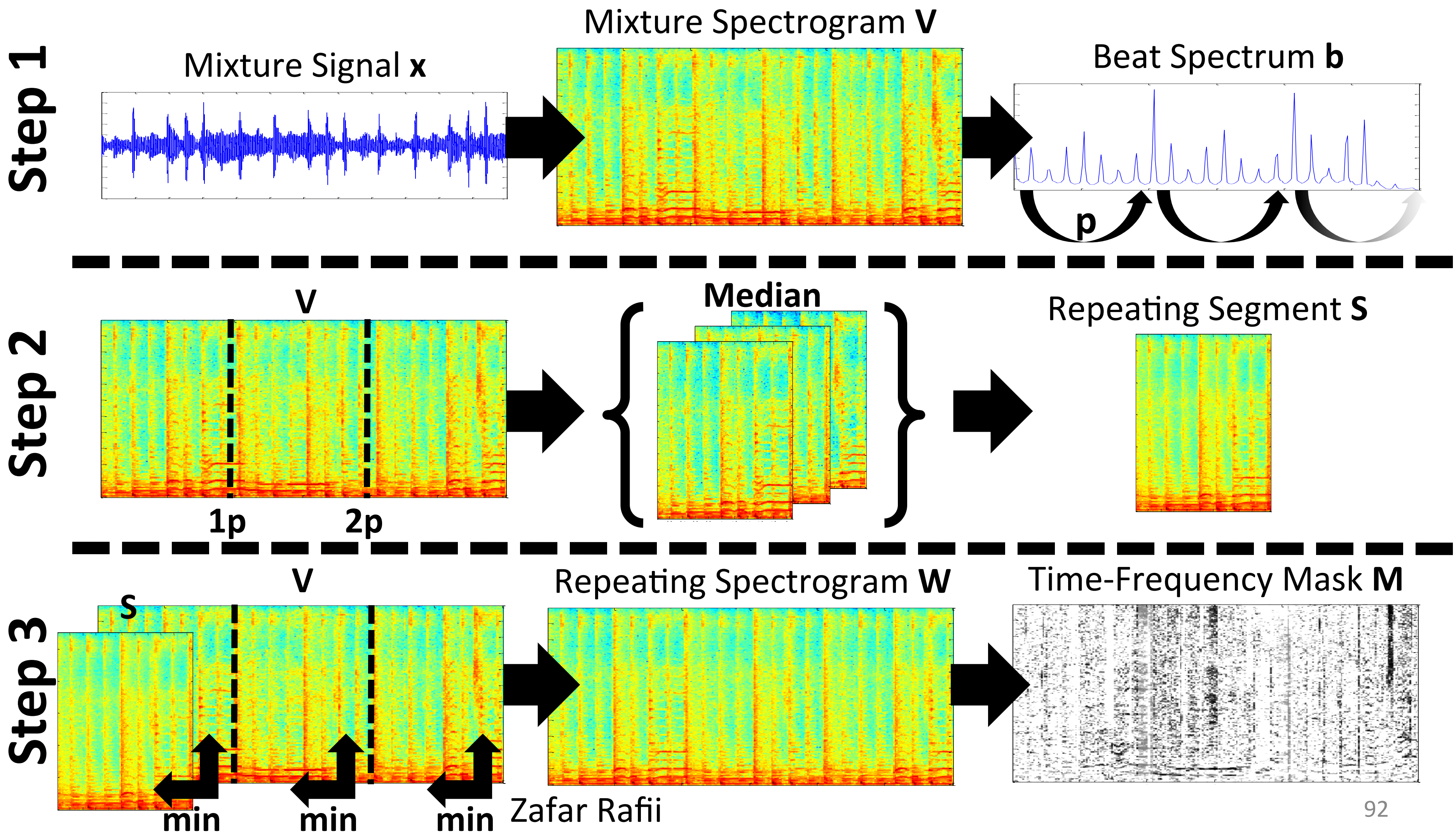
Zafar Rafii

Adaptive REPET



Zafar Rafii

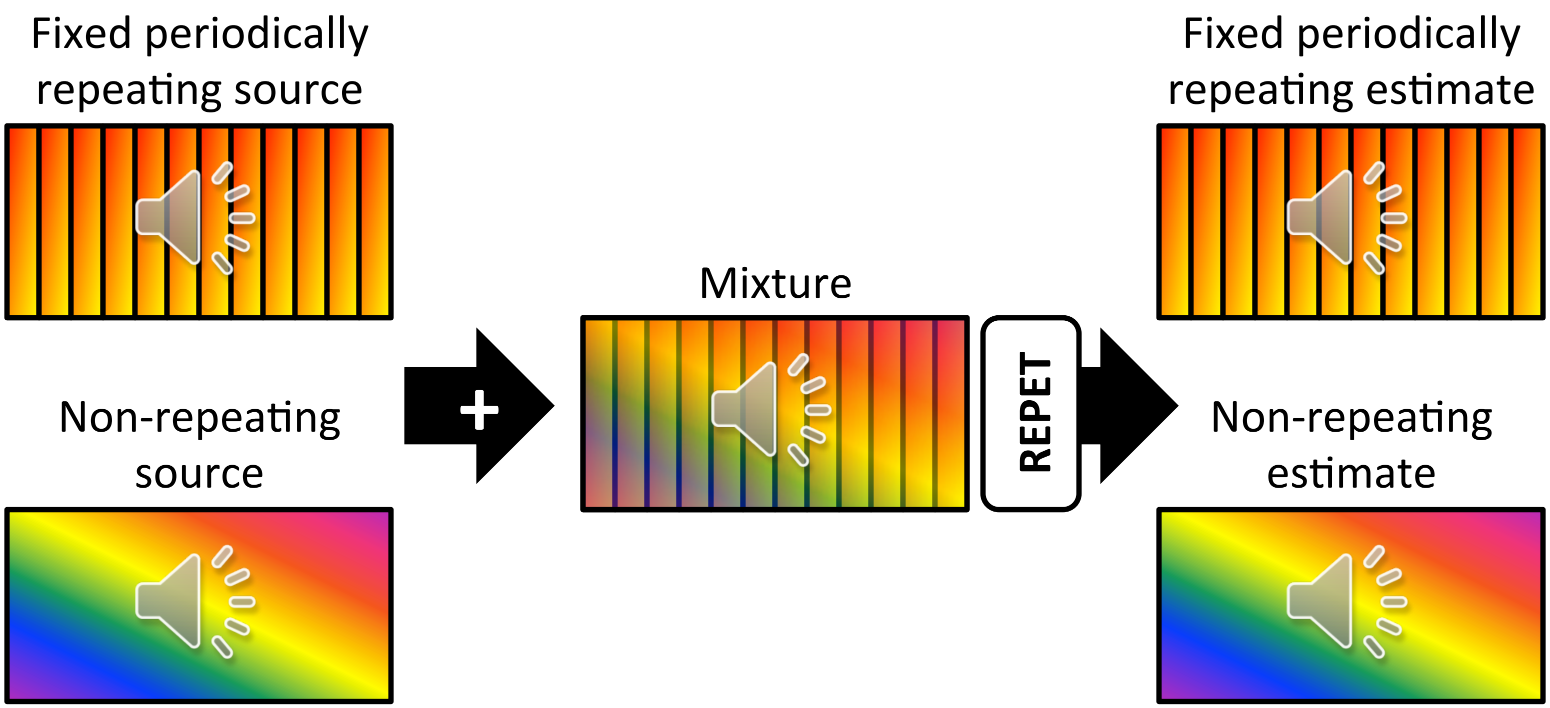
Original REPET



Zafar Rafii

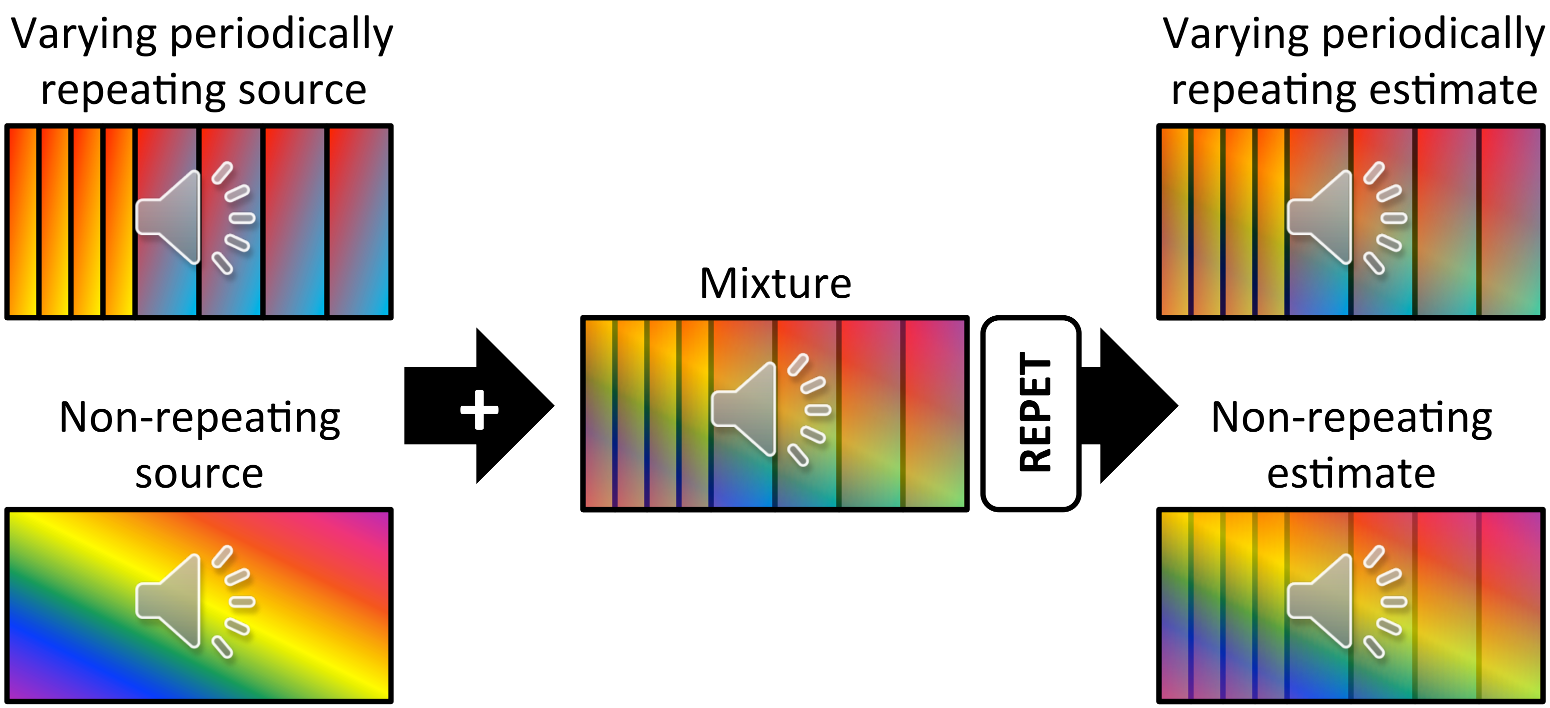
Adaptive vs. original REPET

- REPET assumes a stable repeating background with repetitions occurring at **fixed period rate**



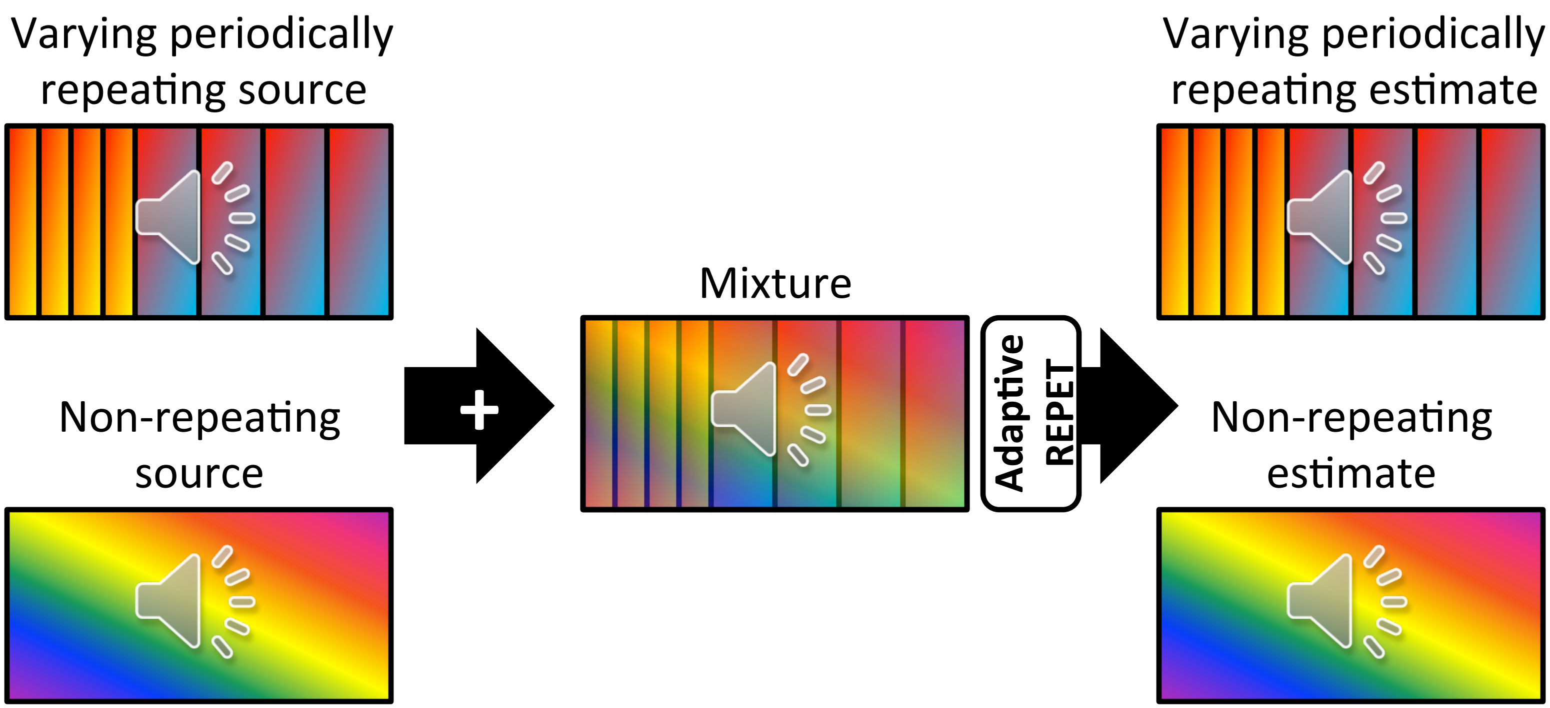
Adaptive vs. original REPET

- The original REPET shows limitations when the repeating background **varies over time**



Adaptive vs. original REPET

- The adaptive REPET can handle **varying repeating structures** (e.g., in full-track songs)



Outline

I. Introduction

II. REPET

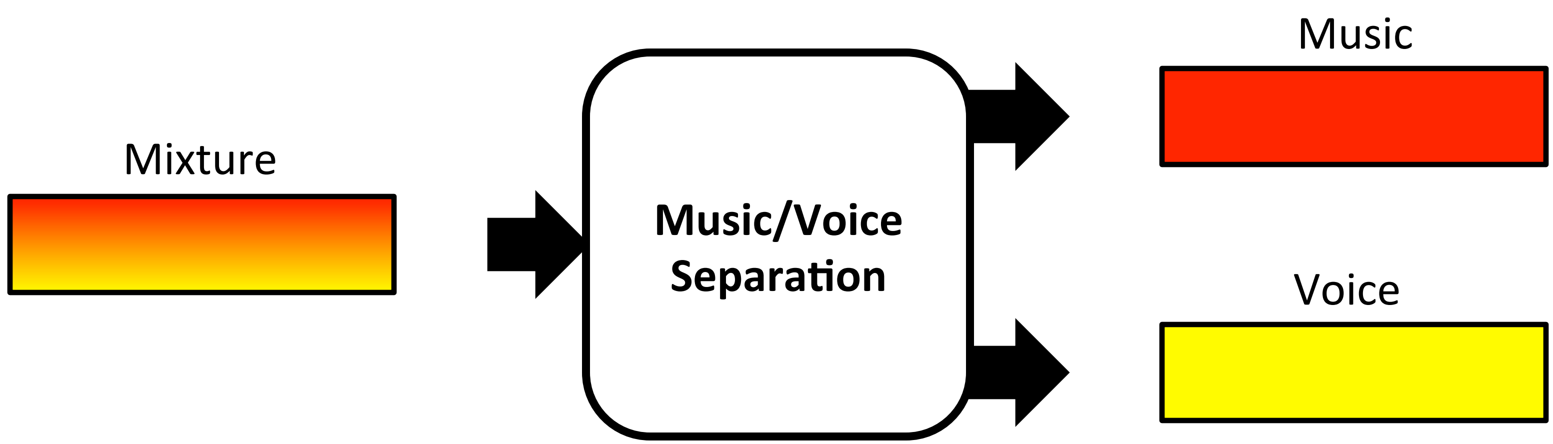
1. Method
2. Extensions
3. **Evaluation**

III. REPET-SIM

IV. Conclusion

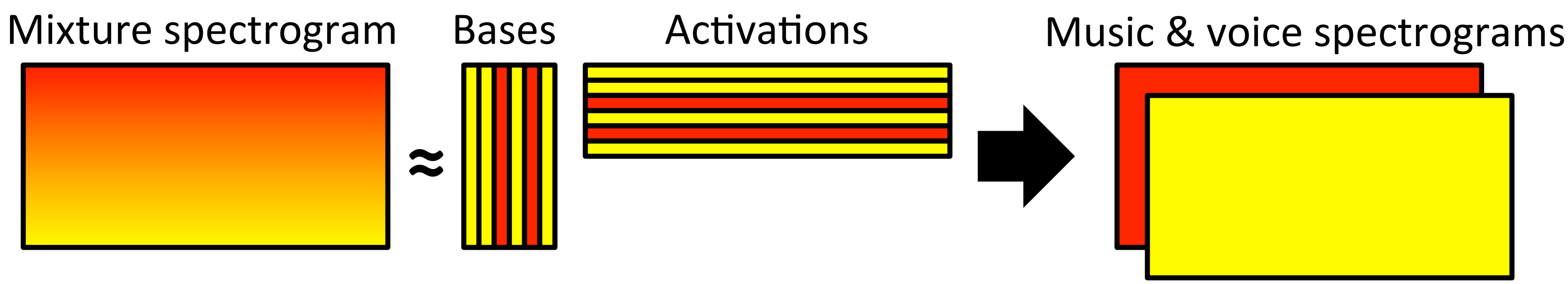
Music/Voice Separation

- **Music/voice separation** systems generally first identify the vocal/non-vocal segments and then use a variety of techniques to separate the music and voice components



Music/Voice Separation

- **Non-negative Matrix Factorization (NMF)**
  - Iterative factorization of the mixture spectrogram into non-negative additive components

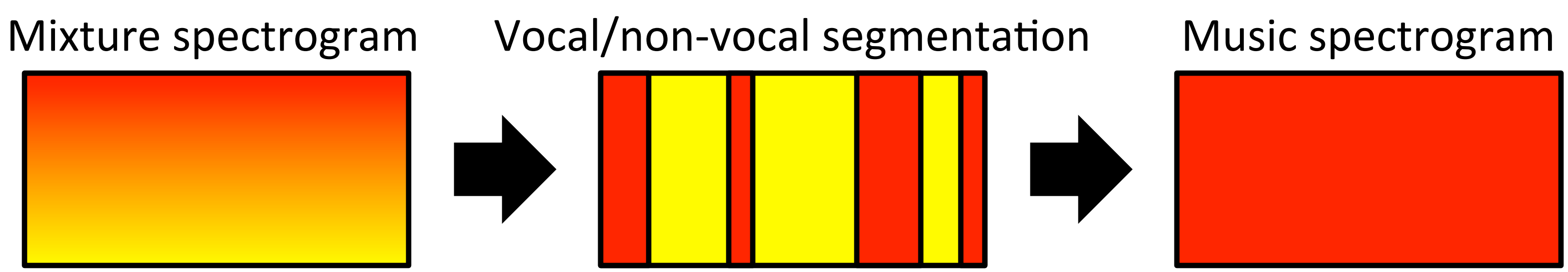


→ Need to know the number of components

→ Need a proper initialization

Music/Voice Separation

- **Accompaniment modeling**
  – Modeling of the musical accompaniment from the non-vocal segments in the mixture

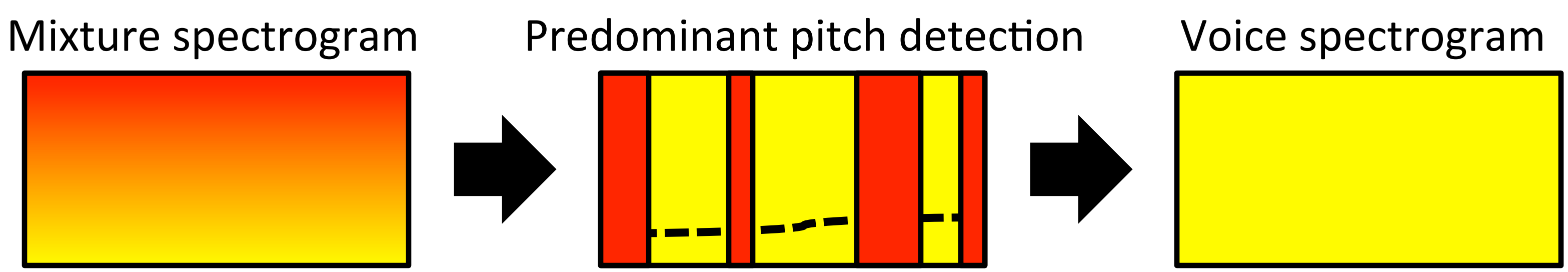


→ Need an accurate vocal/non-vocal segmentation

→ Need a sufficient amount of non-vocal segments

Music/Voice Separation

- **Pitch-based inference**
  - Separation of the vocals using the predominant pitch contour extracted from the vocal segments



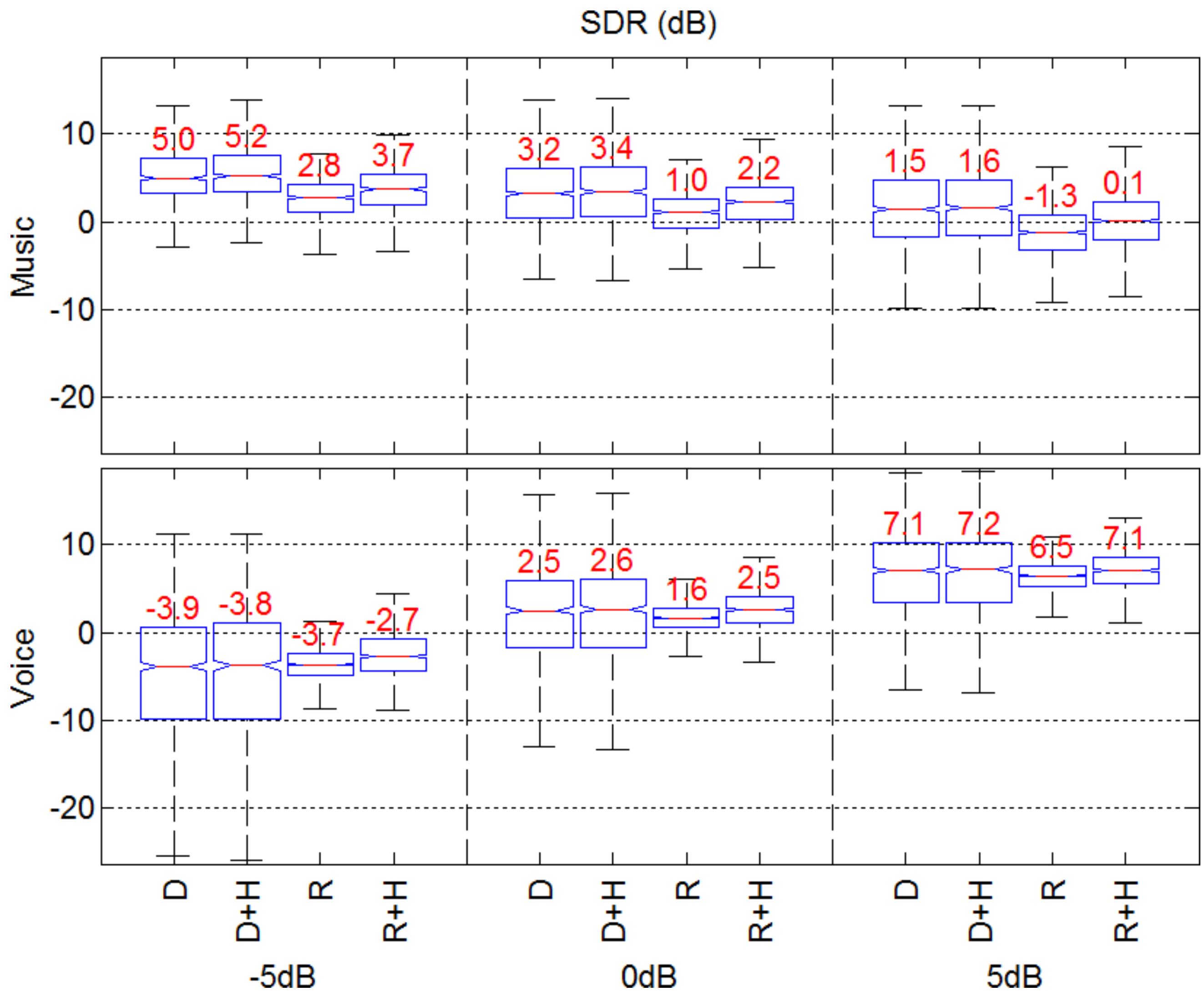
→ Need an accurate predominant pitch detection

