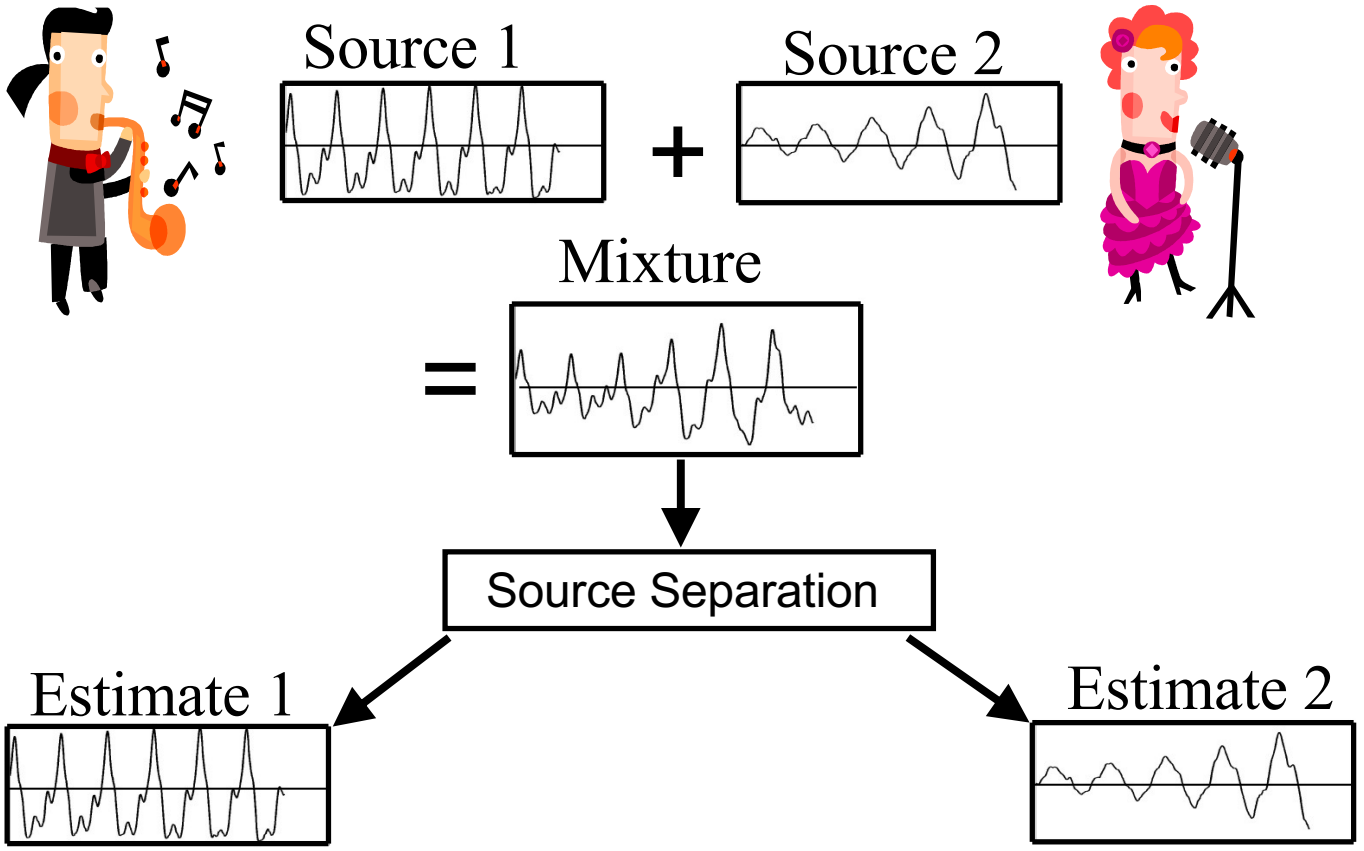
Music/Voice Separation using the Similarity Matrix

Zafar Rafii & Bryan Pardo

NORTHWESTERN
UNIVERSITY