→ Cannot extract unvoiced vocals

Evaluation

- **REPET** [Rafii et al., 2012]
  - Automatic period finder
  - Soft time-frequency masking
- **Competitive method** [Durrieu et al., 2011]
  - Source/filter modeling with NMF framework
  - Unvoiced vocals estimation
- **Data set** [Hsu et al., 2010]
  - 1,000 song clips (from karaoke Chinese pop songs)
  - 3 voice-to-music mixing ratios (-5, 0, and 5 dB)

Evaluation



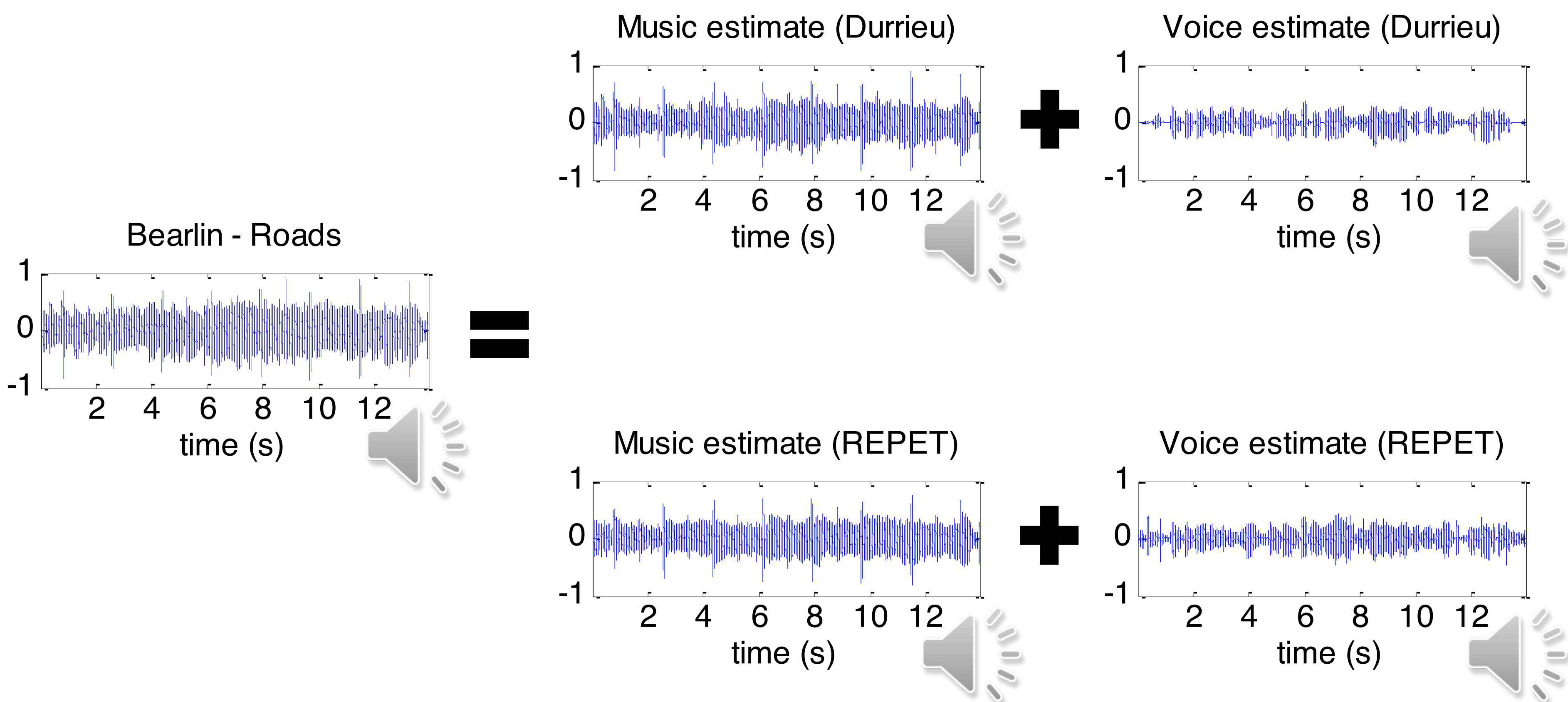
D = Durrieu et al.
D+H = Durrieu + High-pass
R = REPET
R+H = REPET + High-pass

Evaluation

- **Conclusions**
  - REPET can compete with state-of-the-art (and more complex) music/voice separation methods
  - There is room for improvement (+ high-pass, + optimal period, + vocal frames)
  - Average computation time: 0.016 second for 1 second of mixture! (vs. 3.863 seconds for Durrieu)

Example

- REPET vs. Durrieu et al.



Outline

I. Introduction

II. REPET

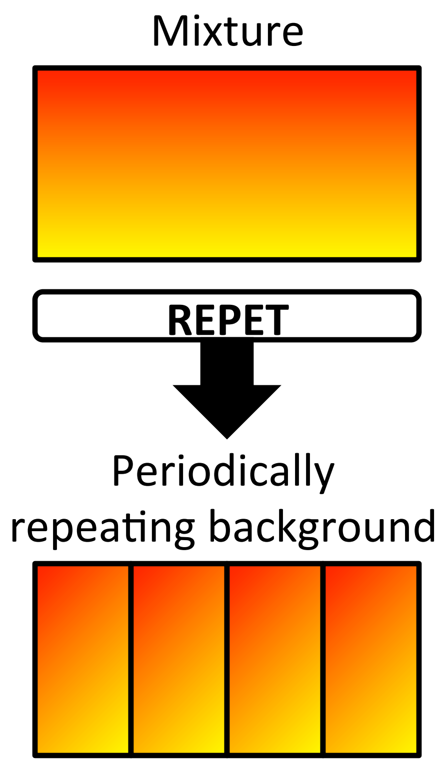
III. REPET-SIM

1. **Similarity**
2. Method
3. Evaluation

IV. Conclusion

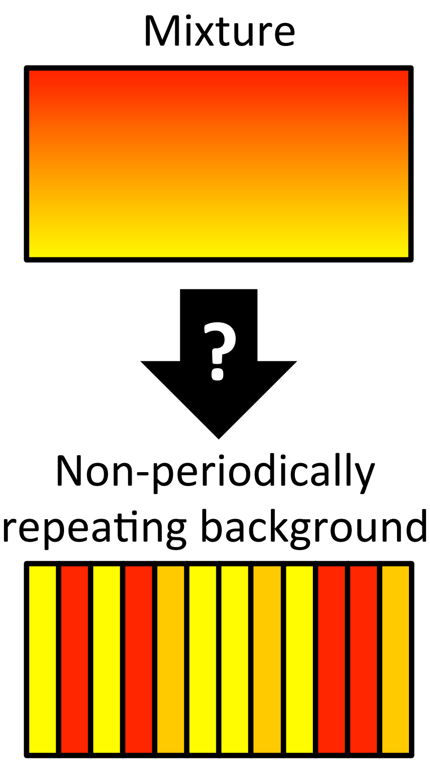
Similarity

- REPET (and its extensions) assume **periodically repeating patterns**



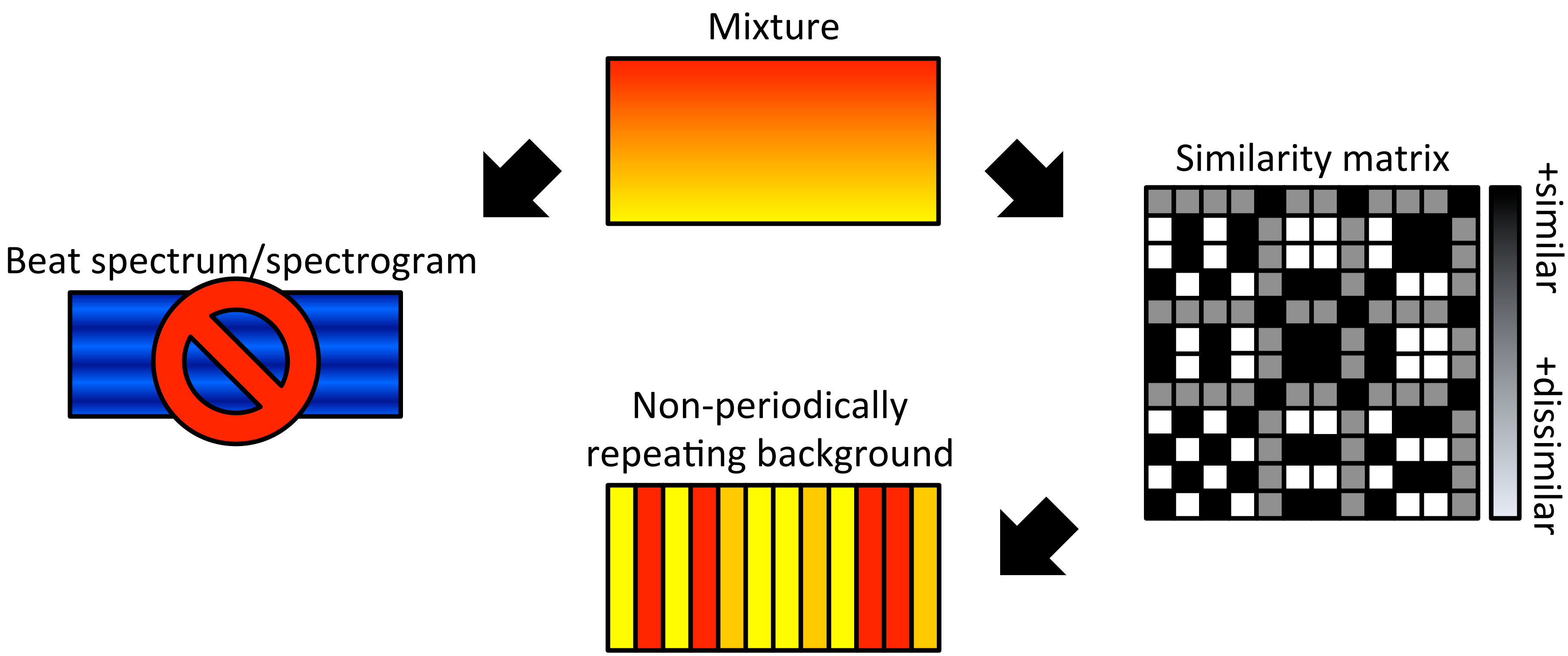
Similarity

- Repetitions can also happen **intermittently** or **without a global (or local) period**



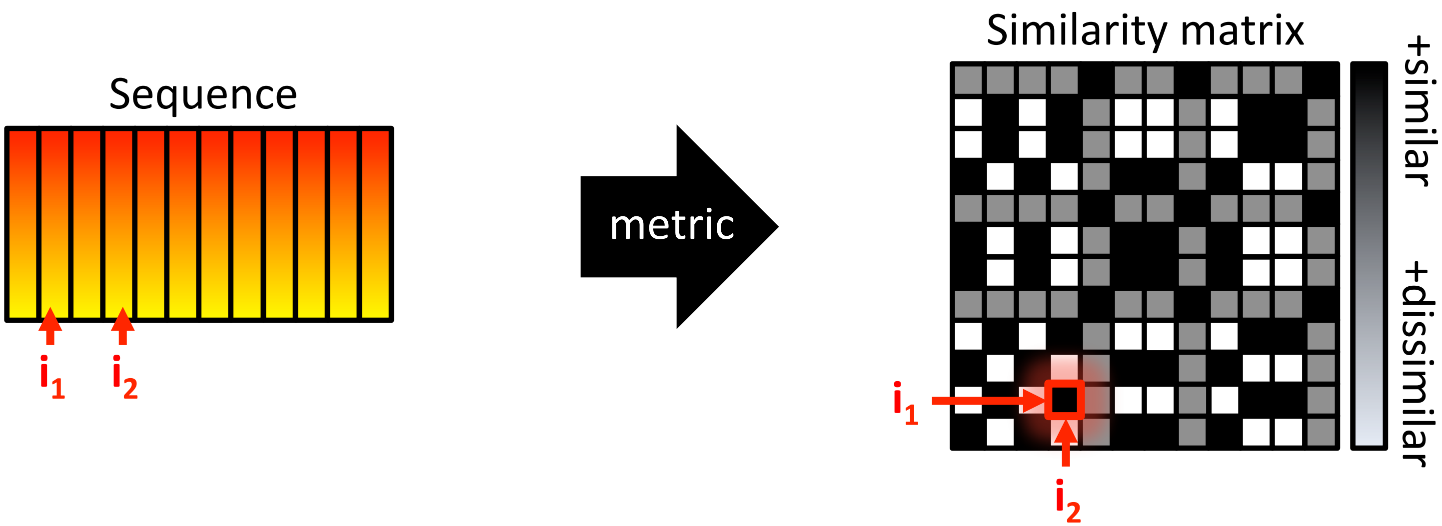
Similarity

- Instead of looking for periodicities, we can look for **similarities**, using a similarity matrix



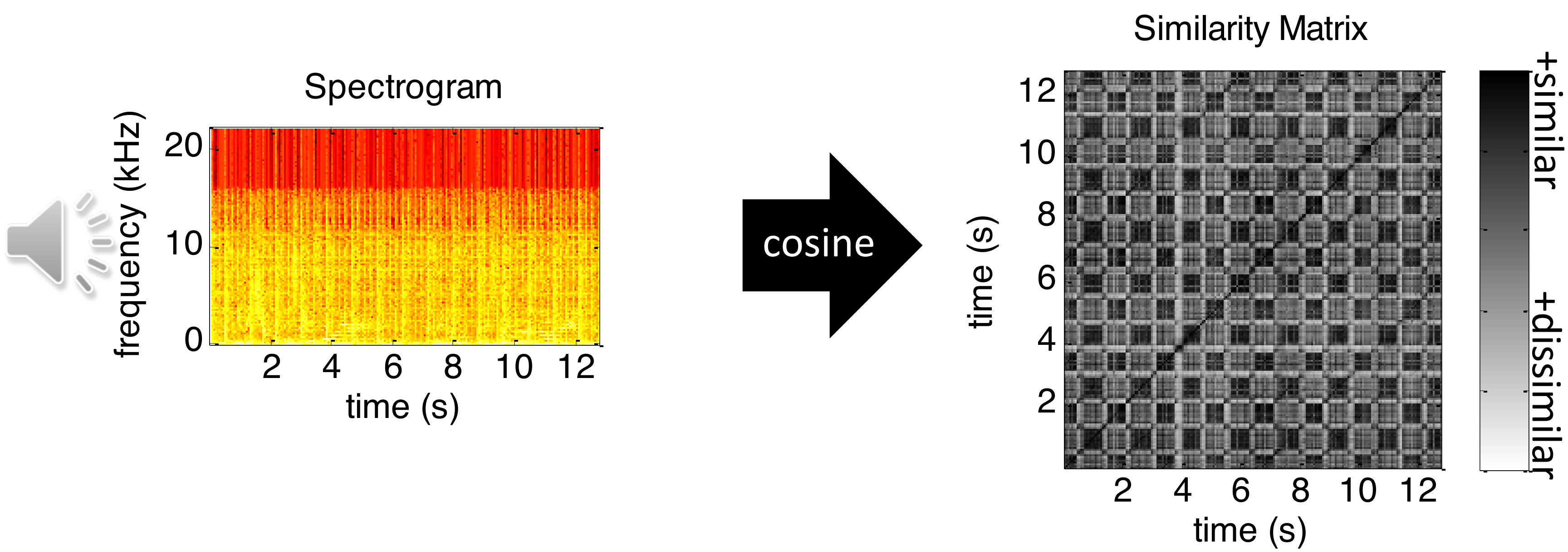
Similarity

- The **similarity matrix** is a matrix where each bin measures the (dis)similarity between any two elements of a sequence given a metric



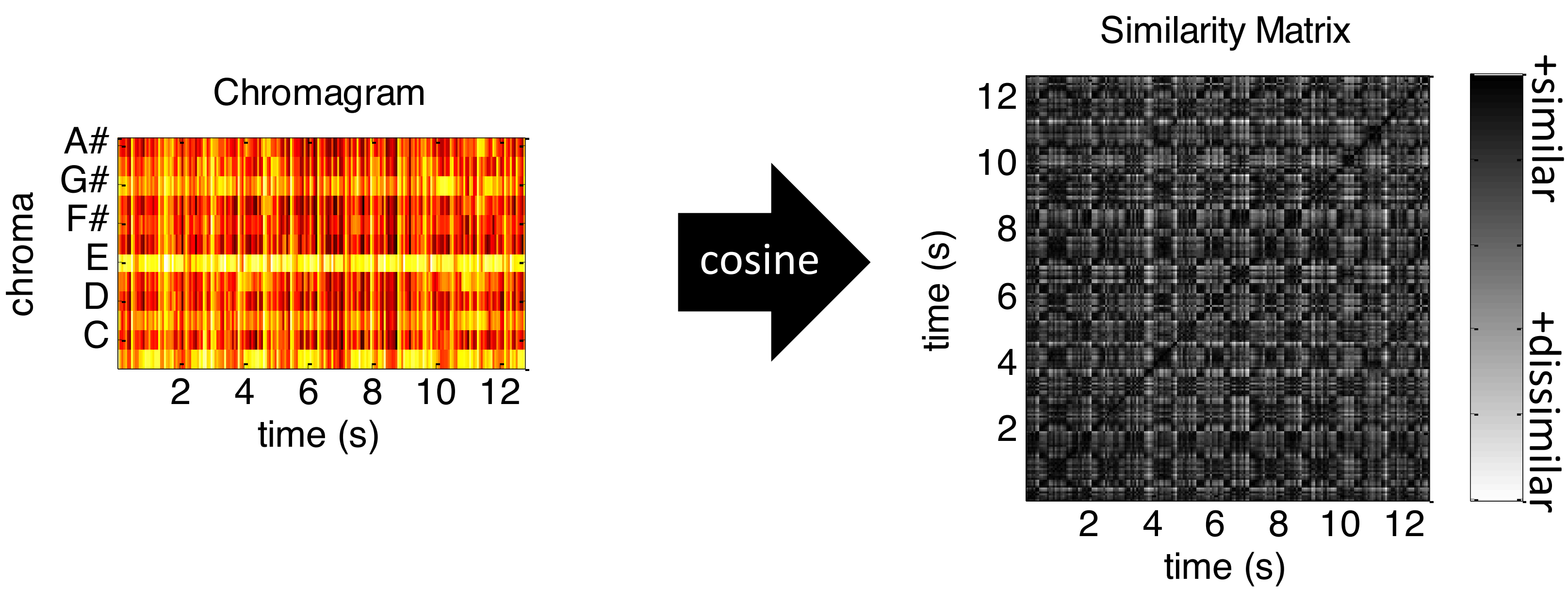
Similarity

- In audio, the SM can help to visualize the time structure and find **repeating/similar patterns**



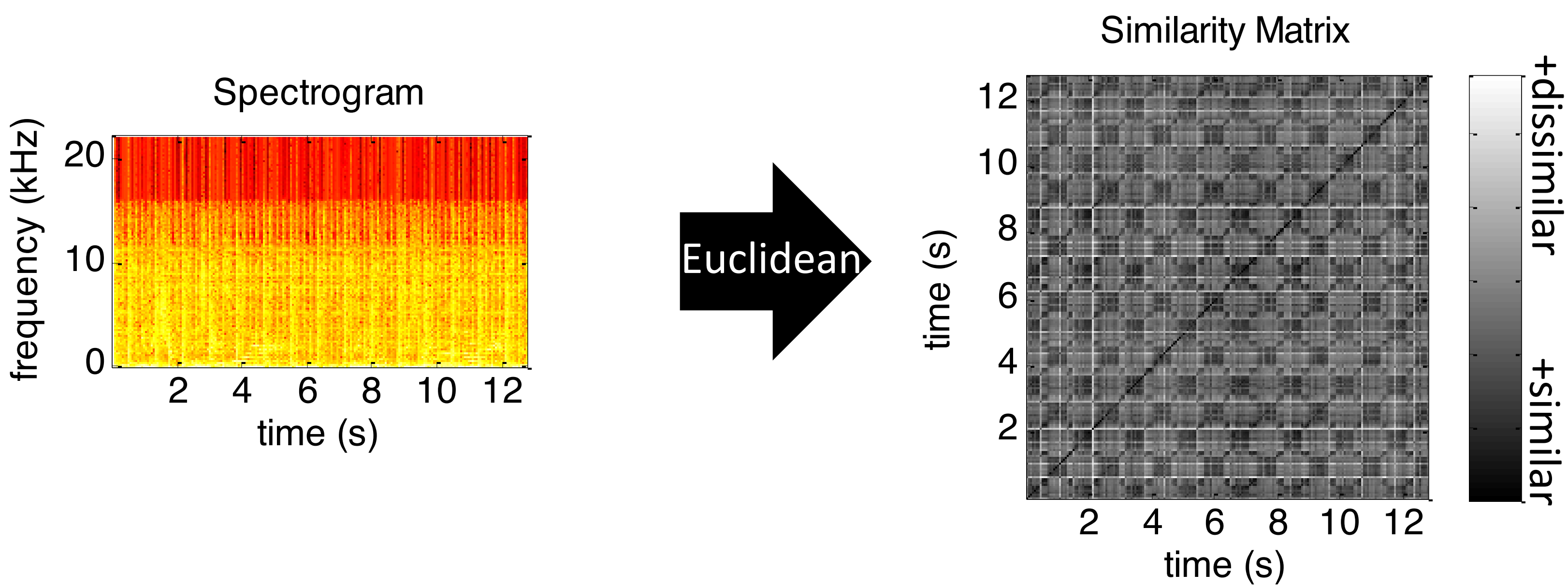
Similarity

- The SM can be built from **different features**: spectrogram, chromagram, etc.



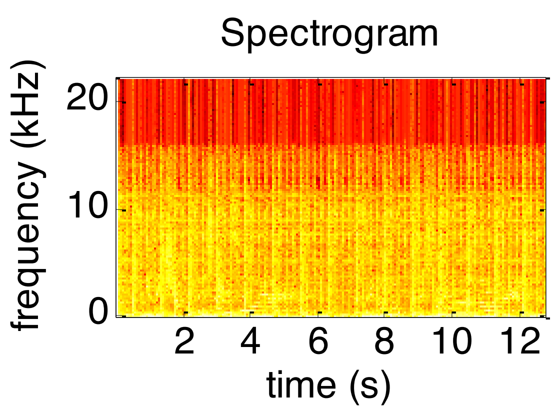
Similarity

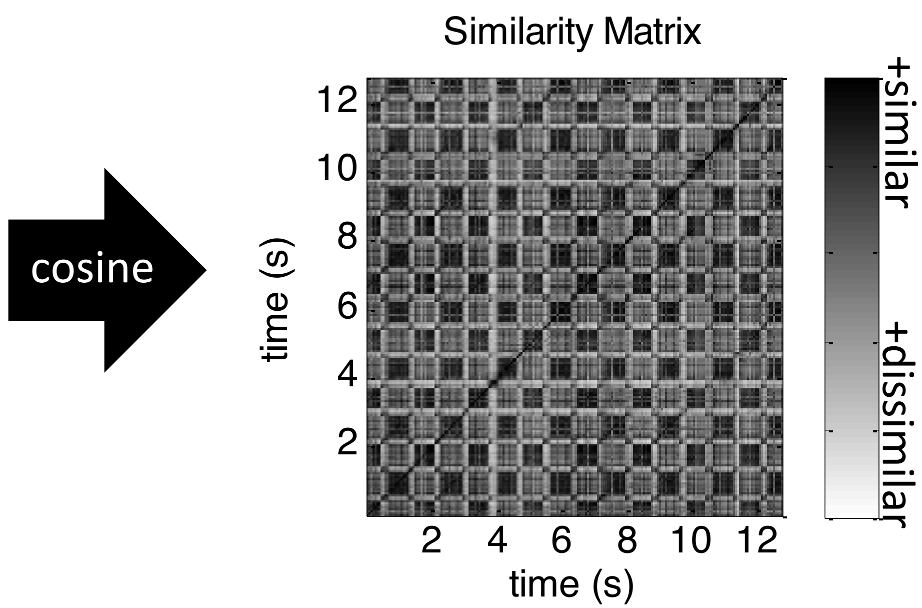
- The SM can be built using **different metrics**: cosine similarity, Euclidean distance, etc.



Similarity

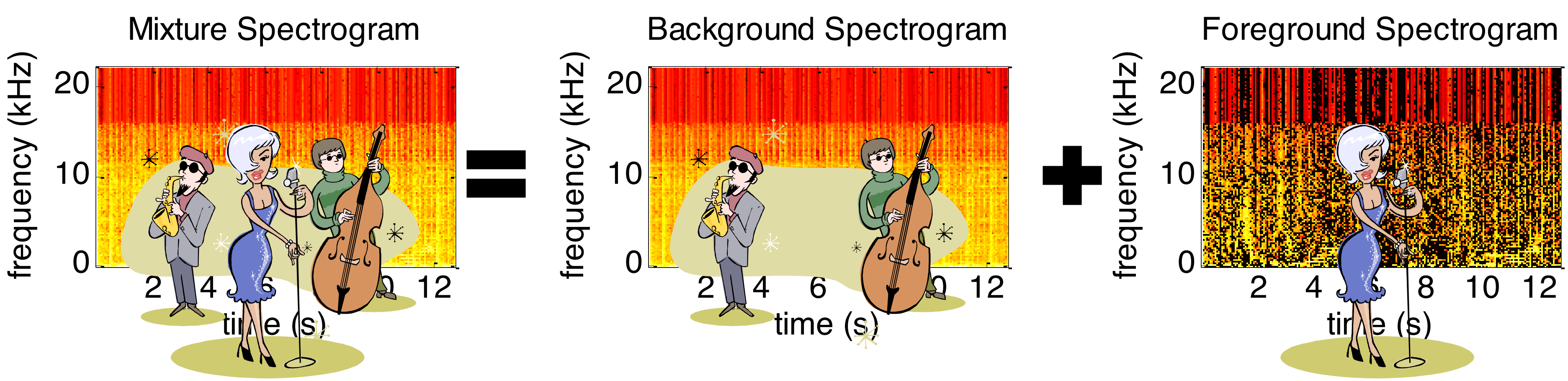
- We choose to simply build the SM from the **spectrogram** using the **cosine similarity**





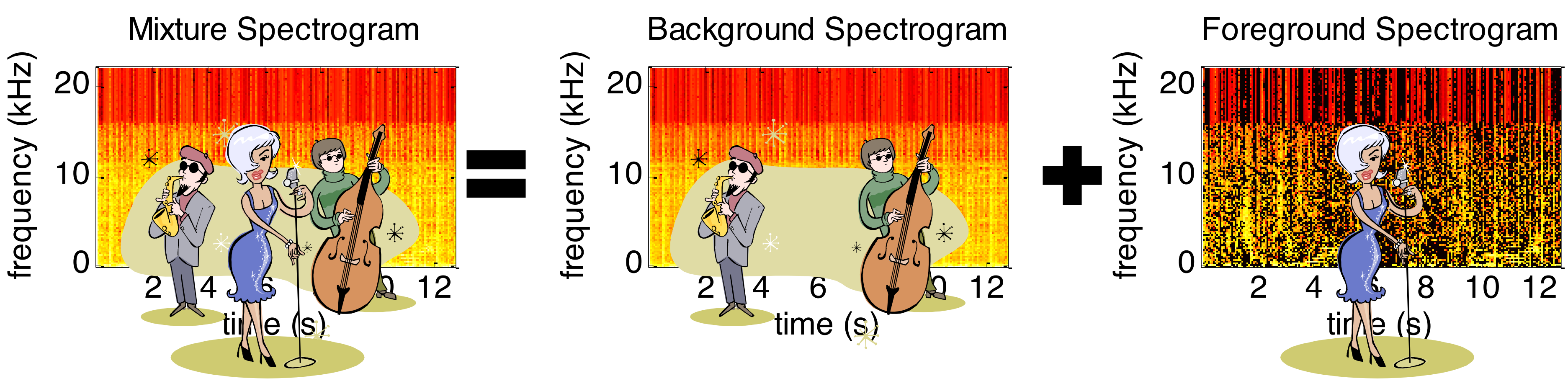
Similarity

- Given a mixture, we (again) assume that:
  - The repeating background is **dense & low-ranked**
  - The non-repeating foreground is **sparse & varied**



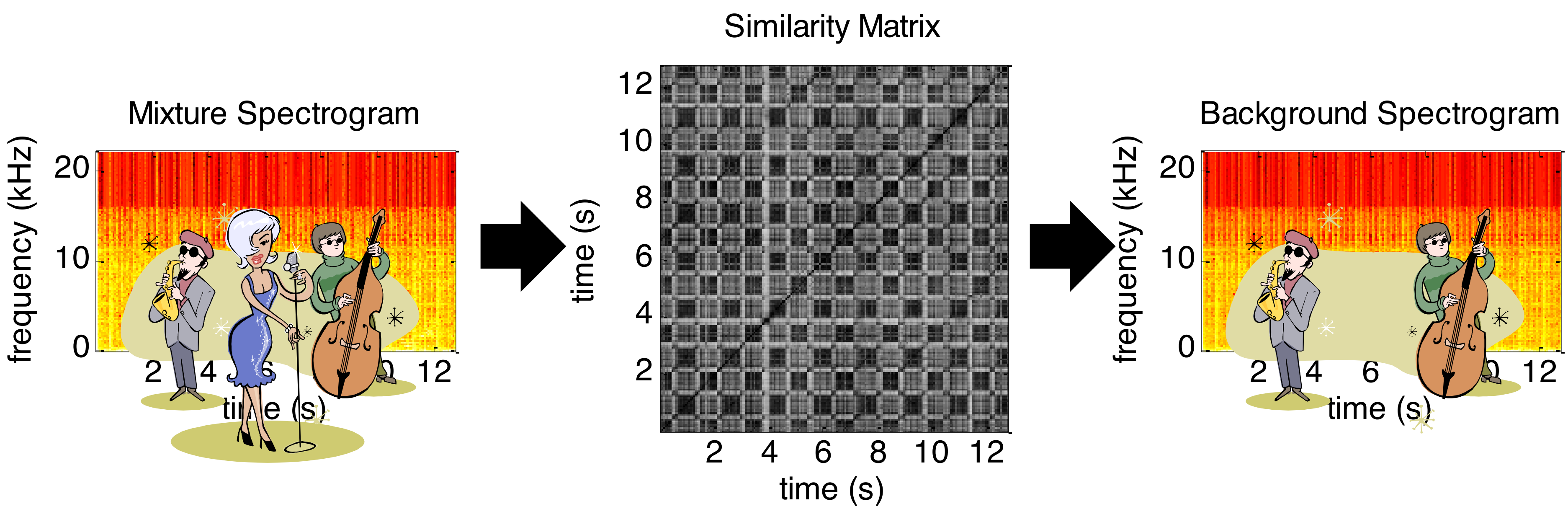
Similarity

- By low-ranked, we now mean the background is repeating, but **not necessarily periodically**



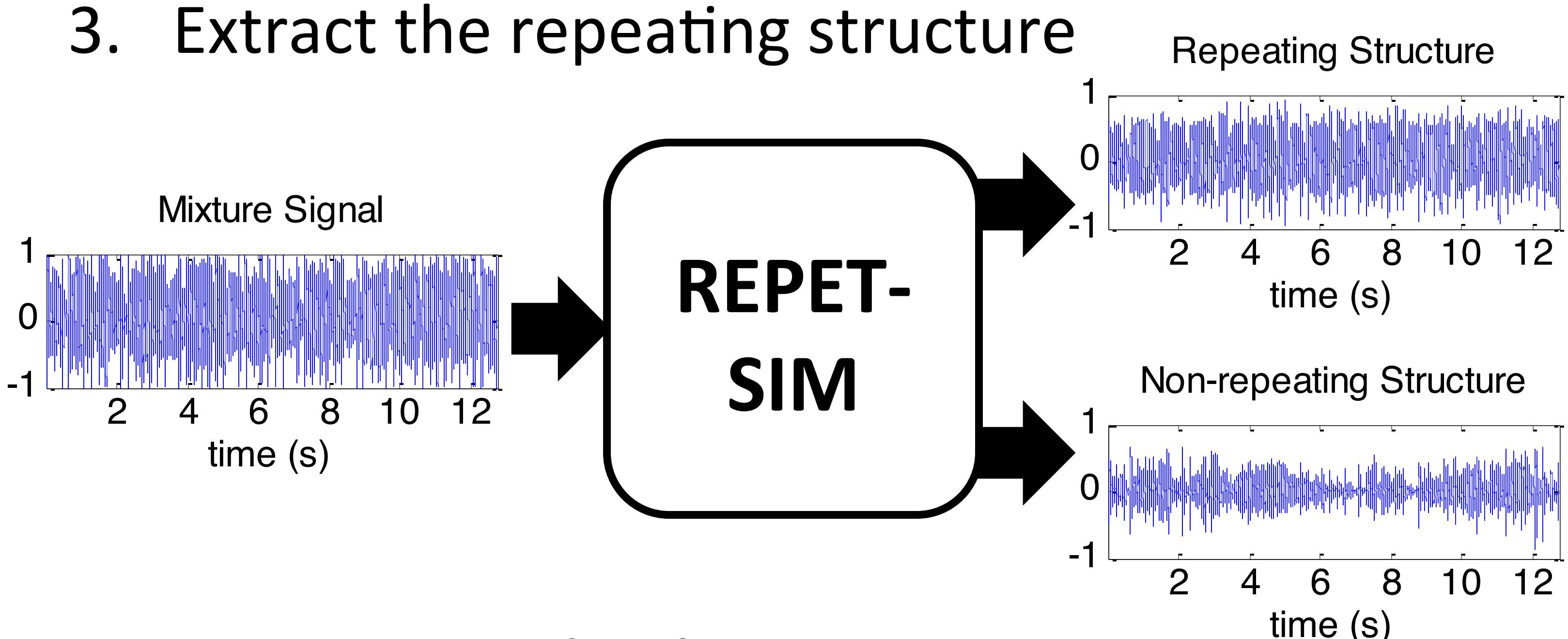
Similarity

- The SM of a mixture is then likely to reveal the structure of the **repeating background**



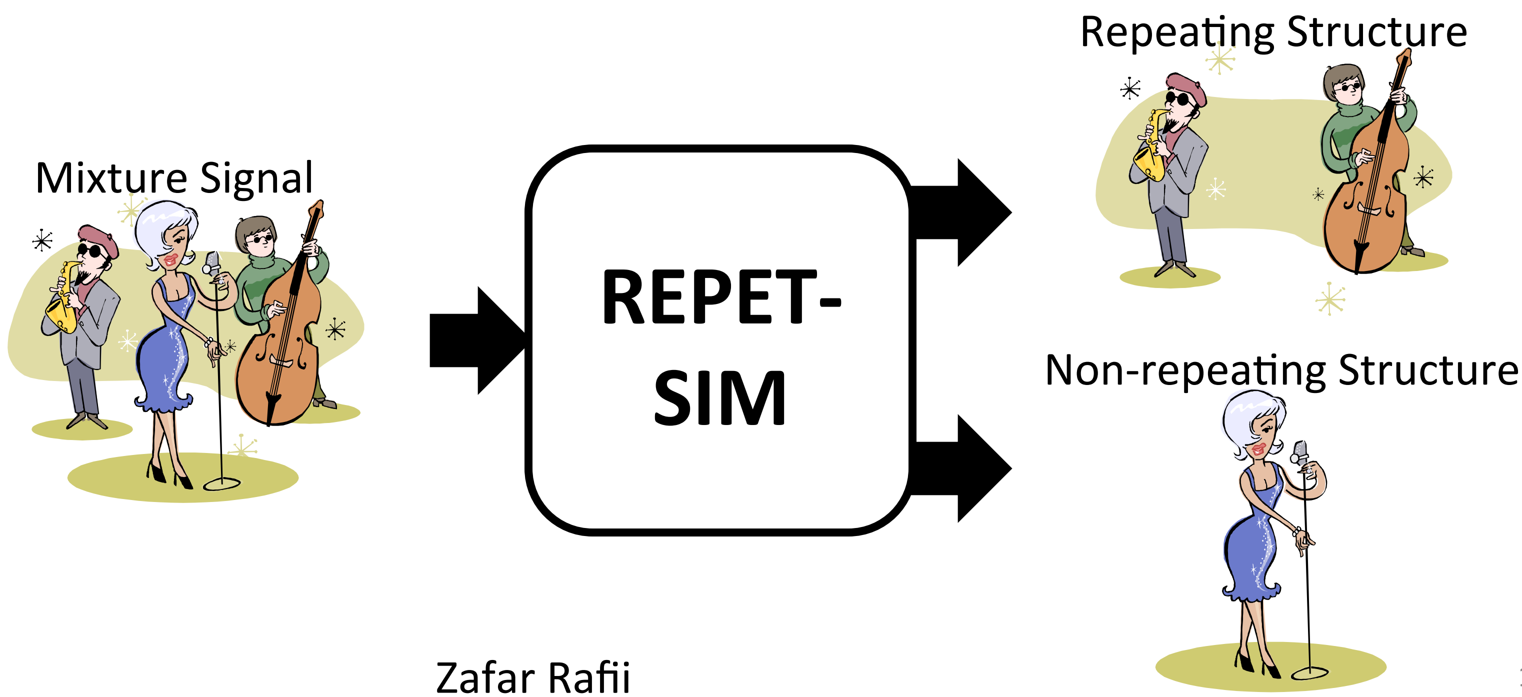
Similarity

- **REPET-SIM!**
  1. Identify the repeating/similar elements
  2. Derive a repeating model
  3. Extract the repeating structure



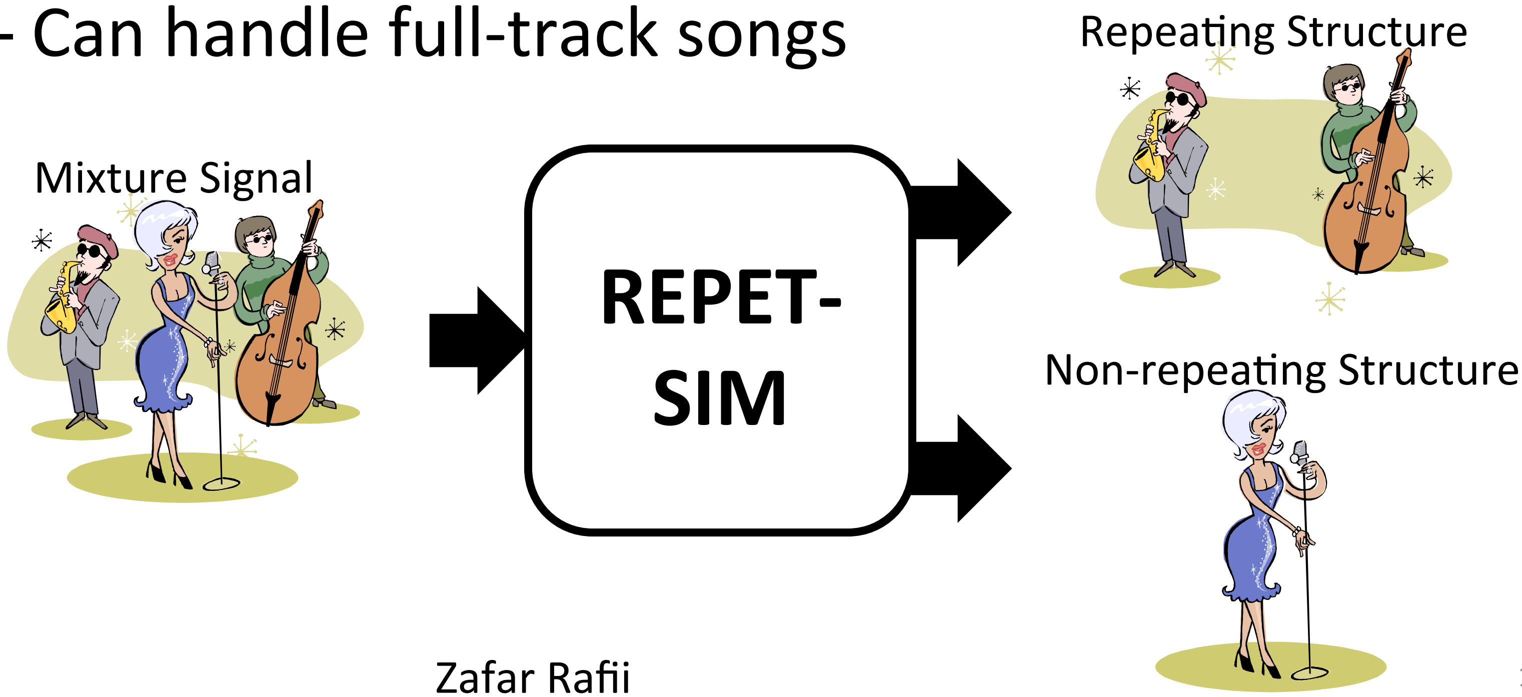
Similarity

- Simple **music/voice separation** method!
  - Repeating structure ≈ musical background
  - Non-repeating structure ≈ vocal foreground



Similarity

- **Advantages** compared with REPET:
  – Can handle intermittent repeating elements
  – Can handle fast-varying repeating structures
  – Can handle full-track songs



Outline

I. Introduction

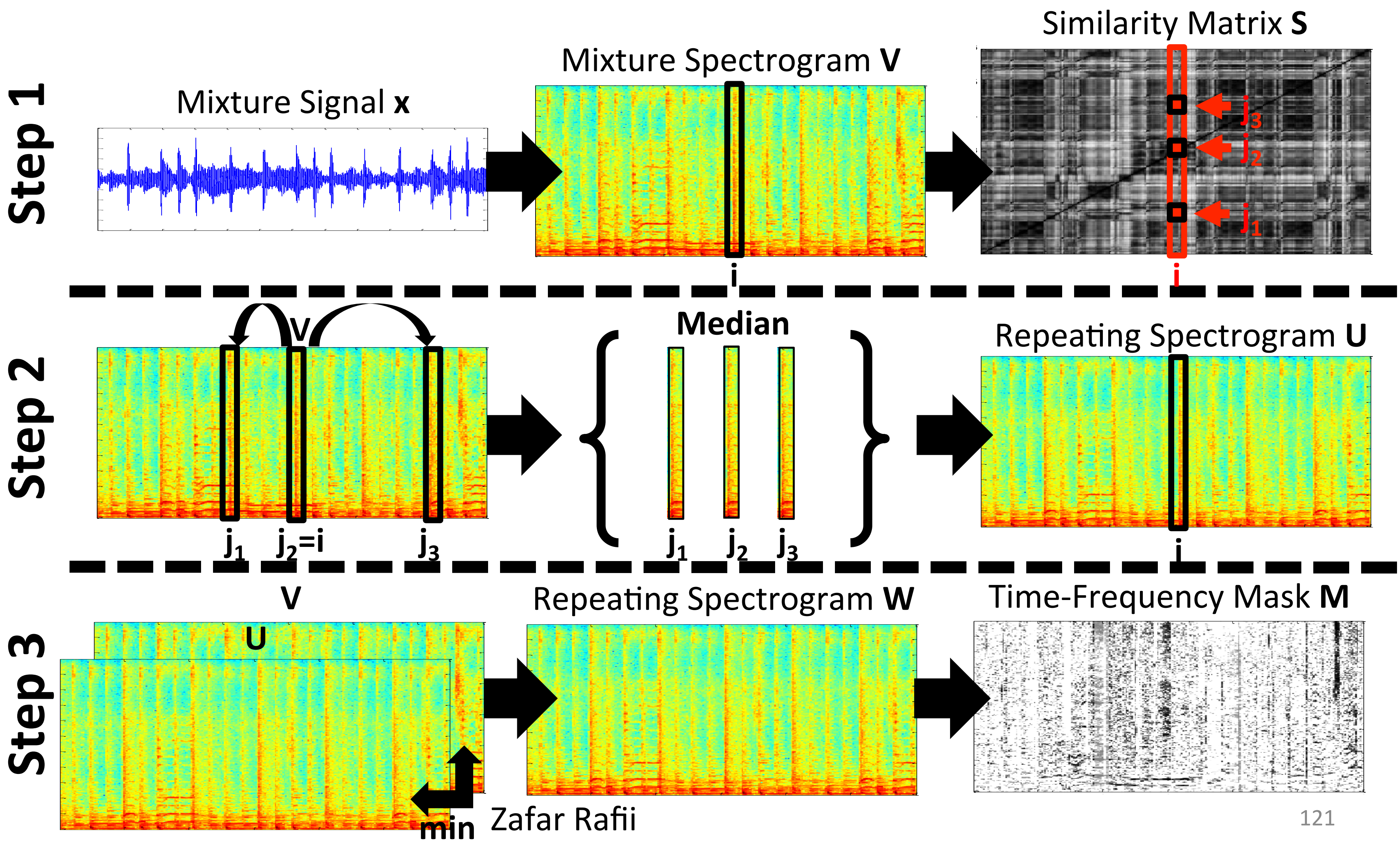
II. REPET

III. REPET-SIM

1. Similarity
2. **Method**
3. Evaluation

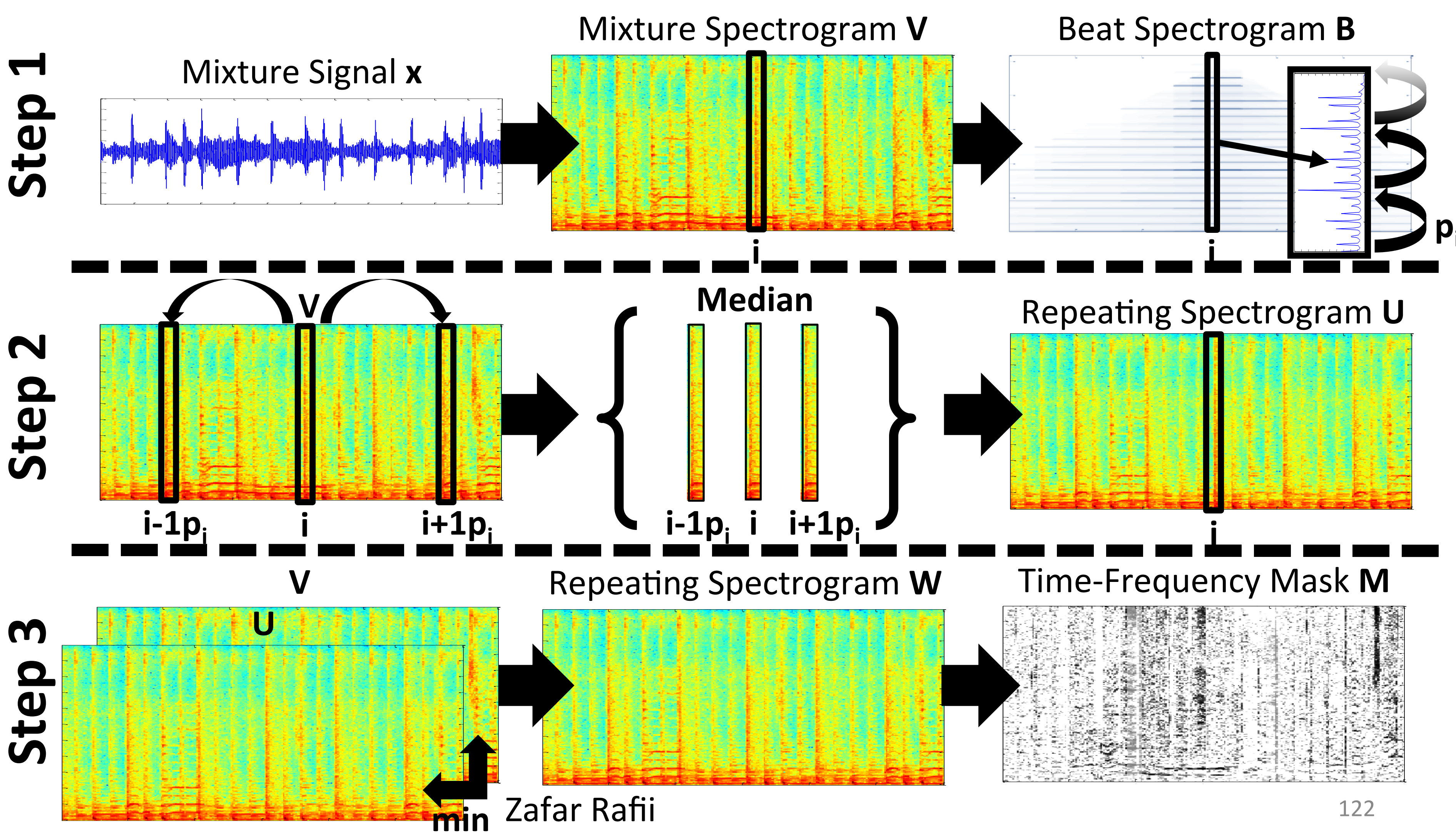
IV. Conclusion

REPET-SIM



Zafar Rafii

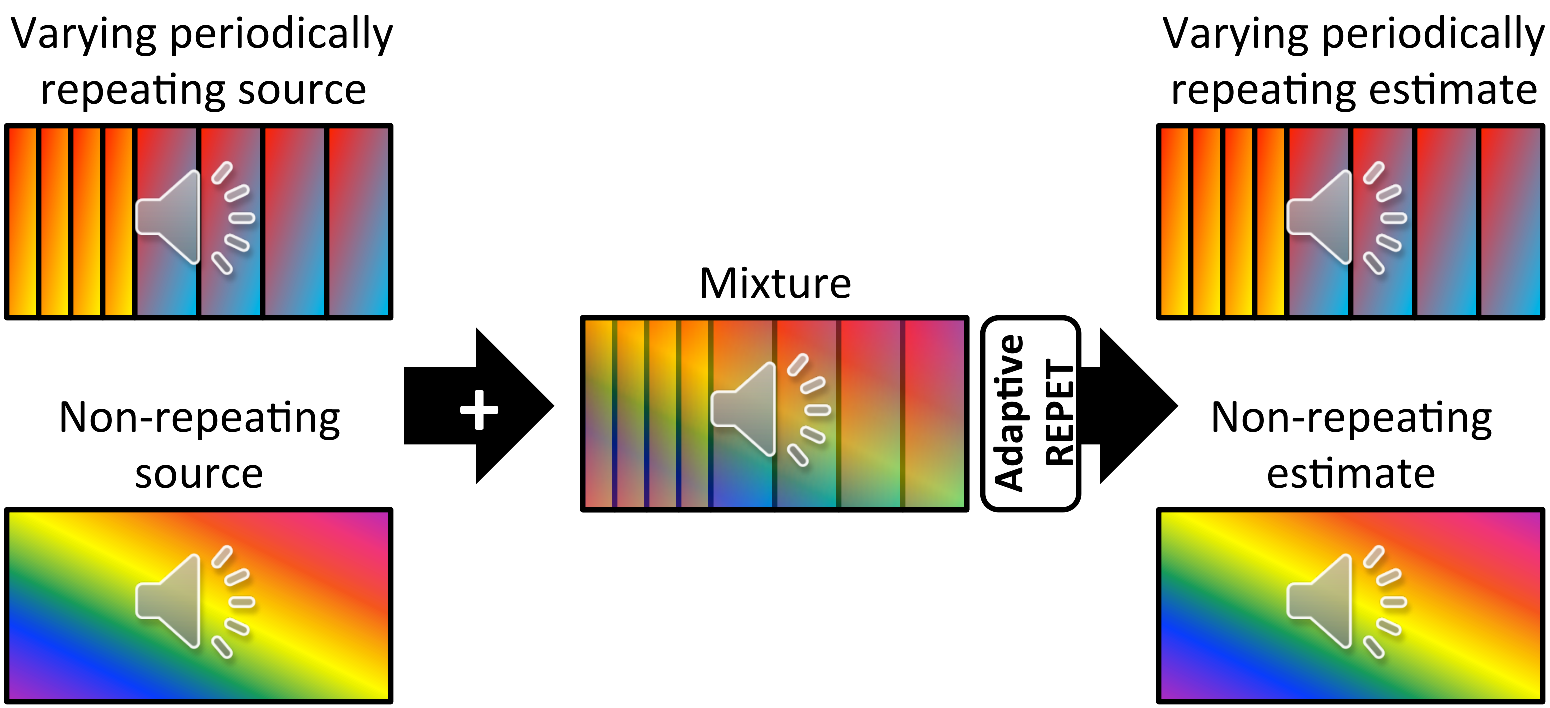
Adaptive REPET



Zafar Rafii

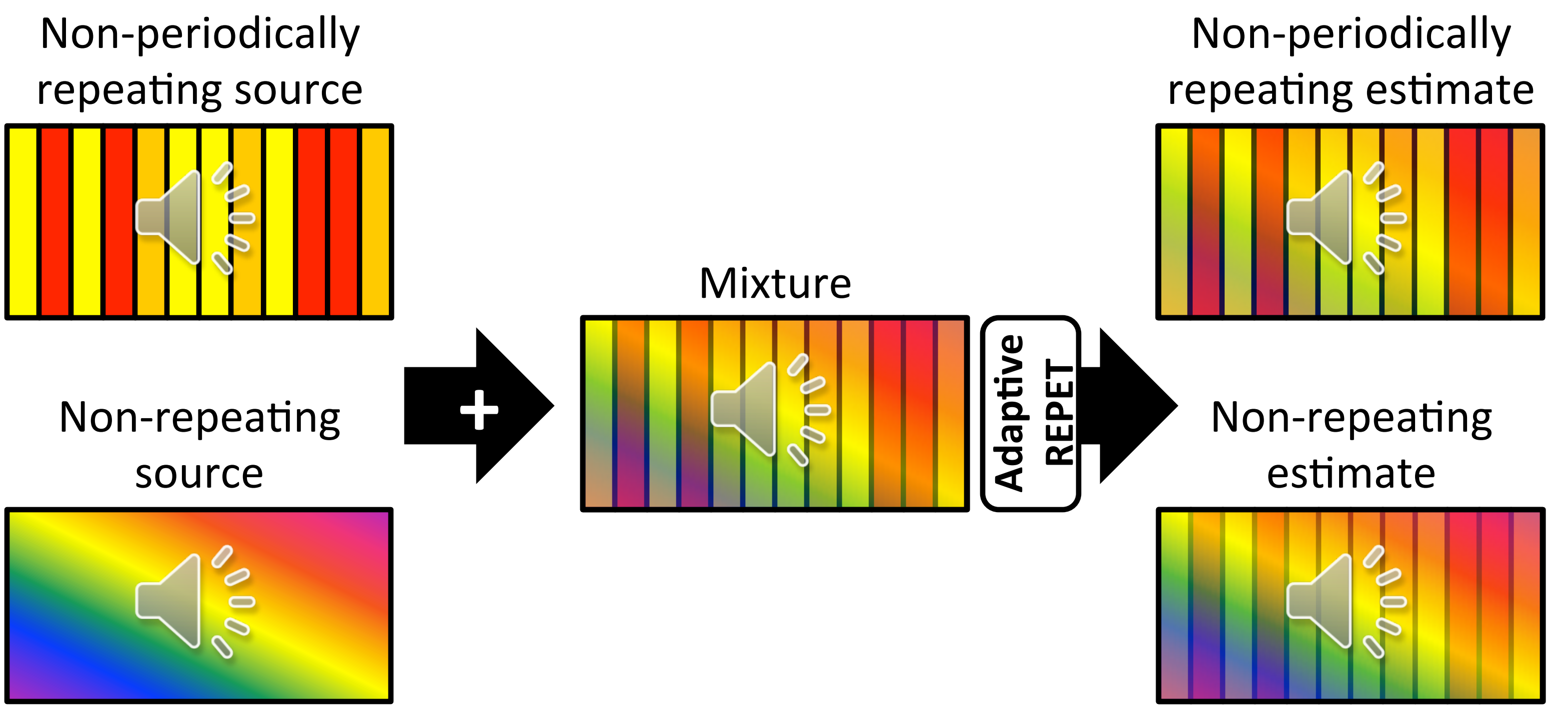
REPET-SIM vs. adaptive REPET

- The adaptive REPET can handle **varying periodically repeating structures**



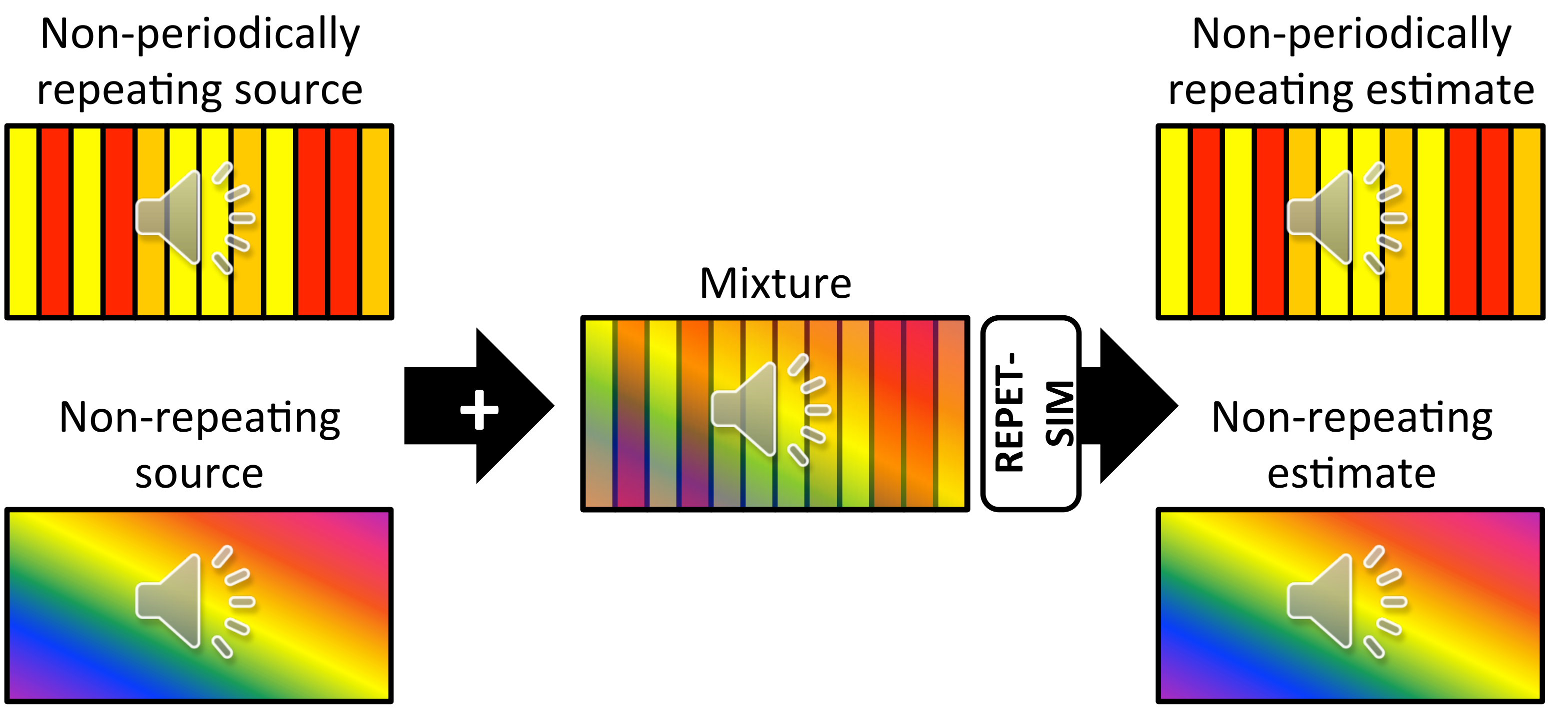
REPET-SIM vs. adaptive REPET

- The adaptive REPET shows limitations when the repeating background **is not periodical**

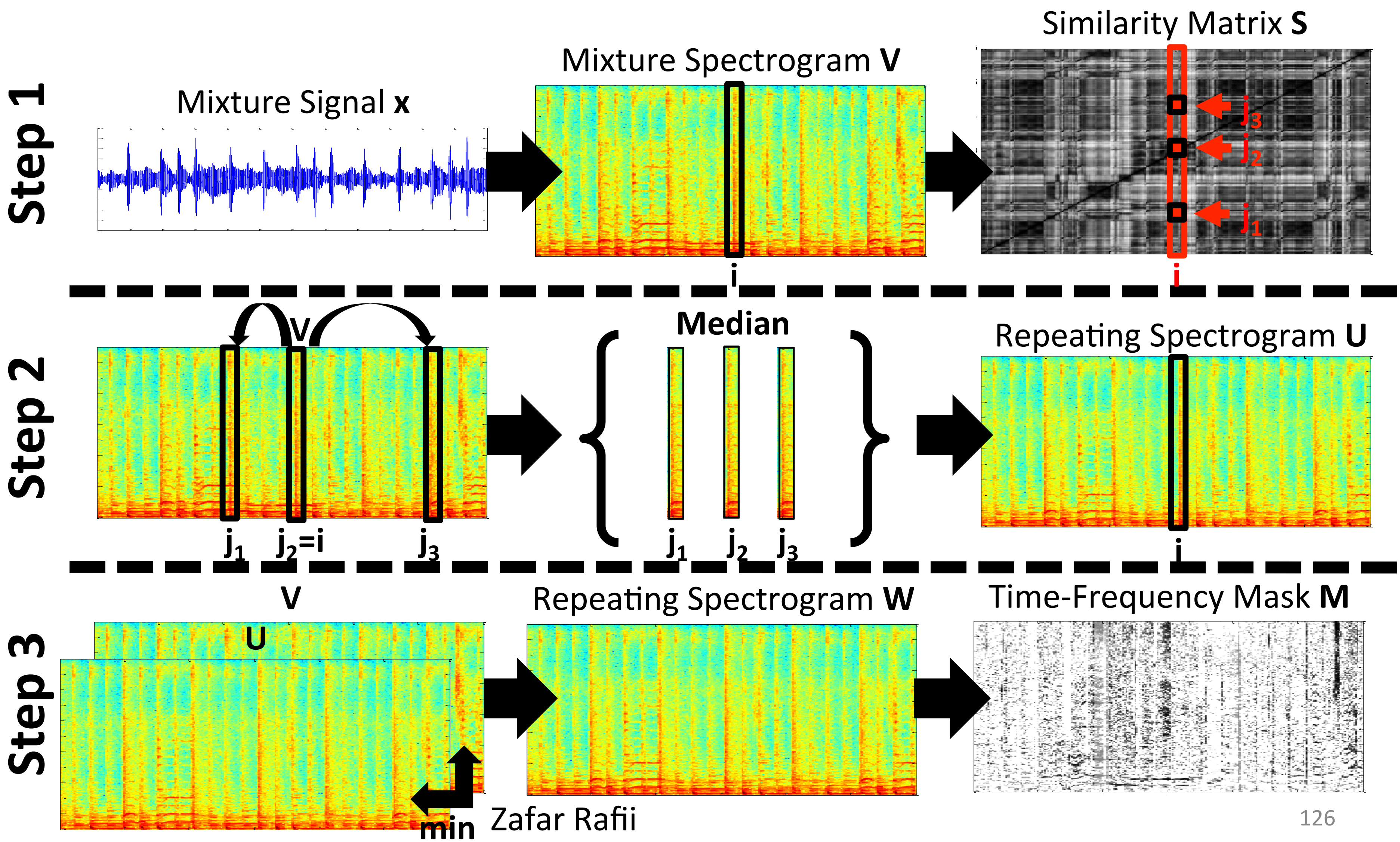


REPET-SIM vs. adaptive REPET

- REPET-SIM can also handle **non-periodically repeating structures** (e.g., in complex songs)

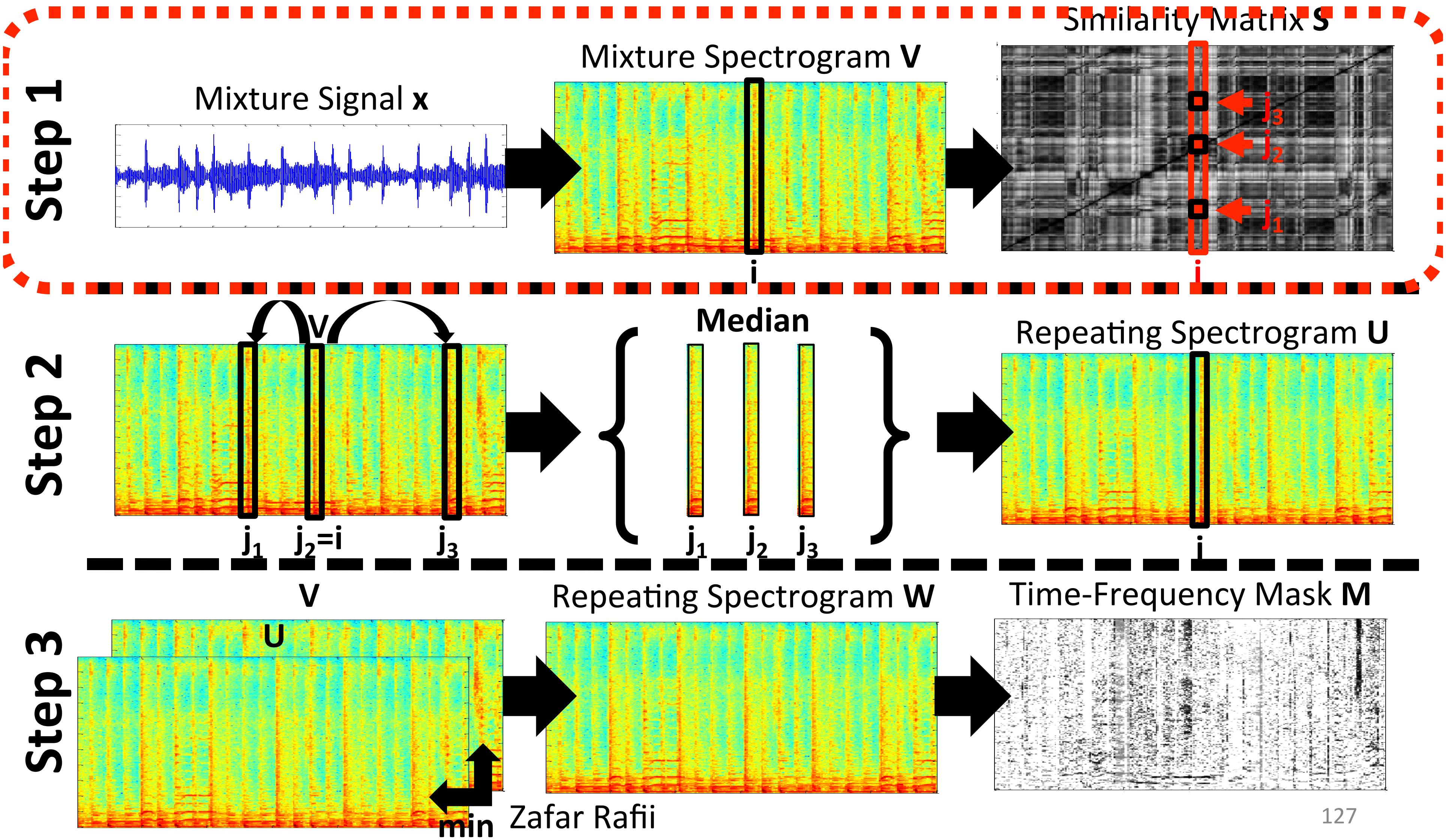


Method



Zafar Rafii

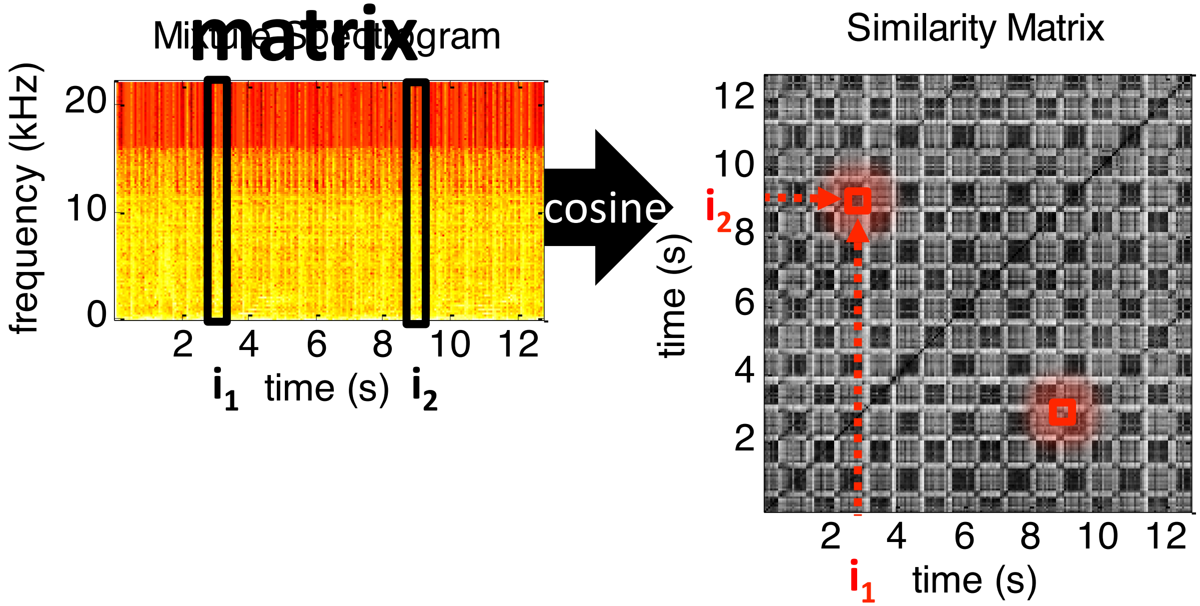
1. Repeating Elements



Zafar Rafii

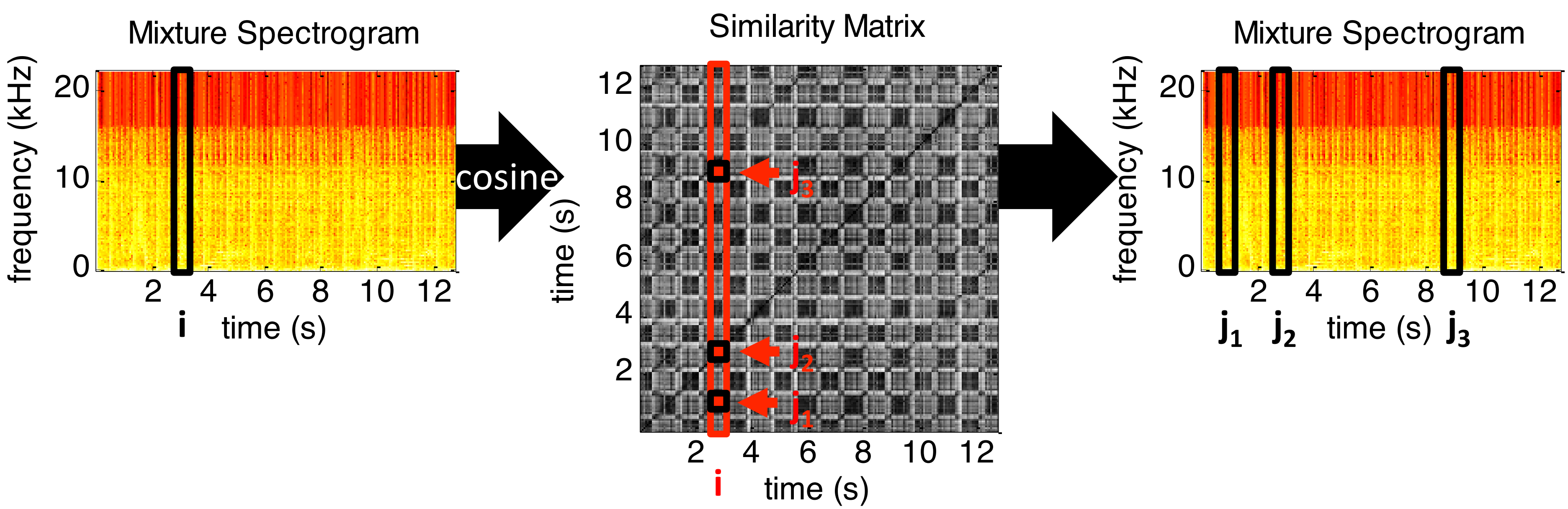
1. Repeating Elements

- We take the cosine similarity between any two pairs of columns and get a **similarity matrix**



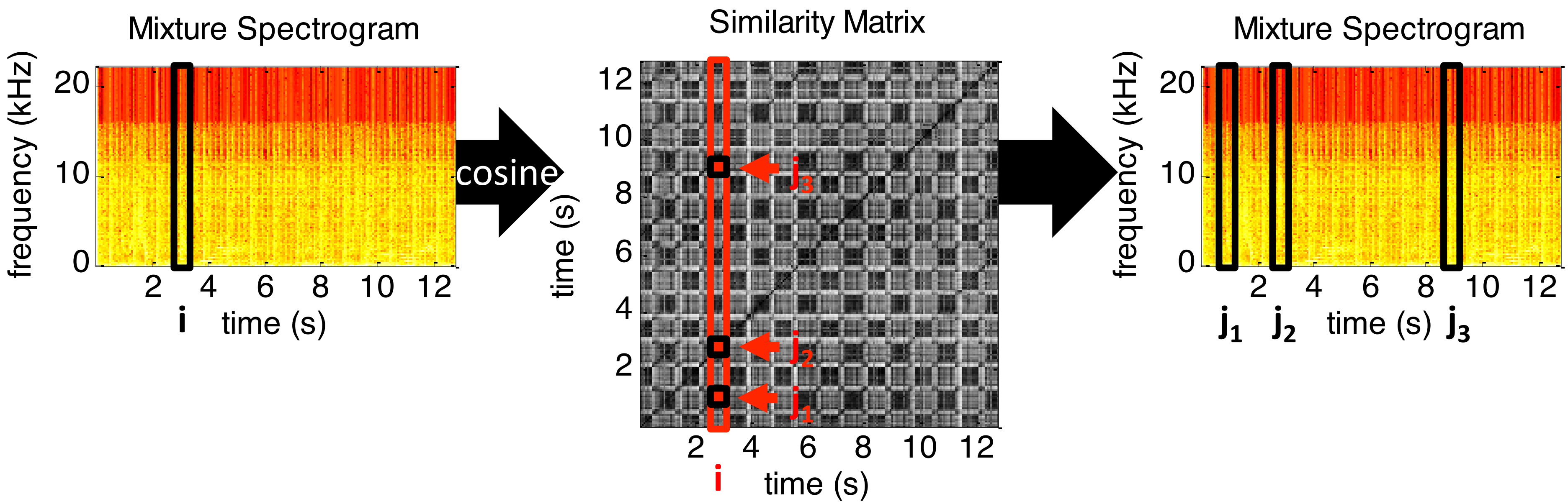
1. Repeating Elements

- The SM reveals for every frame i, the frames j_k that are **the most similar** to frame i

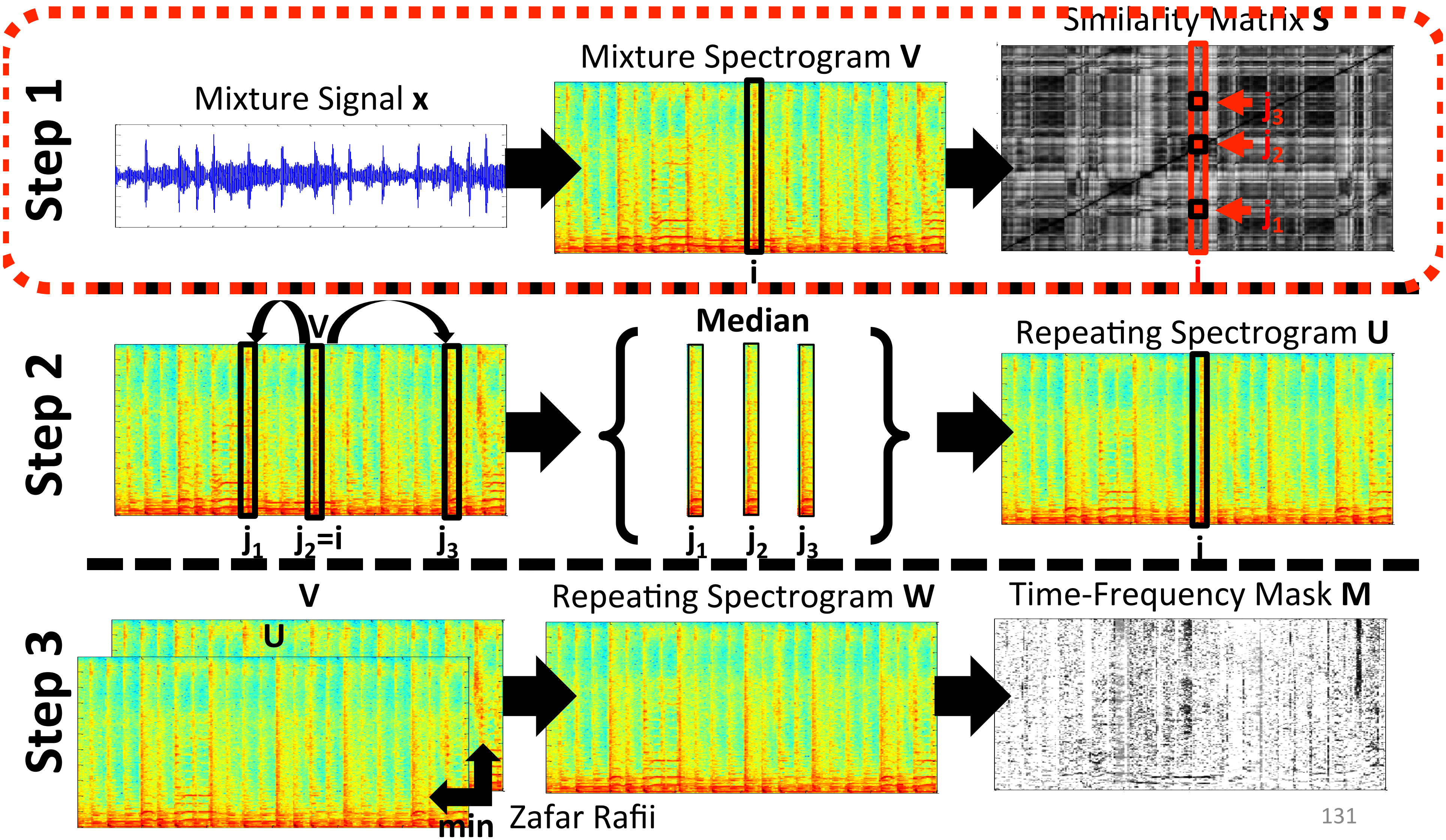


1. Repeating Elements

- We assume here that the background is **more dense and low-ranked** than the foreground

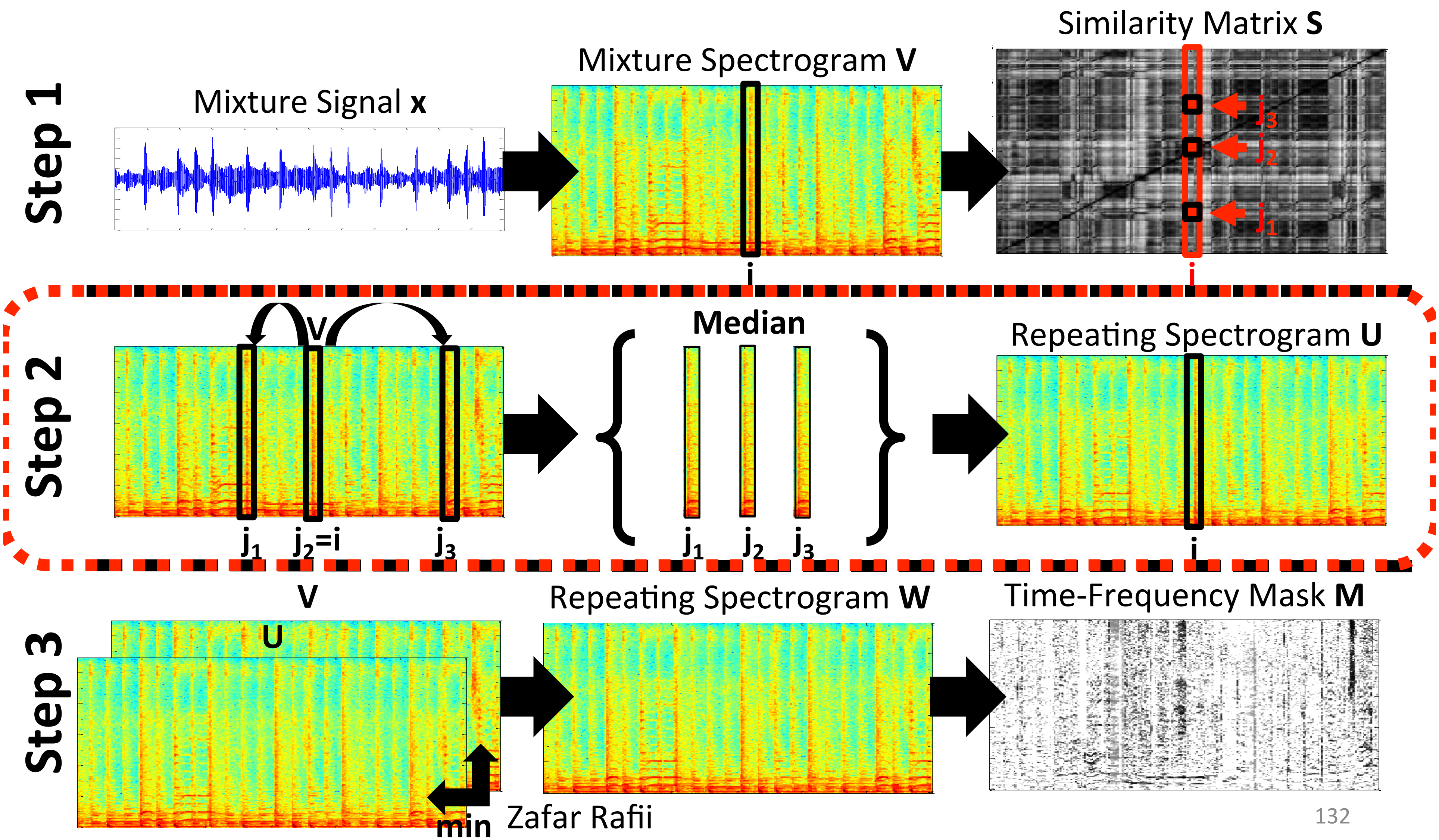


1. Repeating Elements



Zafar Rafii

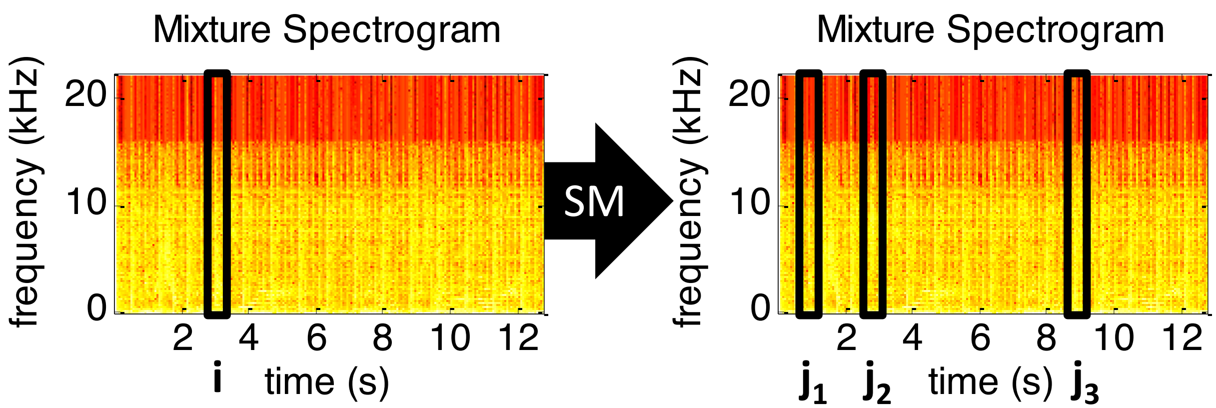
2. Repeating Model



Zafar Rafii

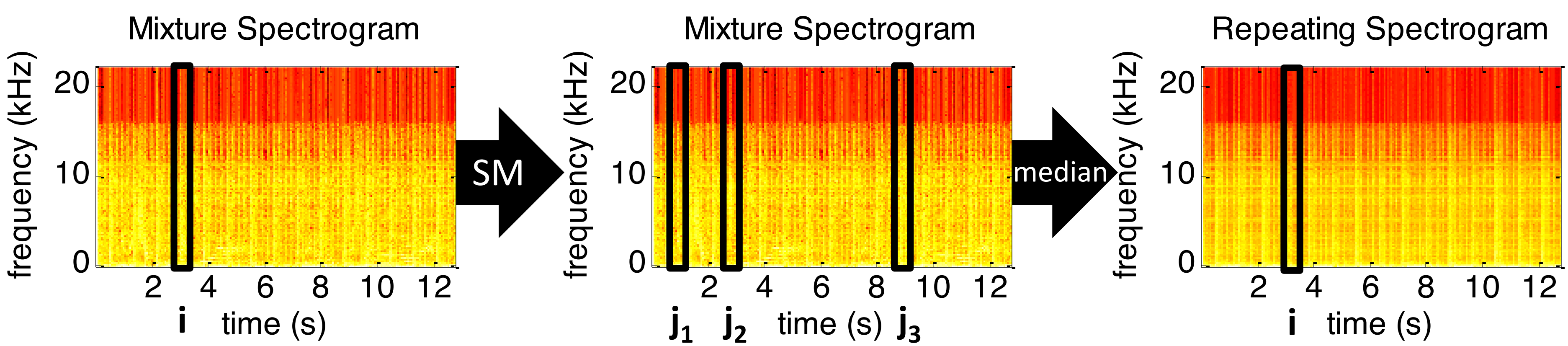
2. Repeating Model

- For every frame i, we take the **median** of the corresponding most similar frames j_k



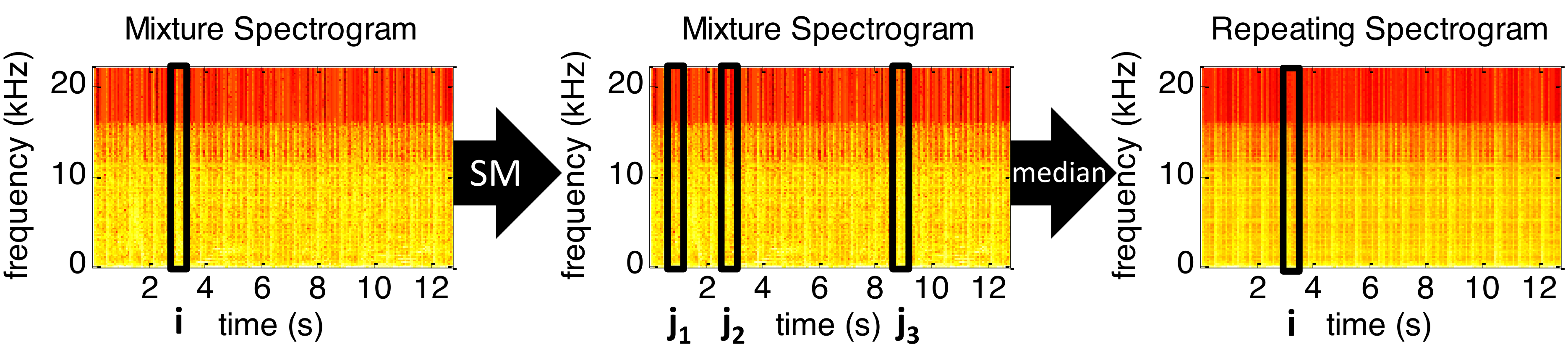
2. Repeating Model

- We obtain an initial **repeating spectrogram model**



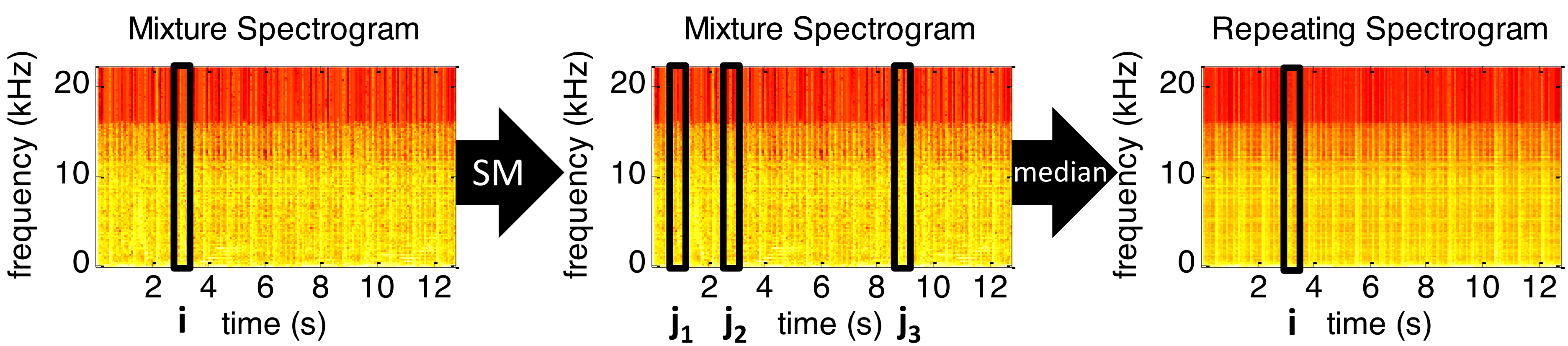
2. Repeating Model

- The **median** helps to derive a clean repeating spectrogram, removing non-repeating outliers

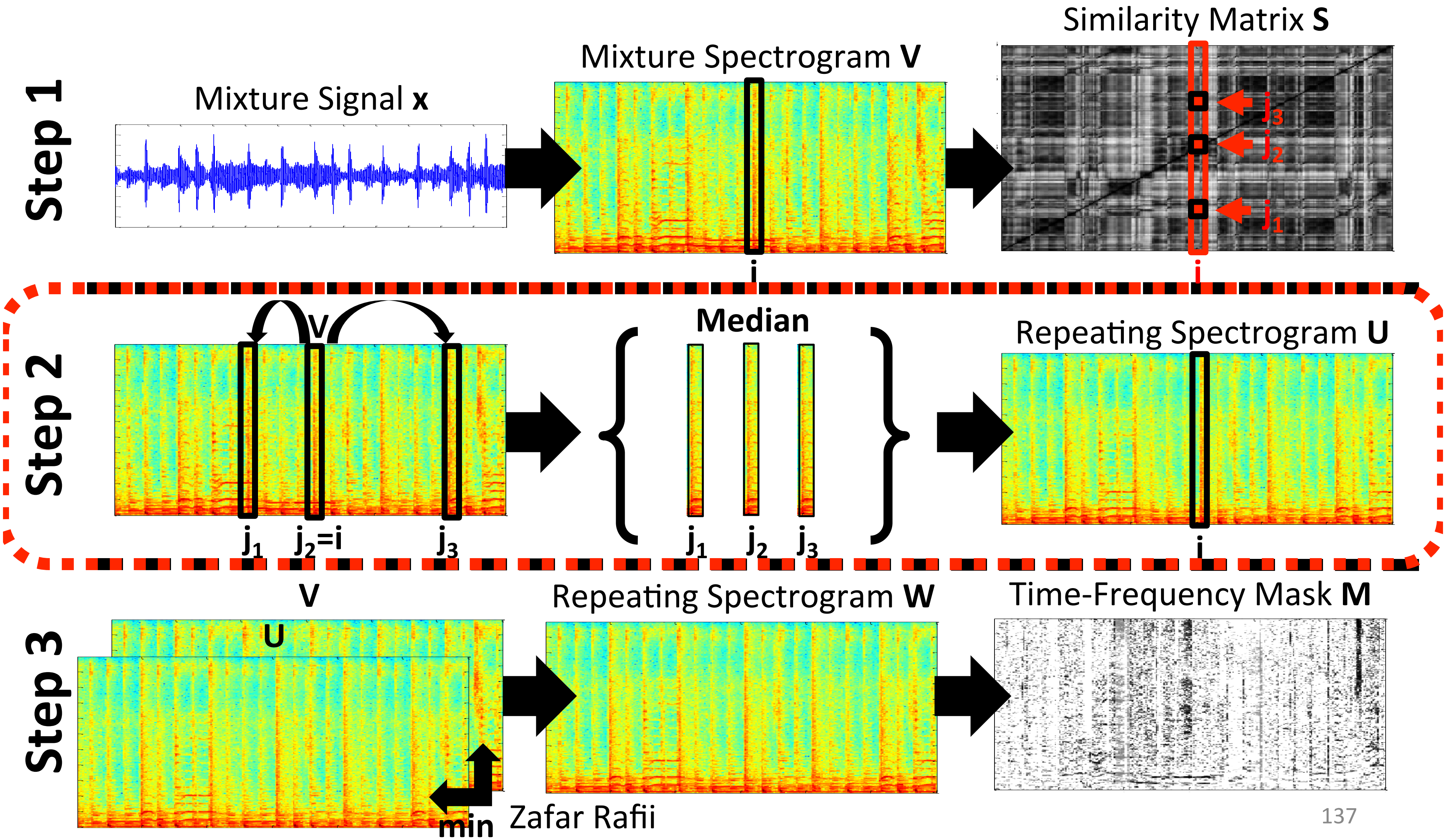


2. Repeating Model

- We assume here that the foreground is **more sparse and varied** than the background

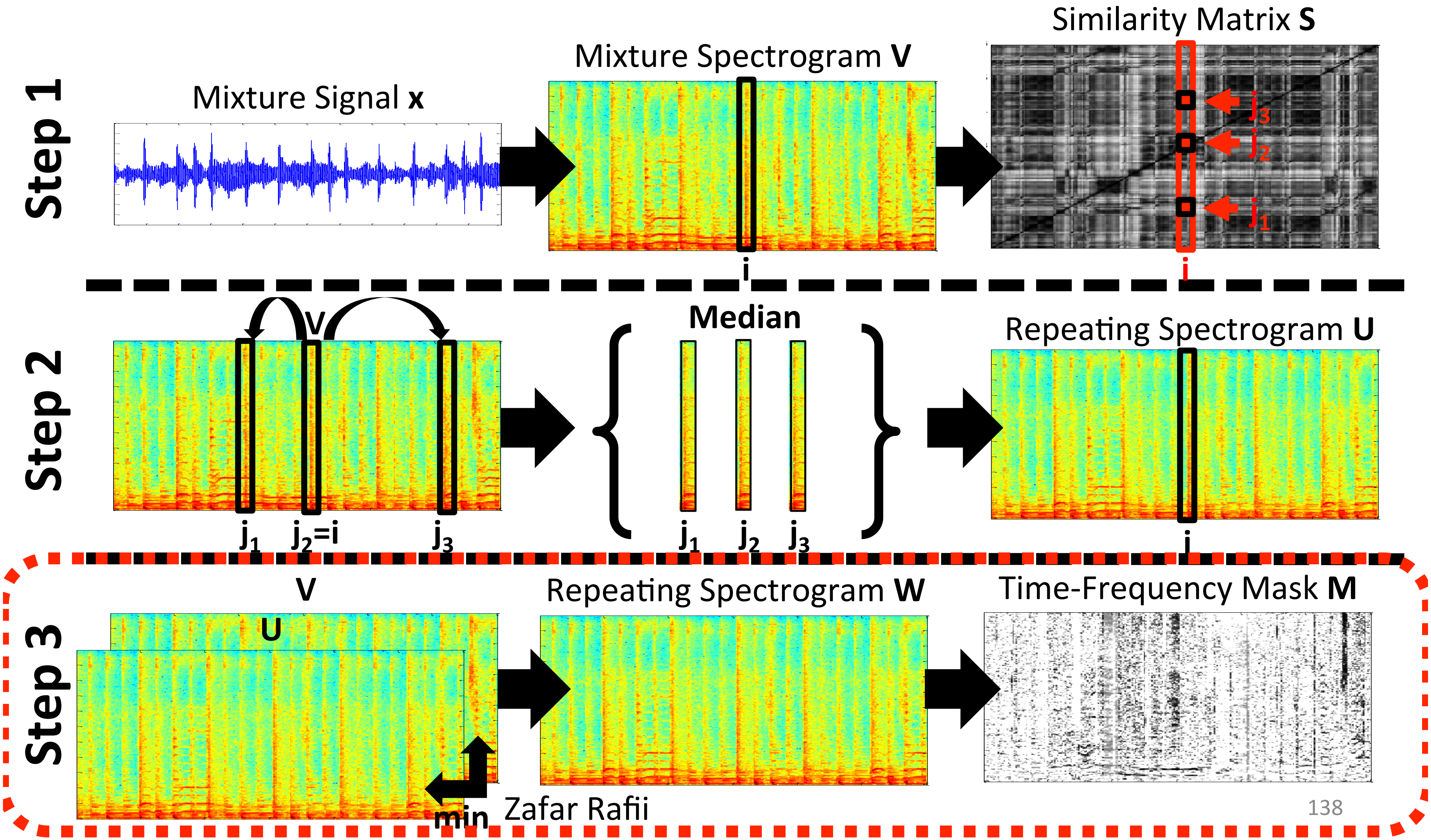


2. Repeating Model



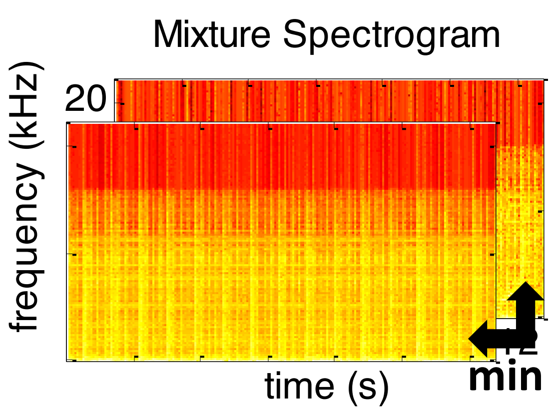
Zafar Rafii

3. Repeating Structure



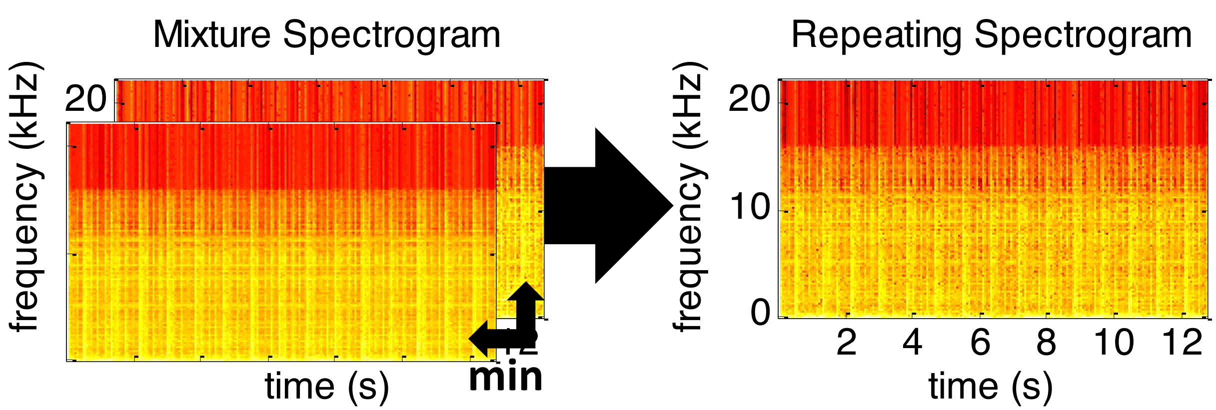
3. Repeating Structure

- We take the element-wise **minimum** between the repeating and mixture spectrograms



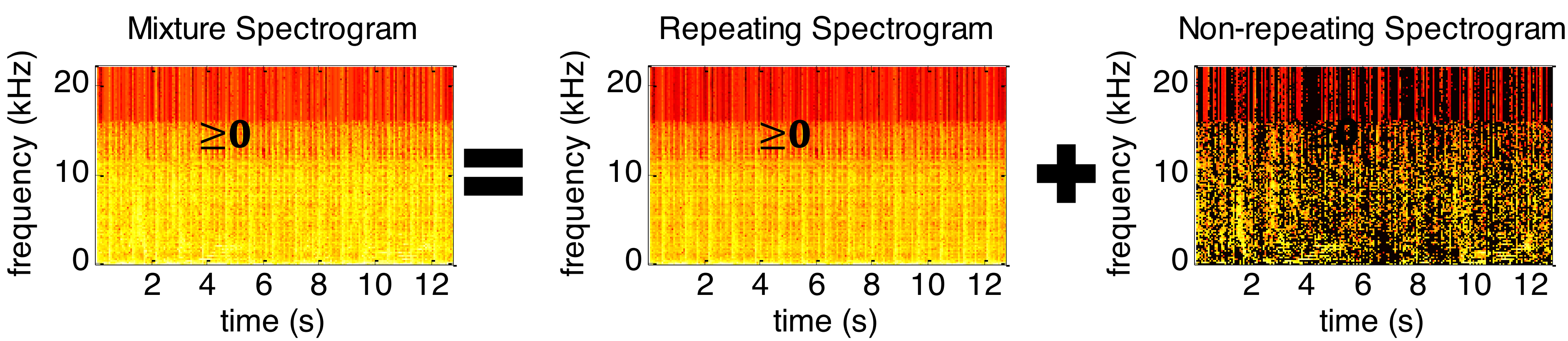
3. Repeating Structure

- We obtain a refined **repeating spectrogram model** for the repeating background



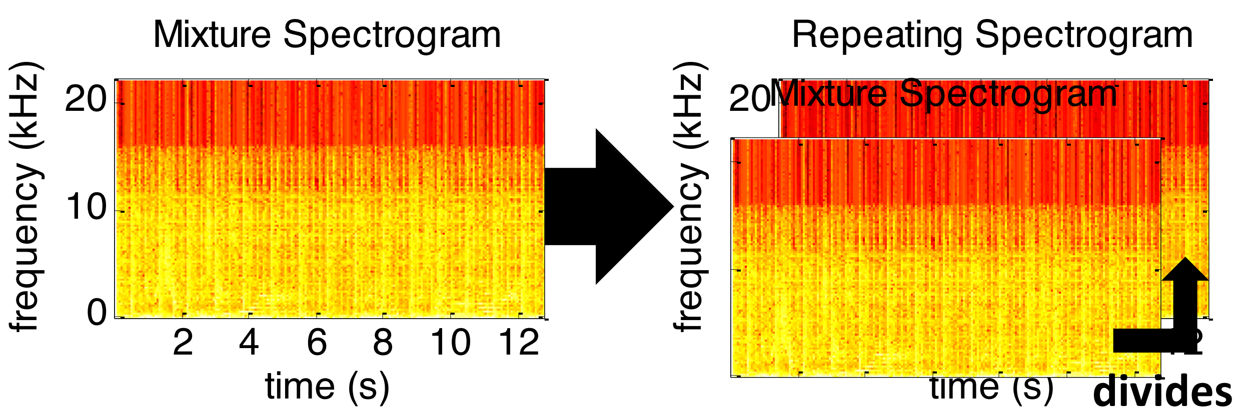
3. Repeating Structure

- The repeating spectrogram **cannot have values higher than** the mixture spectrogram



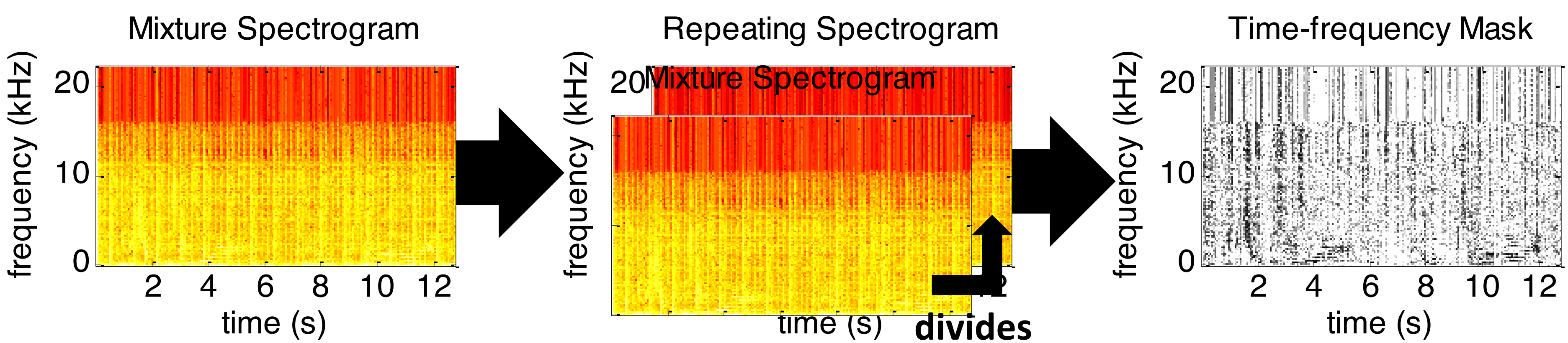
3. Repeating Structure

- We **divide** the repeating spectrogram by the mixture spectrogram, element-wise



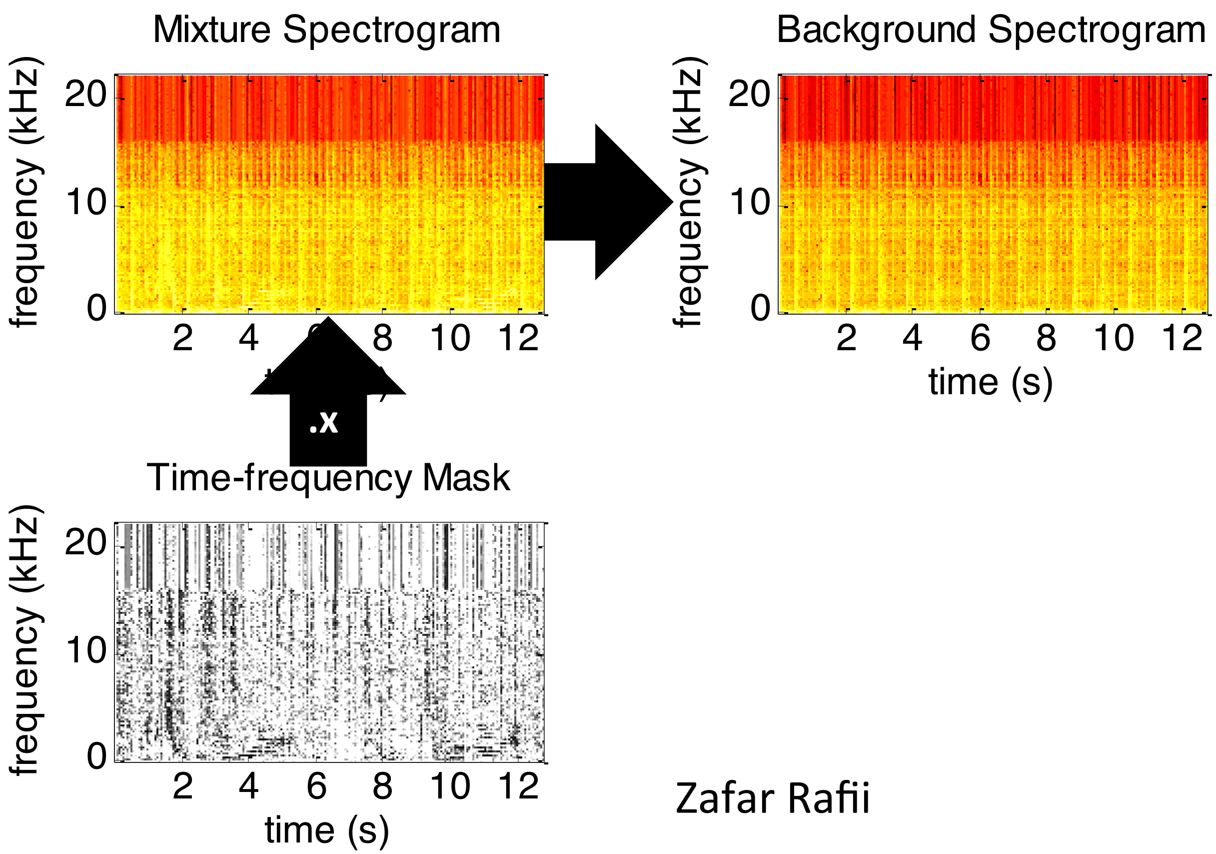
3. Repeating Structure

- We obtain a **soft time-frequency** mask (with values in [0,1])



3. Repeating Structure

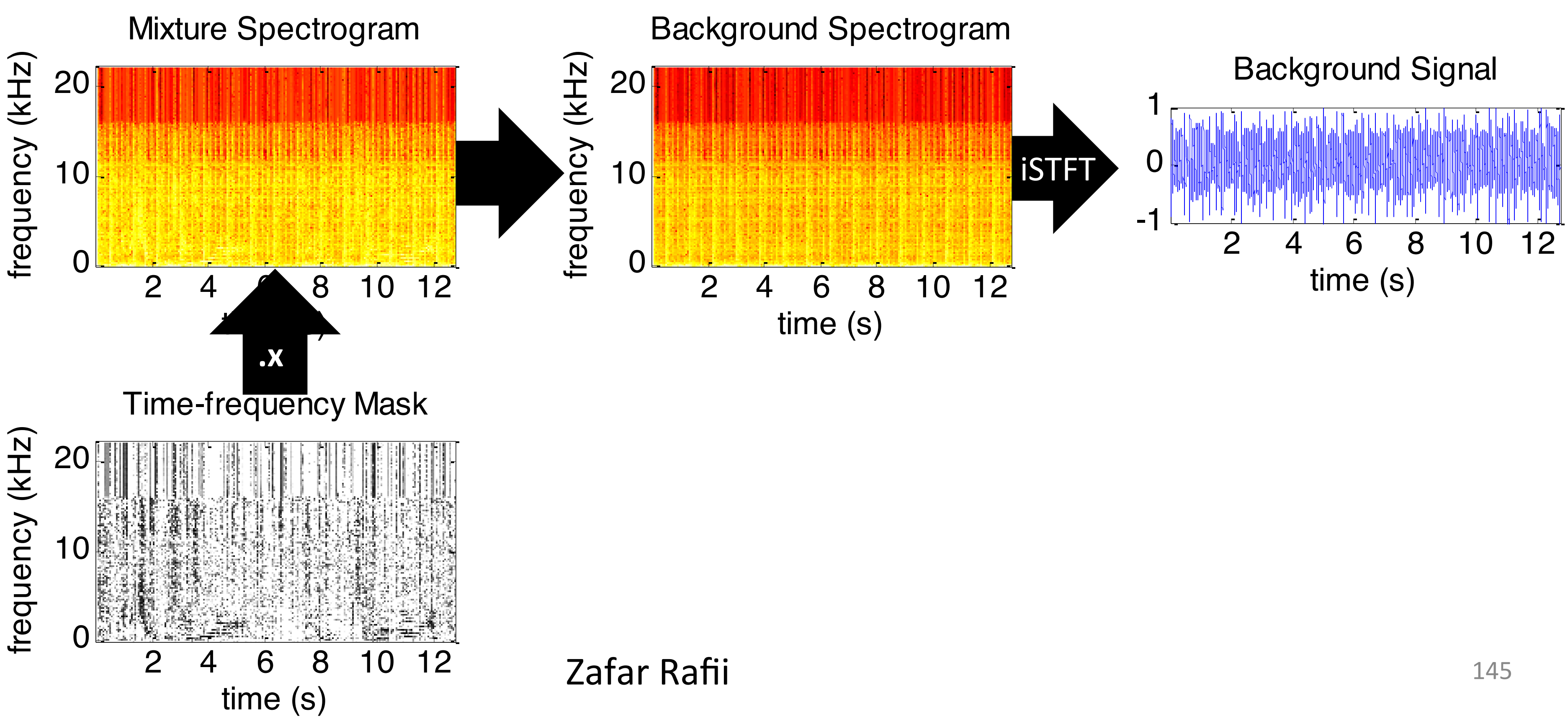
- We **multiplied** the mask with the mixture STFT to extract the repeating background STFT



Zafar Rafii

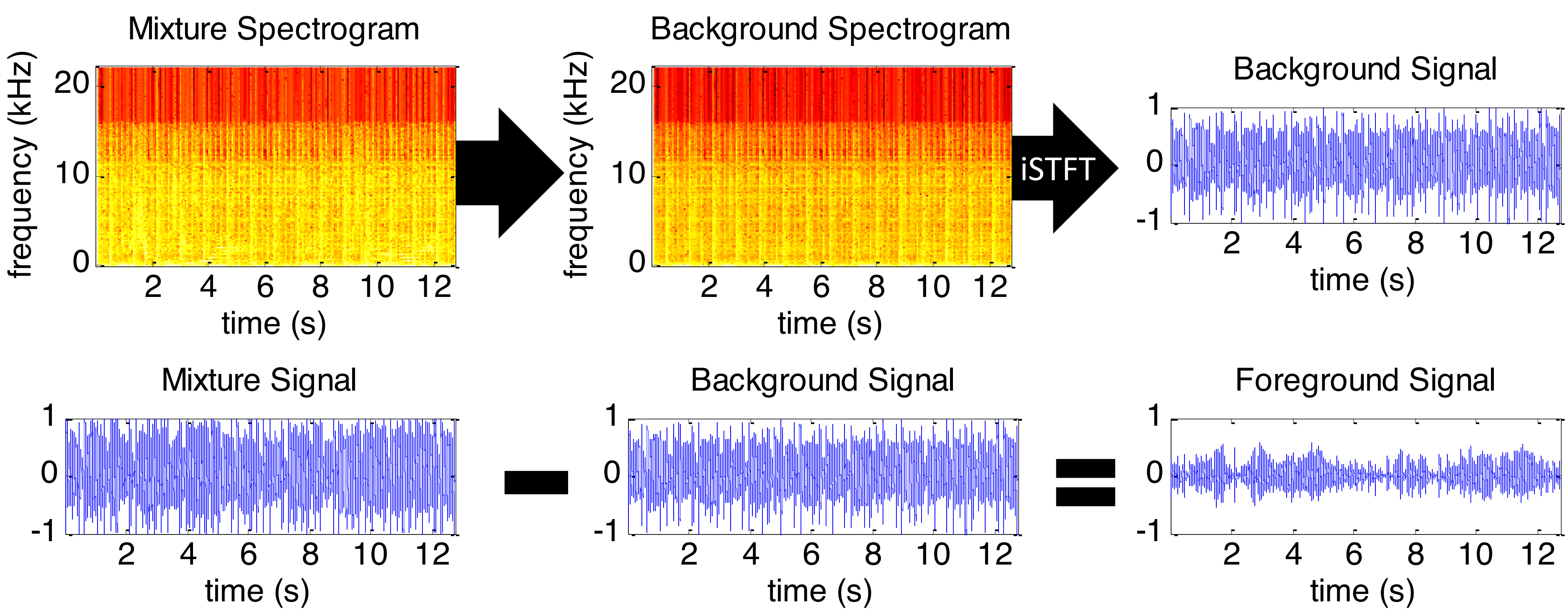
3. Repeating Structure

- The **repeating background** is obtained by inverting its STFT into the time domain



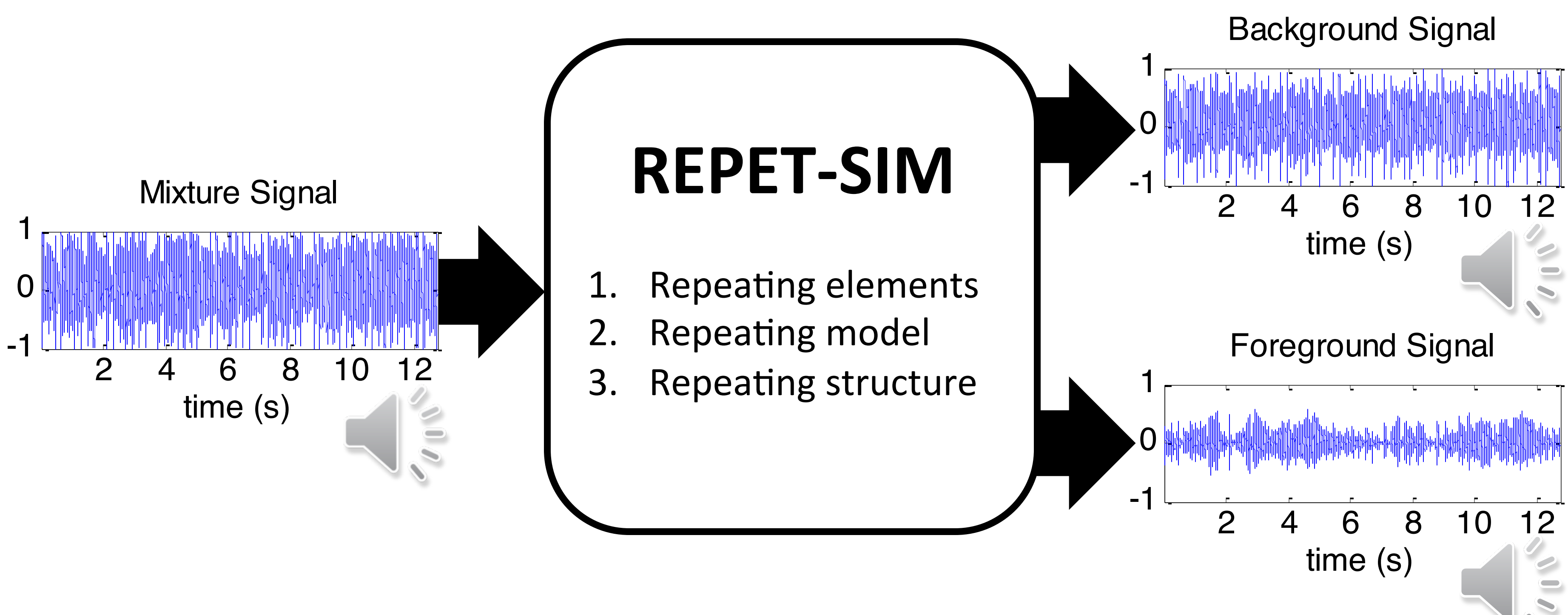
3. Repeating Structure

- The **non-repeating foreground** is obtained by subtracting the background from the mixture



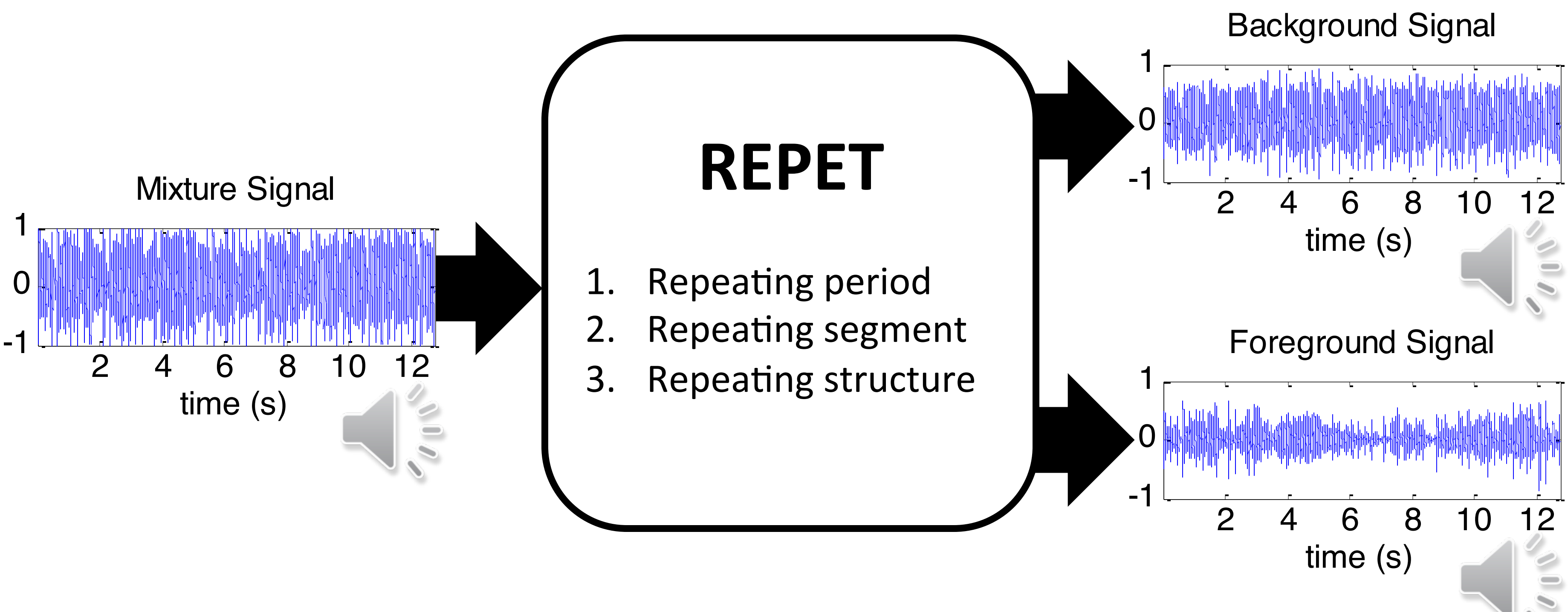
Method

- Repeating background ≈ **music component**
- Non-repeating foreground ≈ **voice component**



Method

- Repeating background ≈ **music component**
- Non-repeating foreground ≈ **voice component**



Outline

I. Introduction

II. REPET

III. REPET-SIM

1. Similarity
2. Method
3. **Evaluation**

IV. Conclusion

Evaluation

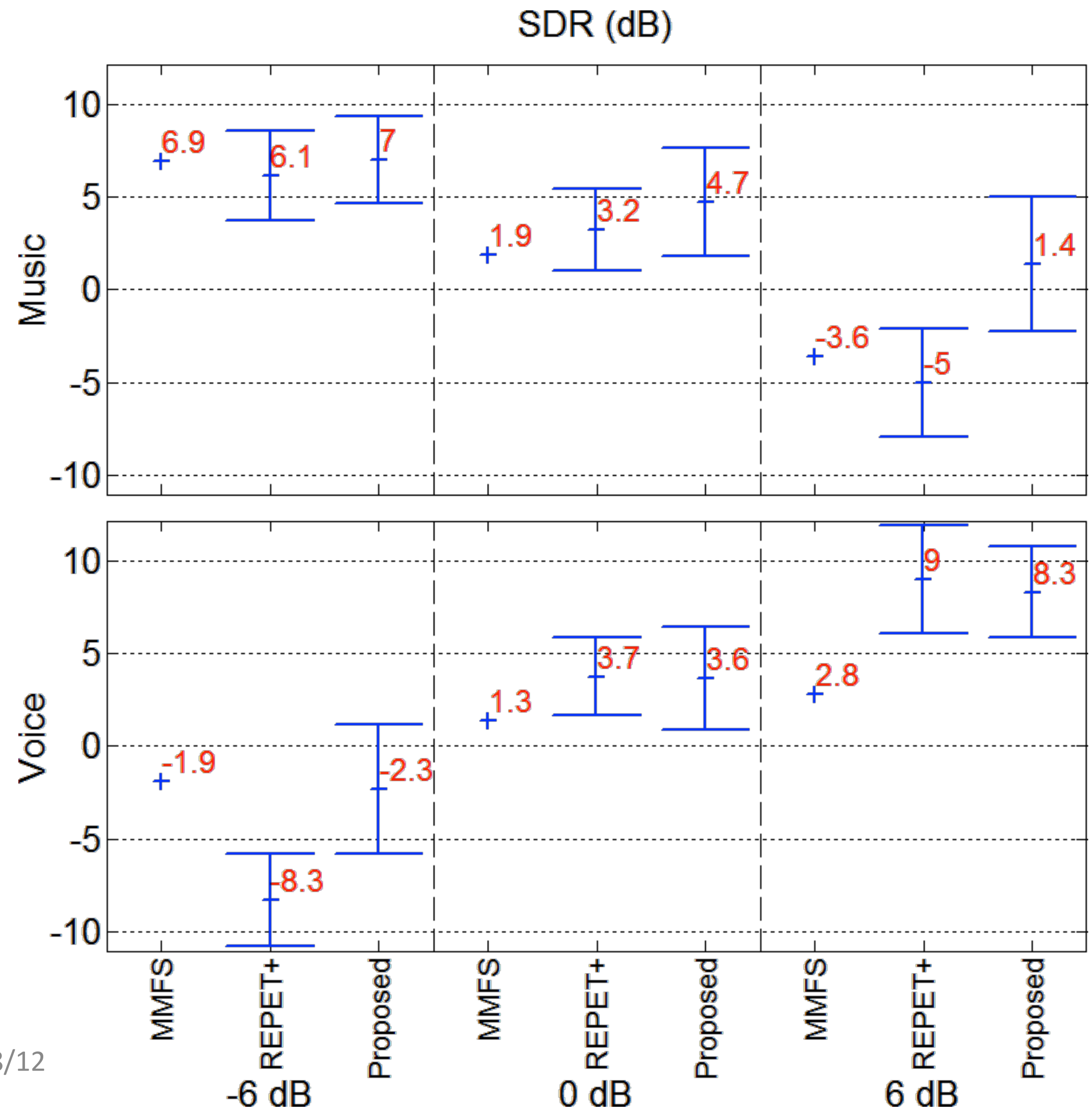
- **REPET-SIM** [Rafii et al., 2012]
  - Cosine similarity
  - Soft time-frequency masking

- **Competitive method 1** [FitzGerald et al., 2010]
  - Median filtering of the spectrogram at different frequency resolutions to extract the vocals

- **Competitive method 2** [Liutkus et al., 2012]
  - Adaptive REPET with automatic periods finder and soft time-frequency masking

- **Data set**
  - 14 full-track real-world songs (from The Beach Boys)
  - 3 voice-to-music mixing ratios (-6, 0, and 6 dB)

Evaluation



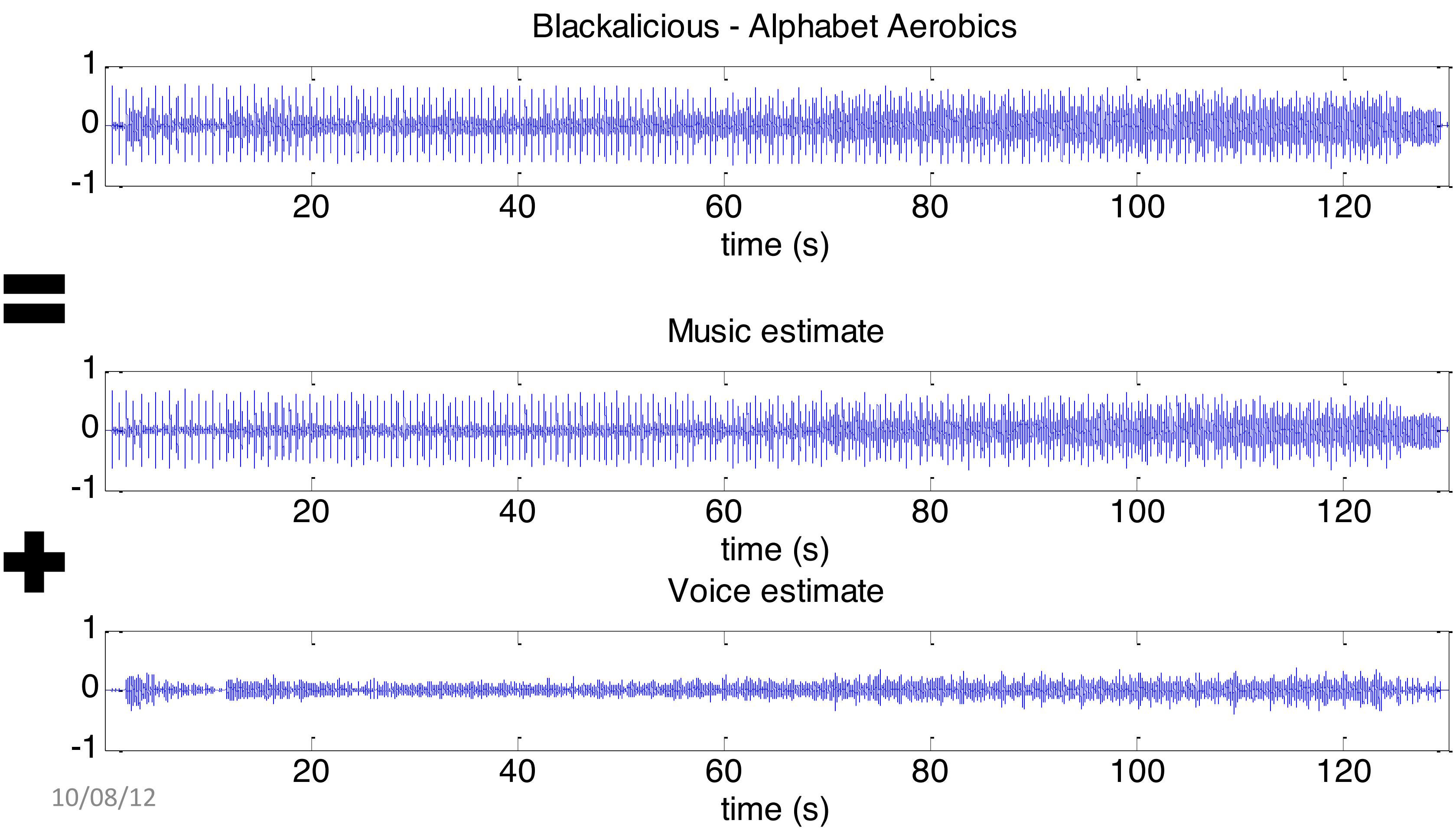
MMFS = FitzGerald et al.
REPET+ = Adaptive REPET
Proposed = REPET-SIM

Evaluation

- **Conclusions**
  - REPET-SIM can compete with a recent music/voice separation method
  - REPET-SIM can perform as well as the adaptive REPET
  - Average computation time: 0.563 second for 1 second of mixture (vs. 1.183 seconds for Adaptive)

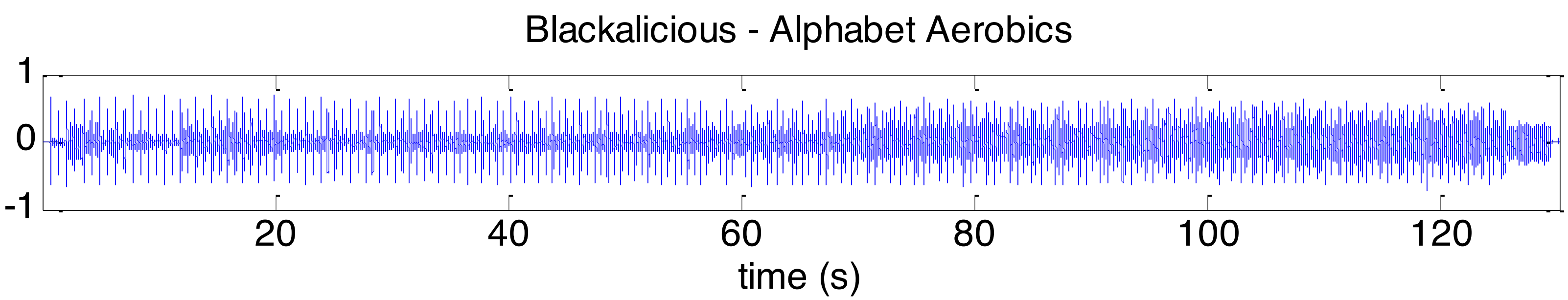
Examples

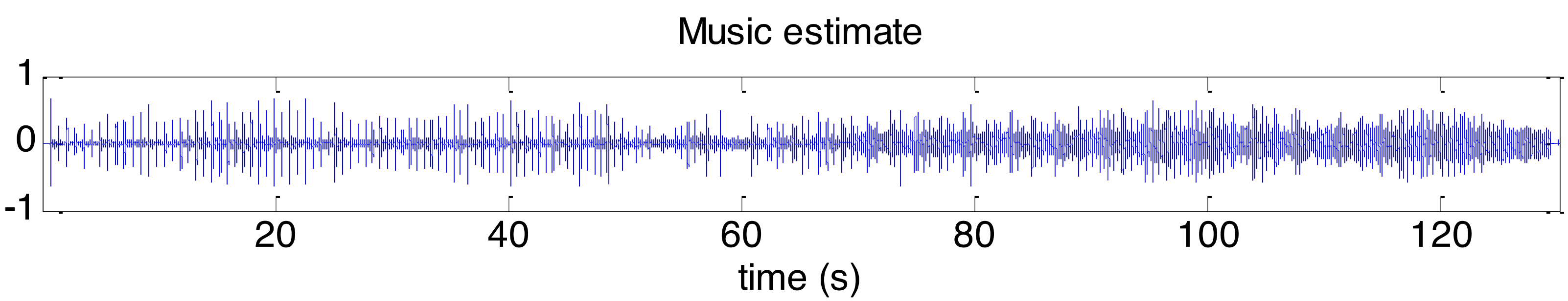
- REPET-SIM

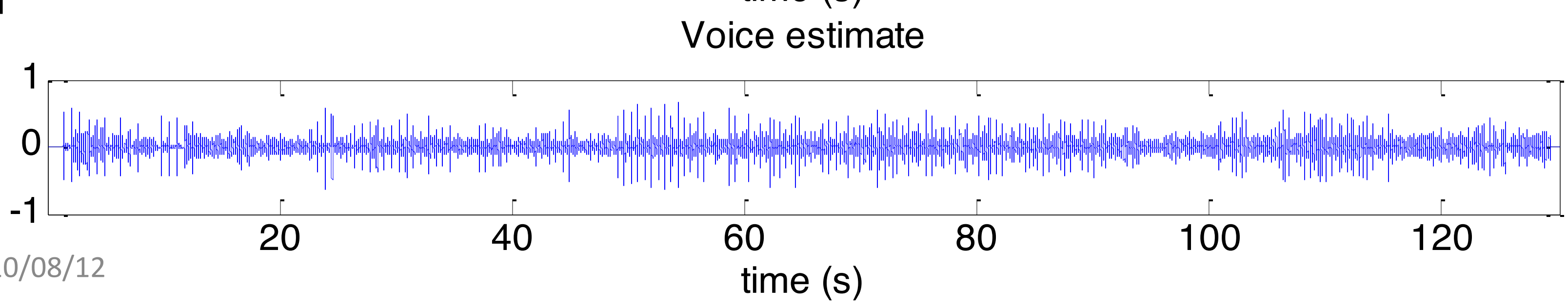


Examples

- Adaptive REPET







Outline

I. Introduction
II. How humans use repetition to identify sound sources (McDermott)
III. Coffee break
IV. Repetition-based algorithms for source separation (Rafii)
V. Links to other methods for source separation
VI. Conclusions/Questions

Links to Other Source Separation Methods

Bryan Pardo
Electrical Engineering & Computer Science
School of Music
Northwestern University

Closely related methods

- Nearest Neighbor Median Filtering

- Robust Principal Component Analysis

Nearest Neighbor Median Filtering
(Fitzgerald 2012)

- Essentially identical to REPET-SIM

  differences include:

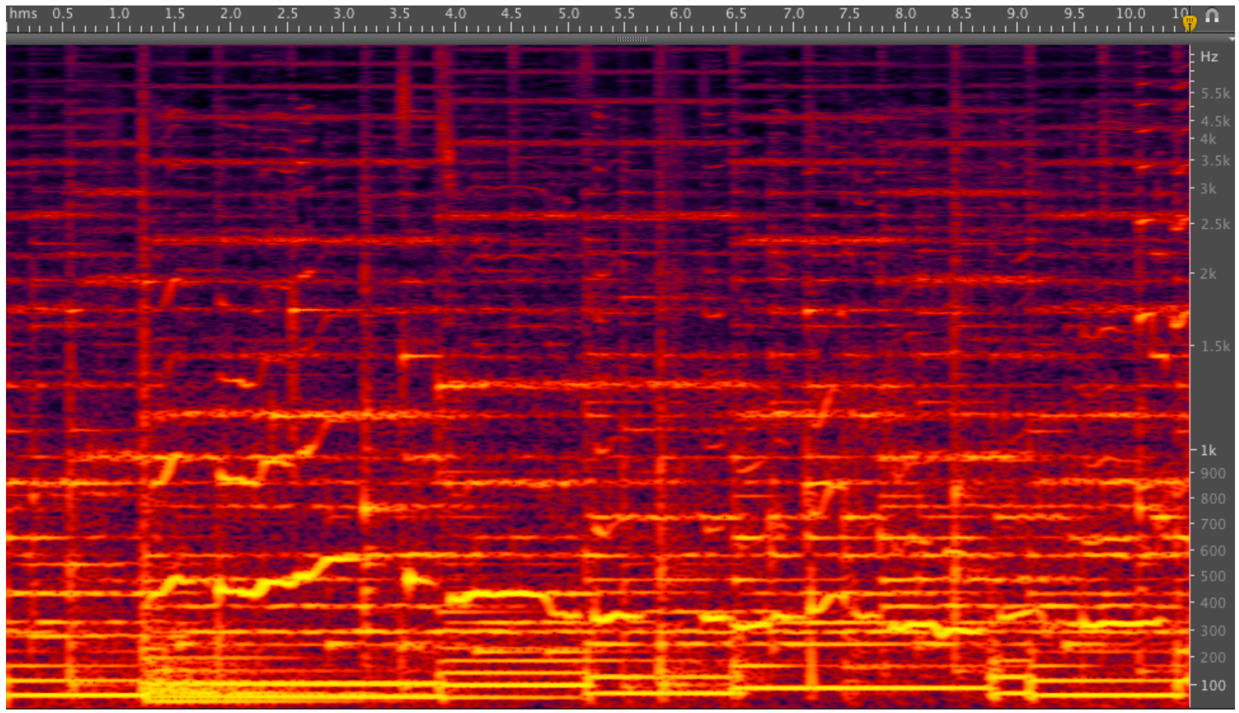
  Squared Euclidean distance replaces Cosine similarity

  No prohibition on using immediate temporal neighbor frames as repetitions

  Let's see what allowing temporal neighbors as repetitions does…

Robust Principal Component Analysis
(Candes 2009, Huang 2012)

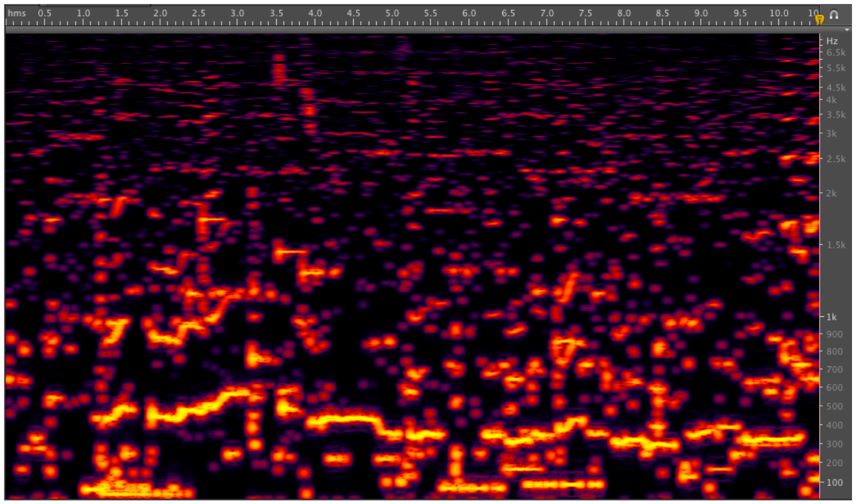
Separate an original matrix *M* into…

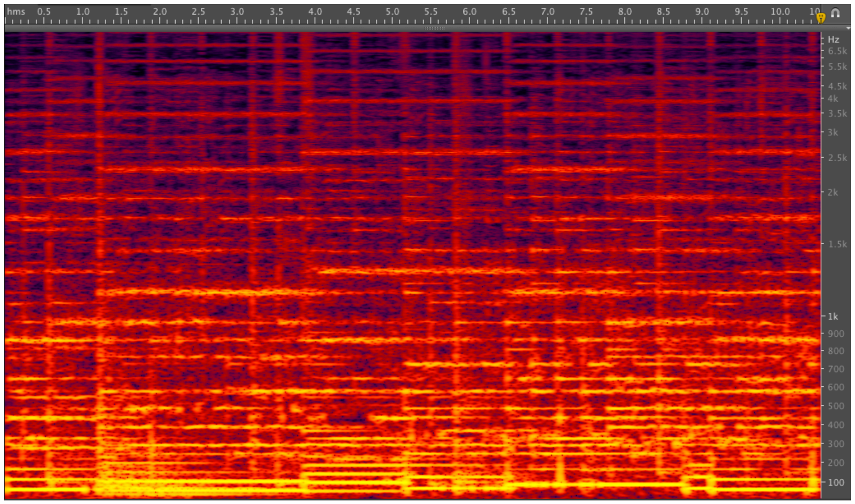


(We'll hear this example later)

Robust Principal Component Analysis

A Sparse Matrix, *S* AND A Low-rank Matrix, *L*





How? Minimize $\| L \|_* + \lambda \| S \|_1$

...subject to constraing $L + S \approx M$

Similar goals to REPET-SIM

RPCA Assumptions

- Sparse matrix *S* must NOT be low rank

  Translation: Non repeating elements must be distributed throughout the audio.

  Problematic example: Repeated funk riff with the occasional "good god"

- Low rank matrix *L* must NOT be sparse

  Translation: It works better if your accompaniment occupies a lot of the spectrum (chords, snare drums)

  Problematic example: Voice + Acoustic Bass

Slow

- Original approach used Iterative Thresholding.
- Converges extremely slowly

  About 10^4 iterations to converge

  Each iteration requires one singular value decomposition.

  A matrix of m = 800, took 8 hours on a PC from 2009.
- Accelerated Proximal Gradient is 50x faster

  About 10 minutes for the same matrix

Faster

- Huang et al (ICASSP 2012) use the Augmented Lagrange Multiplier (ALM) method for RPCA.
- Not an exact method…but 250 times faster than Iterative Thresholding
- Approx real-time on 16 bit audio at 16 kHz
- Let's compare/contrast with

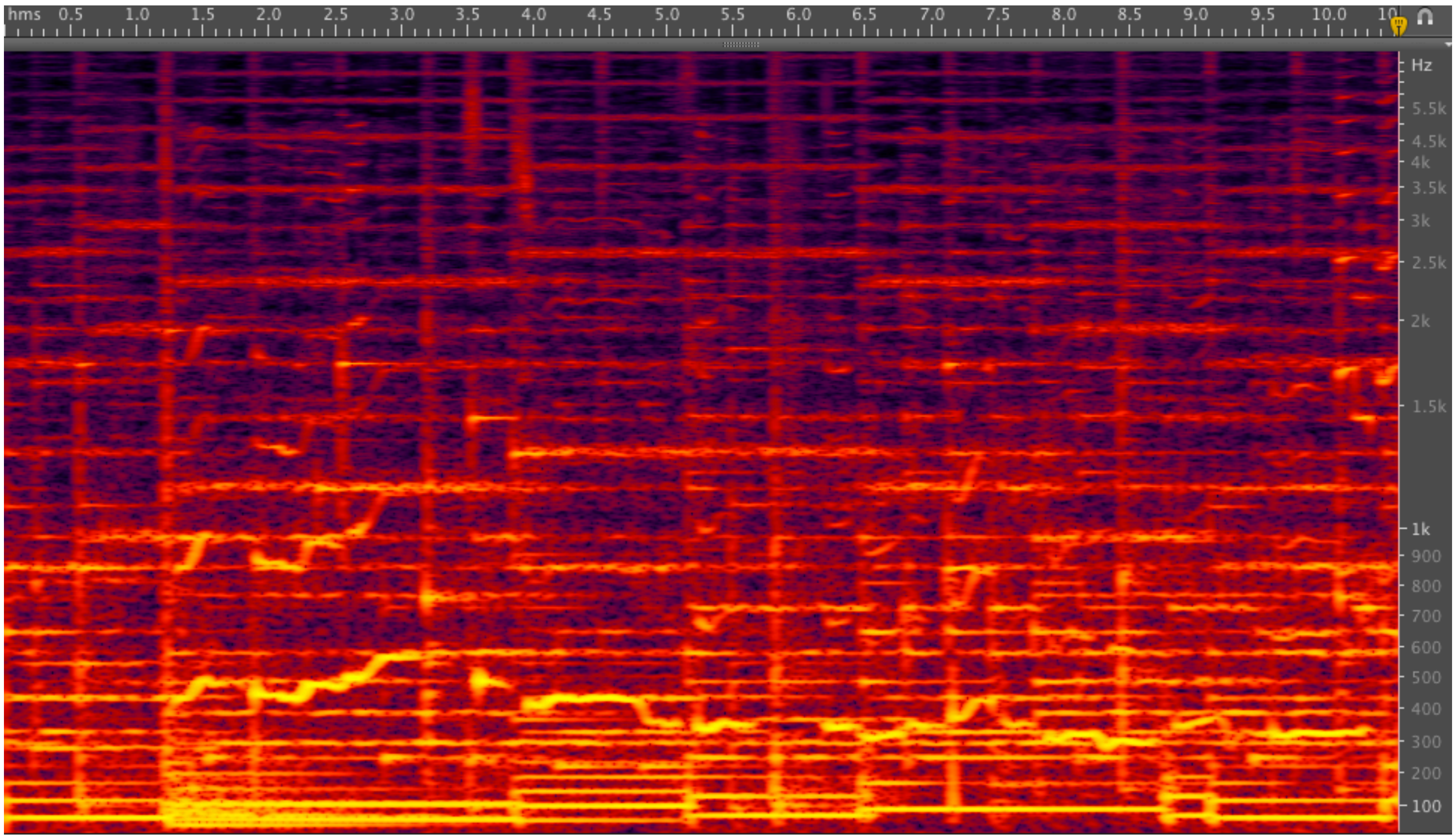
  A periodic method (REPET)

  A Similarity Matrix method (REPET SIM)

Example 1: Singer + Synthesizer

Background: (horizontal lines): low rank, aperiodic, not sparse

Foreground: (squiggly lines) : sparse, aperiodic, not low rank, broadly distributed



10.5 seconds
16 bits
16 kHz

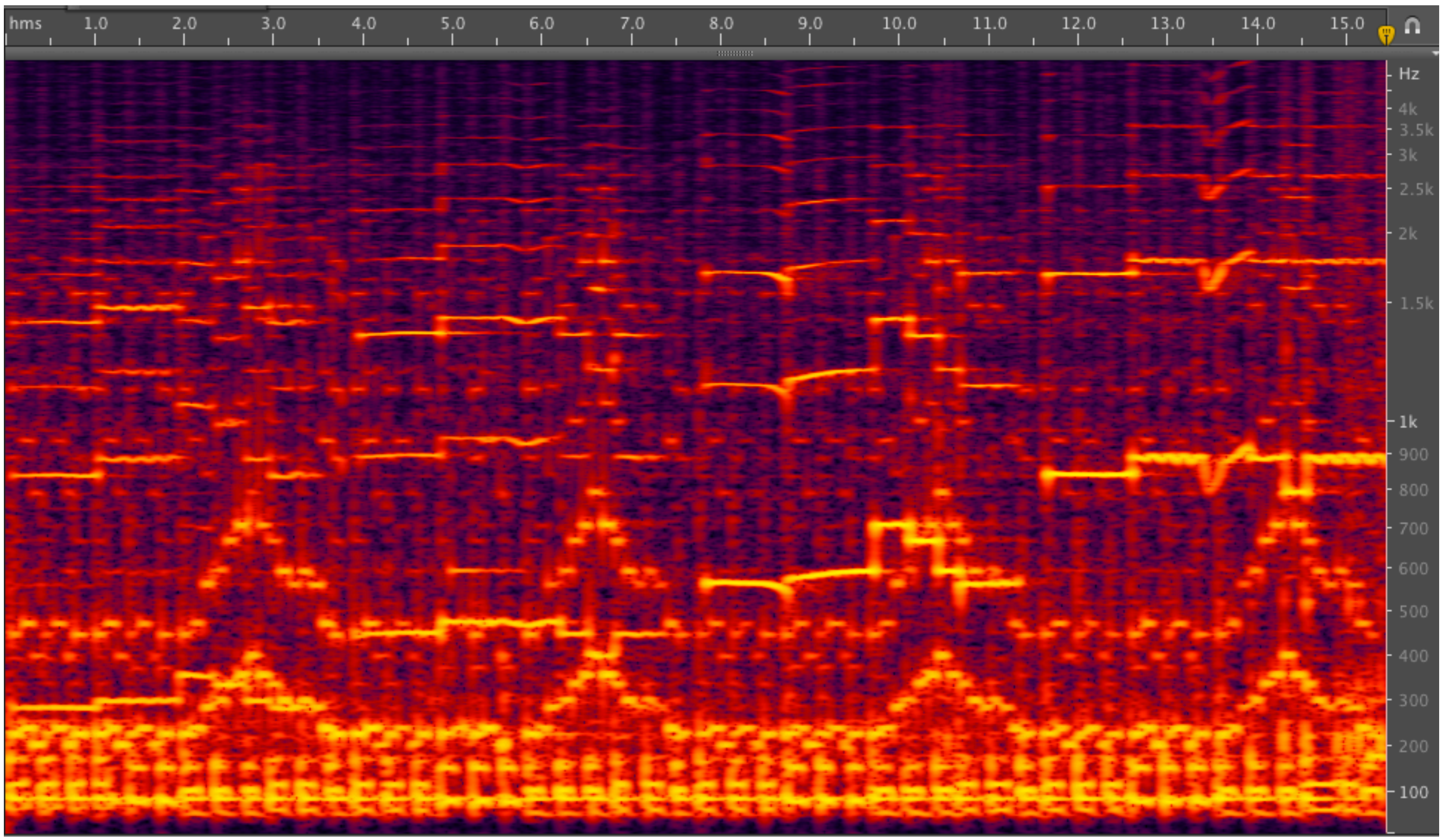
Processing Time in sec.

RPCA	13.9
REPET SIM	1.0
REPET	0.8

Example 2: Clarinet + Guitar/Bass/Snare

Background: (short, horiz. lines making triangles): low rank, sparse, periodic

Foreground: (long horiz. lines): sparse, not periodic, not low rank, broadly distributed



15.4 seconds
16 bits
11.025 kHz

Processing Time in sec.

RPCA	14.6
REPET SIM	1.7
REPET	1.4

When to use…

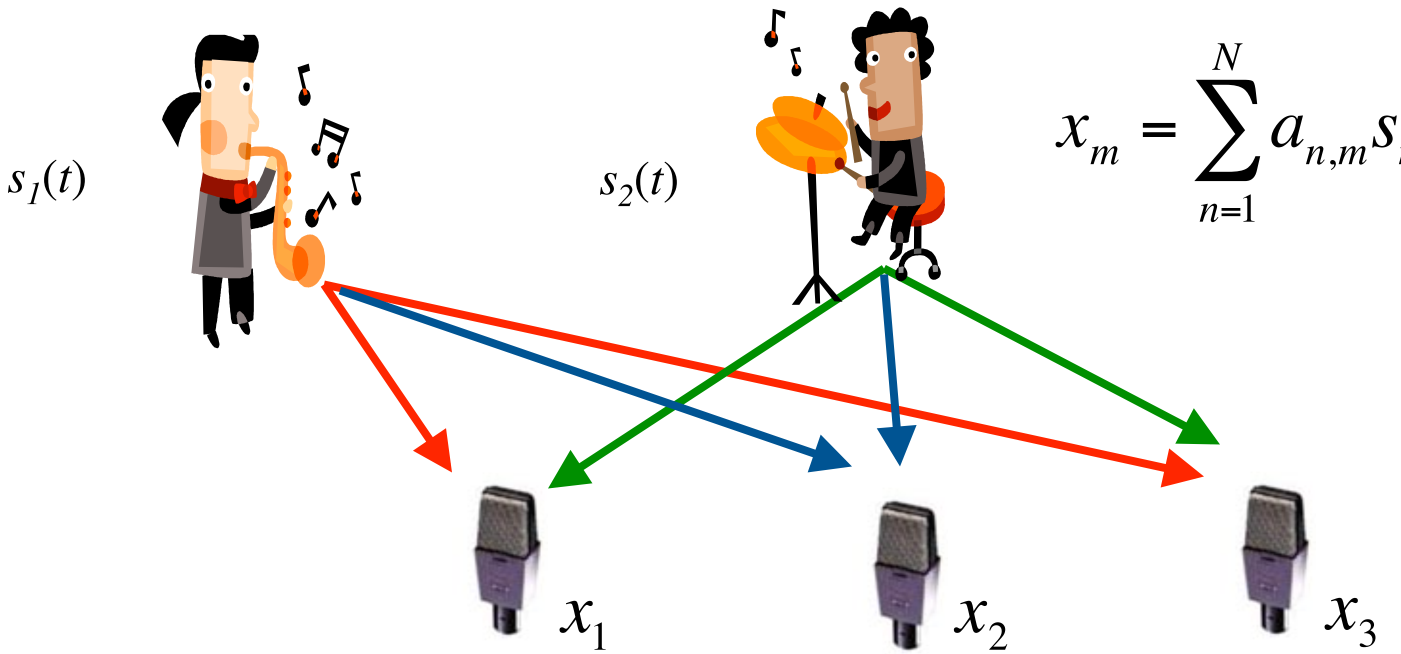
		REPET	REPET-SIM	RPCA
FOREGROUND	periodic	never	don’t care	don’t care
	low rank	don’t care	never	never
	sparse	helps	helps	required
	broadly distributed	don’t care	don’t care	required
BACKGROUND	periodic	required	don’t care	don’t care
	low rank	implied by periodic	required	required
	sparse	don’t care	don’t care	never
	broadly distributed	don’t care	don’t care	helps

Repetition to Augment Separation

- Repetition is a powerful cue for source separation
- It works in isolation (e.g. REPET)
- How can we leverage repetition to improve other approaches to source separation?

Independent Component Analysis (ICA)

- Assumes statistically independent sources
- Number of mixtures cannot be less than the number of sources



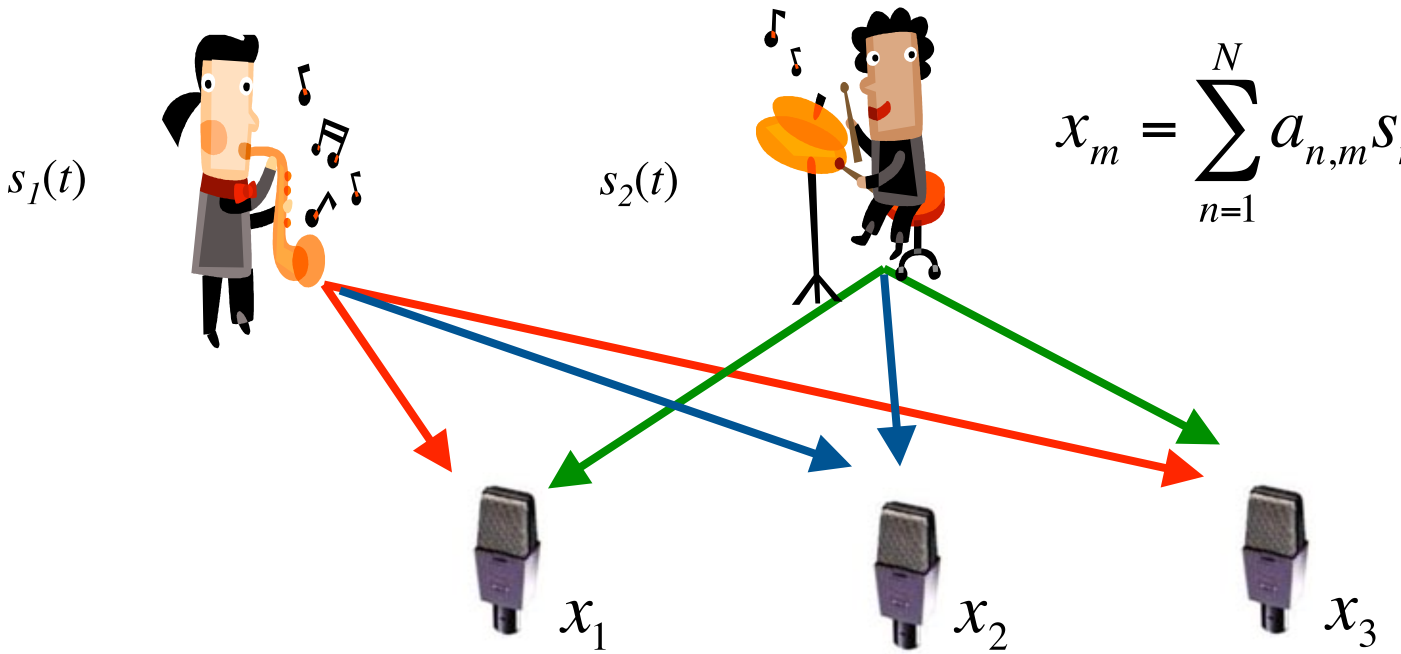
$$x_m = \sum_{n=1}^{N} a_{n,m} s_{n,m}(t - \delta_{n,m})$$

Independent Component Analysis (ICA)

- Probably not how people do it

  People have 2 ears. Scenes often have >2 sources.

- Not useful when there aren't enough mics



$$x_m = \sum_{n=1}^{N} a_{n,m} s_{n,m}(t - \delta_{n,m})$$

ICA and Repetition

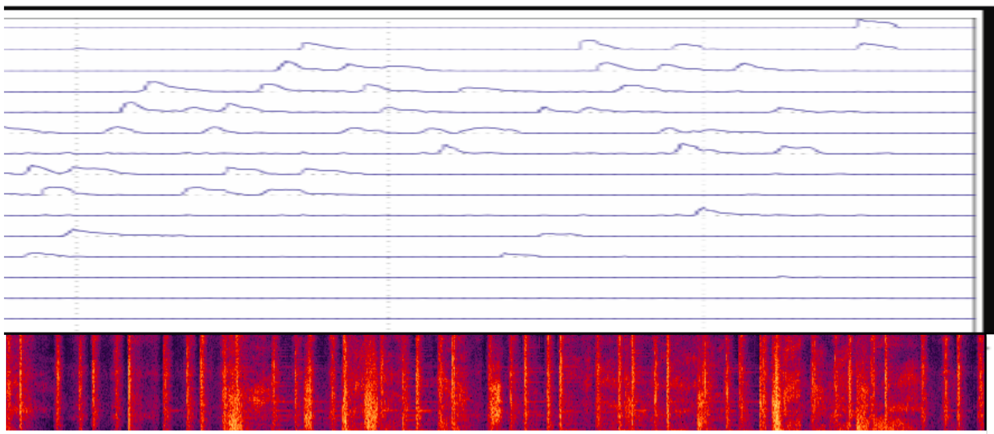
- Hedayiogl et al (ICASSP 2011) found a way to leverage repetition for single-mixture ICA
  1. Assume periodically repeating sources
     (e.g. heart beat patterns)
  2. Record the audio with a single microphone
  3. Segment the audio at period of repetition
  4. Call each segment a channel
  5. Do ICA, just like usual

Nonnegative Matrix Factorization (NMF)

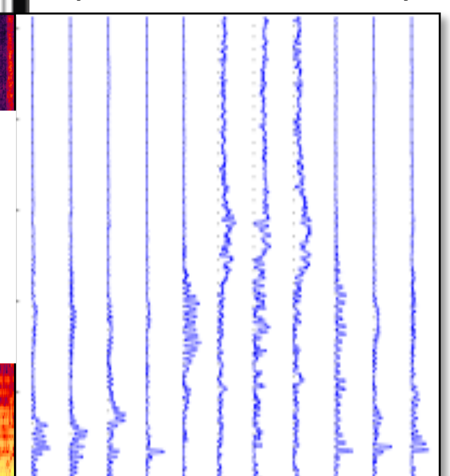
…and its probabilistic reframing, known as
Probabilistic Latent Component Analysis (PLCA)

Just find WH = X and we're done

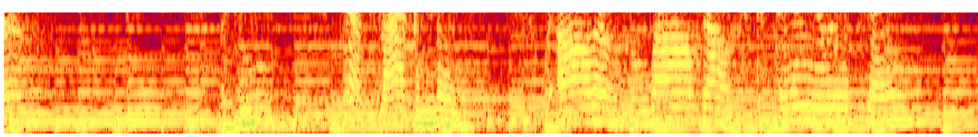
Activation matrix H



Spectral Dictionary W



CATCH: Without special care (setting priors, picking good examples) dictionary elements often represent parts of sources and/or mixes of sources. IE we didn't actually do source isolation.



Magnitude Spectrogram X

NMF & REPET

- Both assume a lower-rank encoding of (some of the) data is possible

- NMF/PLCA assume a fixed size spectral dictionary prior to processing

  Picking a good dictionary size is a black art

- REPET's "dictionary" size depends on the period of the audio

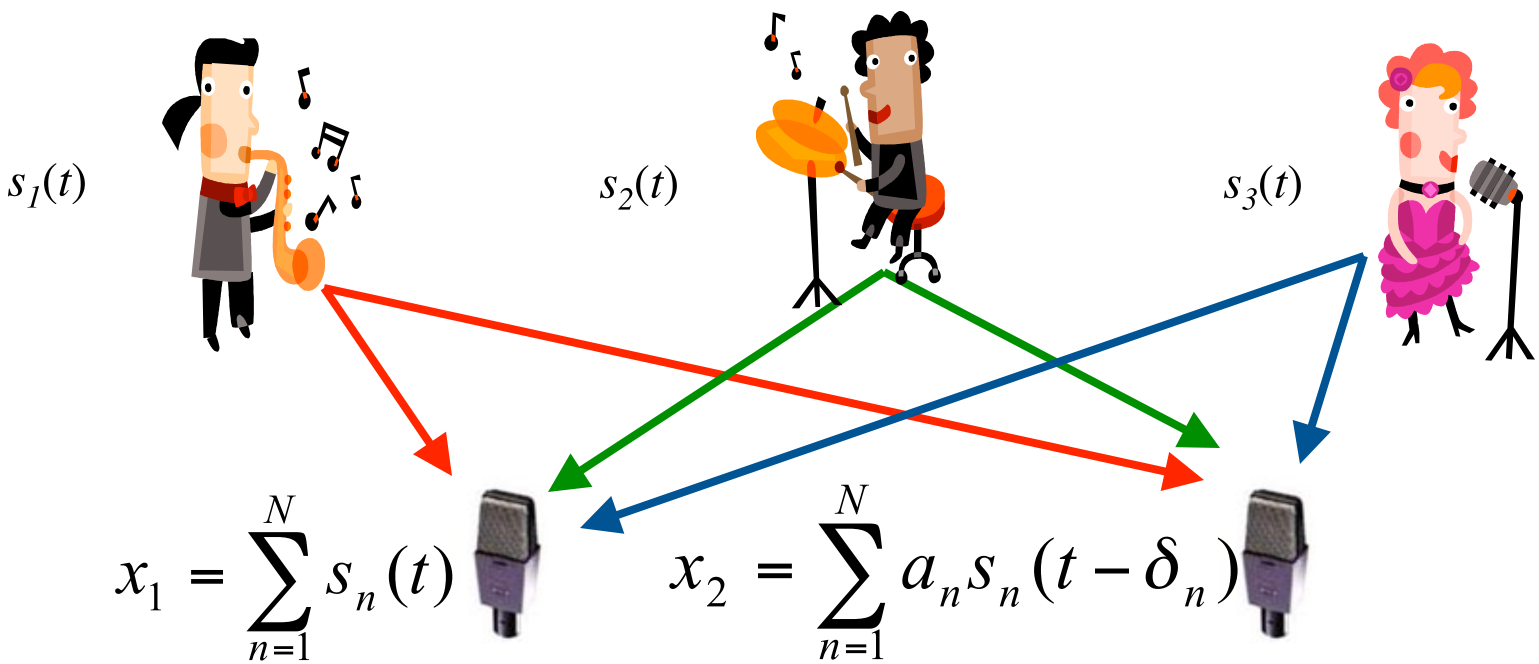
Improving NMF with Repetition

- Could we find a good dictionary size for NMF by finding the period of repetition prior to processing?

- Could we seed the dictionary for NMF with the repeating spectrum segment calculated by REPET?

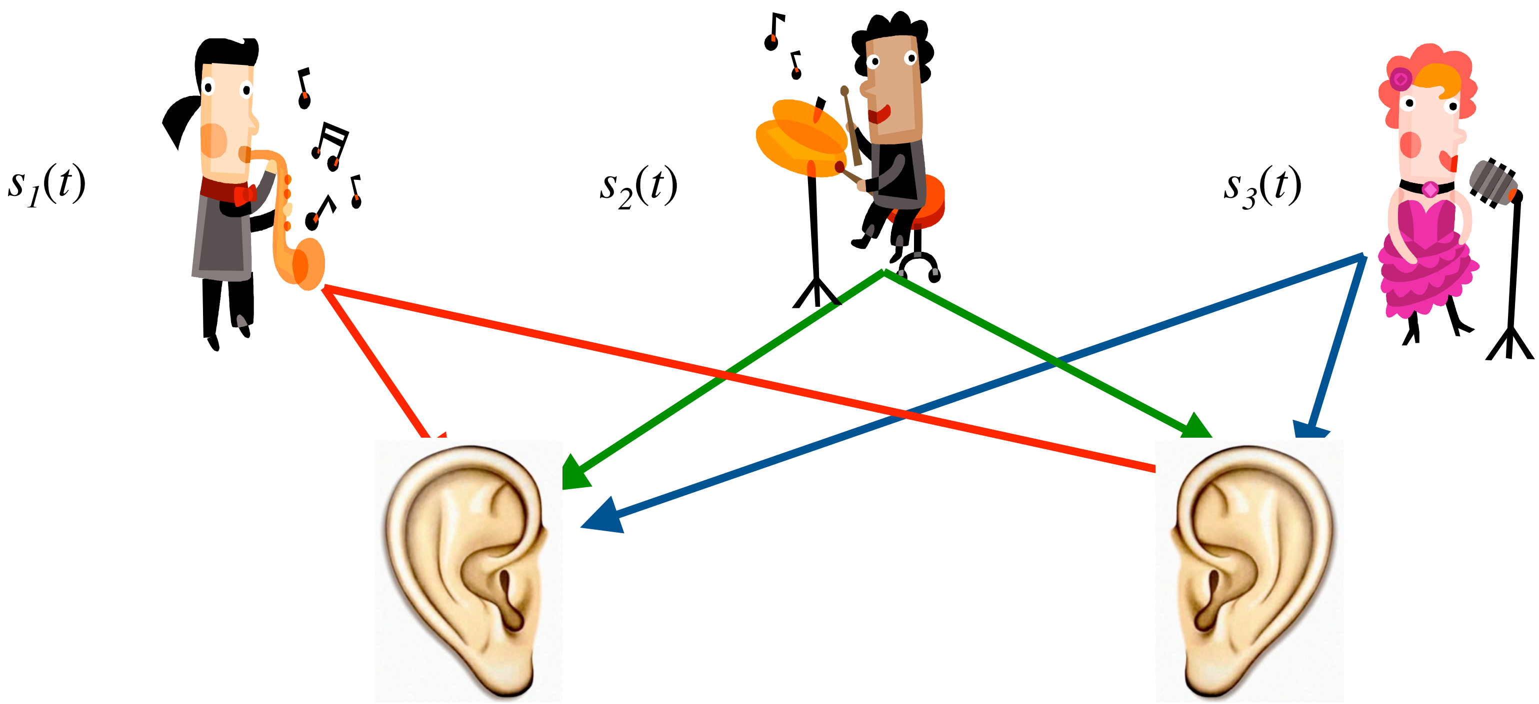
Using Spatial Cues: DUET

- Each source location has a unique cross-channel amplitude scaling $\boldsymbol{a_n}$ and time-shift $\boldsymbol{\delta_n}$
- Find those and you can separate your sources with a mask (e.g. DUET)



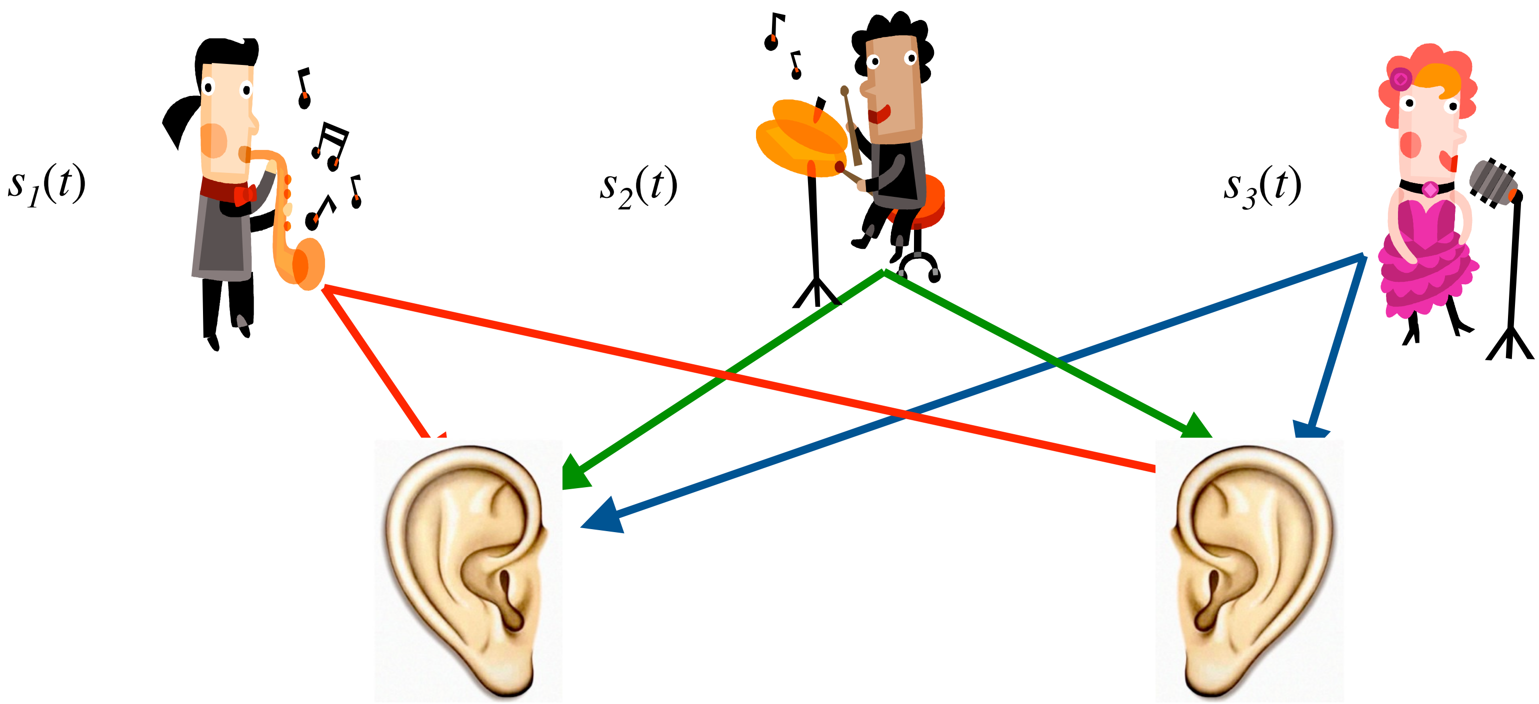
Approach: Using Spatial Cues

- Translation: Sound **closer** to the left ear hits it **sooner** and **louder**. Use that.



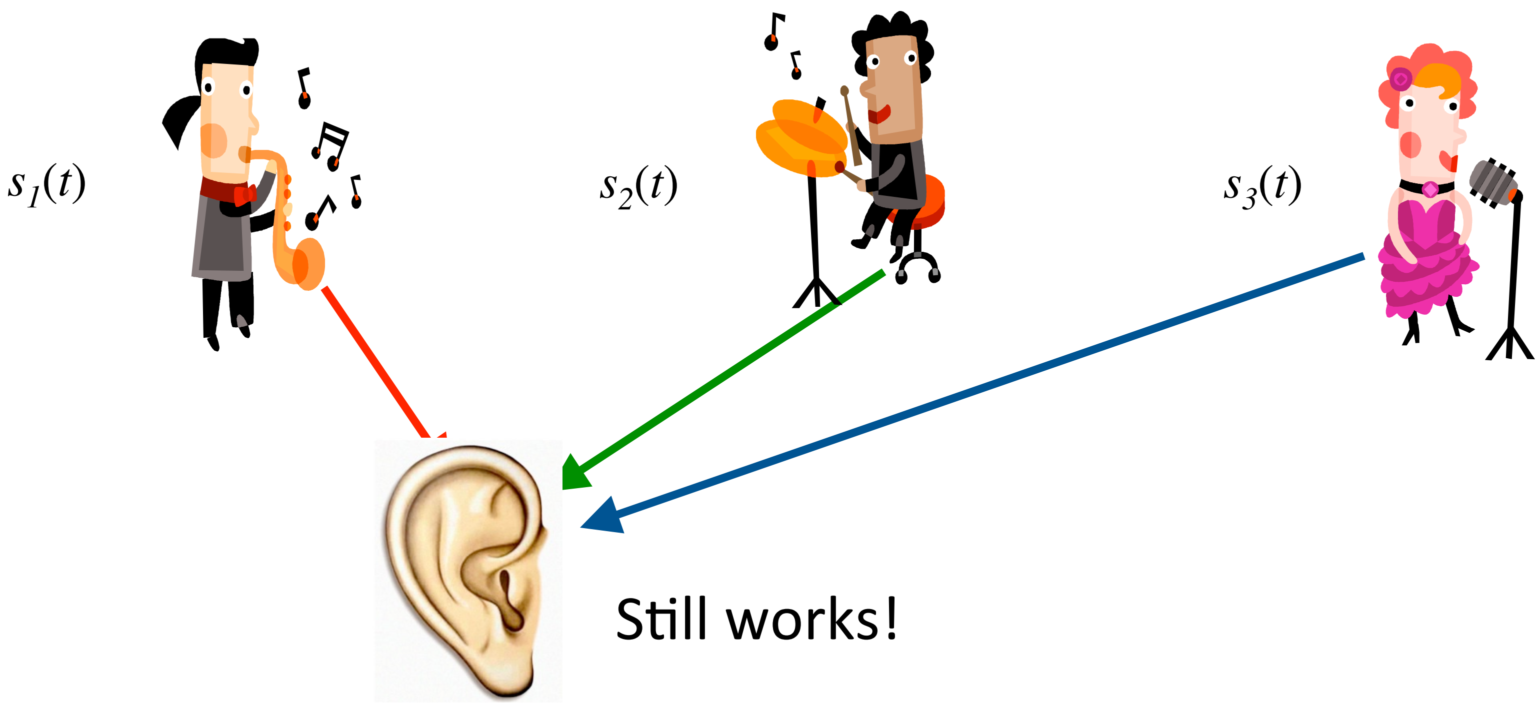
Approach: Using Spatial Cues

- Can have more sources than microphones
- Assumes sources don't move
- Has great difficulty with reverberation



Approach: Using Spatial Cues

- Can have more sources than microphones
- Assumes sources don't move
- Has great difficulty with reverberation
- People don't need 2 ears to follow sounds in a mix



Repetition and Duet

- Could the same game played with ICA be done with DUET?
  - Move a single microphone around
  - Align the recordings at the period of repetition
  - Run DUET
- Could we combine DUET and REPET to overcome reverberation issues?

If you like any of those ideas…

….maybe you'd like to collaborate?

Our contact information is at the start of these slides.

Outline

I. Introduction
II. How humans use repetition to identify sound sources (McDermott)
III. Coffee break
IV. Repetition-based algorithms for source separation (Rafii)
V. Links to other methods for source separation
VI. Conclusions/Questions

Conclusions

- Repetition is a fundamental element in generating and perceiving structure in audio
- Repeating structure can be used to effectively segment audio scenes
- Algorithms based on repetition are related to those seeking low-rank decompositions
- The assumptions they make are different than existing approaches
- Therefore, they complement existing approaches

Getting Source Code

- REPET

http://music.cs.northwestern.edu/research.php?project=repet

- REPET SIM

http://music.cs.northwestern.edu/research.php?project=repet

- RPCA

https://sites.google.com/site/singingvoiceseparationrpca/

Questions/Discussion

?

References

- Emmanuel J. Candes, Xiaodong Li, Yi Ma, and John Wright,"Robust principal component analysis?," *J. ACM*, vol. 58, pp.11:1–11:37, Jun. 2011.
- J.-L. Durrieu, B. David, and G. Richard, "A Musically Motivated Mid-level Representation for Pitch Estimation and Musical Audio Source Separation," *IEEE Journal on Selected Topics on Signal Processing*, vol. 5, no. 6, pp. 1180-1191, October 2011.
- D. FitzGerald and M. Gainza, "Single Channel Vocal Separation using Median Filtering and Factorisation Techniques," *ISAST Transactions on Electronic and Signal Processing*, vol. 4, no. 1, pp. 62-73, 2010.
- J. Foote, "Visualizing Music and Audio using Self-Similarity," *ACM International Conference on Multimedia*, Orlando, FL, USA, October 30-November 5, 1999.
- J. Foote and S. Uchihashi, "The Beat Spectrum: A New Approach to Rhythm Analysis," *IEEE International Conference on Multimedia and Expo*, Tokyo, Japan, August 22-25, 2001.
- F. Hedayioglu, M. Jafari, S. Mattos, M. Plumley and M. Coimbra, "Separating sources from sequentially acquired mixtures of heart signals, " *IEEE International Conference on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, May 22-27, 2011.
- C.-L. Hsu and J.S. R. Jang, "On the Improvement of Singing Voice Separation for Monaural Recordings Using the MIR-1K Dataset," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 310-319, February 2010.
- P. Huang, S. Deeann Chen, P. Smaragdis, M. Hasegawa-Johnson, "SINGING-VOICE SEPARATION FROM MONAURAL RECORDINGS USING ROBUST PRINCIPAL COMPONENT ANALYSIS" , *IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, March 25-30, 2012.
- Z. Lin, M. Chen, L. Wu, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low rank matrices," Tech. Rep. UILU-ENG-09-2215, UIUC, Nov.2009
- A. Liutkus, Z. Rafii, R. Badeau, B. Pardo, and G. Richard, "Adaptive Filtering for Music/Voice Separation exploiting the Repeating Musical Structure," *IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, March 25-30, 2012.
- McDermott, J.H., Wrobleski, D., Oxenham, A.J. (2011) Recovering sound sources from embedded repetition. Proceedings of the National Academy of Sciences, 108, 1188-1193.
- Z. Rafii and B. Pardo, "A Simple Music/Voice Separation Method based on the Extraction of the Repeating Musical Structure," *IEEE International Conference on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, May 22-27, 2011.
- Z. Rafii and B. Pardo, "REpeating Pattern Extraction Technique (REPET): A Simple Method for Music/Voice Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, in press.
- Z. Rafii and B. Pardo, "Music/Voice Separation using the Similarity Matrix," *13th International Society for Music Information Retrieval*, Porto, Portugal, October 8-12, 2012.
- Schwartz, A., McDermott, J.H., Shinn-Cunningham, B. (2012) Spatial cues alone produce inaccurate sound segregation: The effect of interaural time differences. Journal of the Acoustical Society of America, 132, 357-368.
- O. Yimlaz and S. Rikard, "Blind Separation of Speech Mixtures via Time-frequency Masking," IEEE Transactions on Signal Processing, July 2004, Vol. 52(7), 1830-1847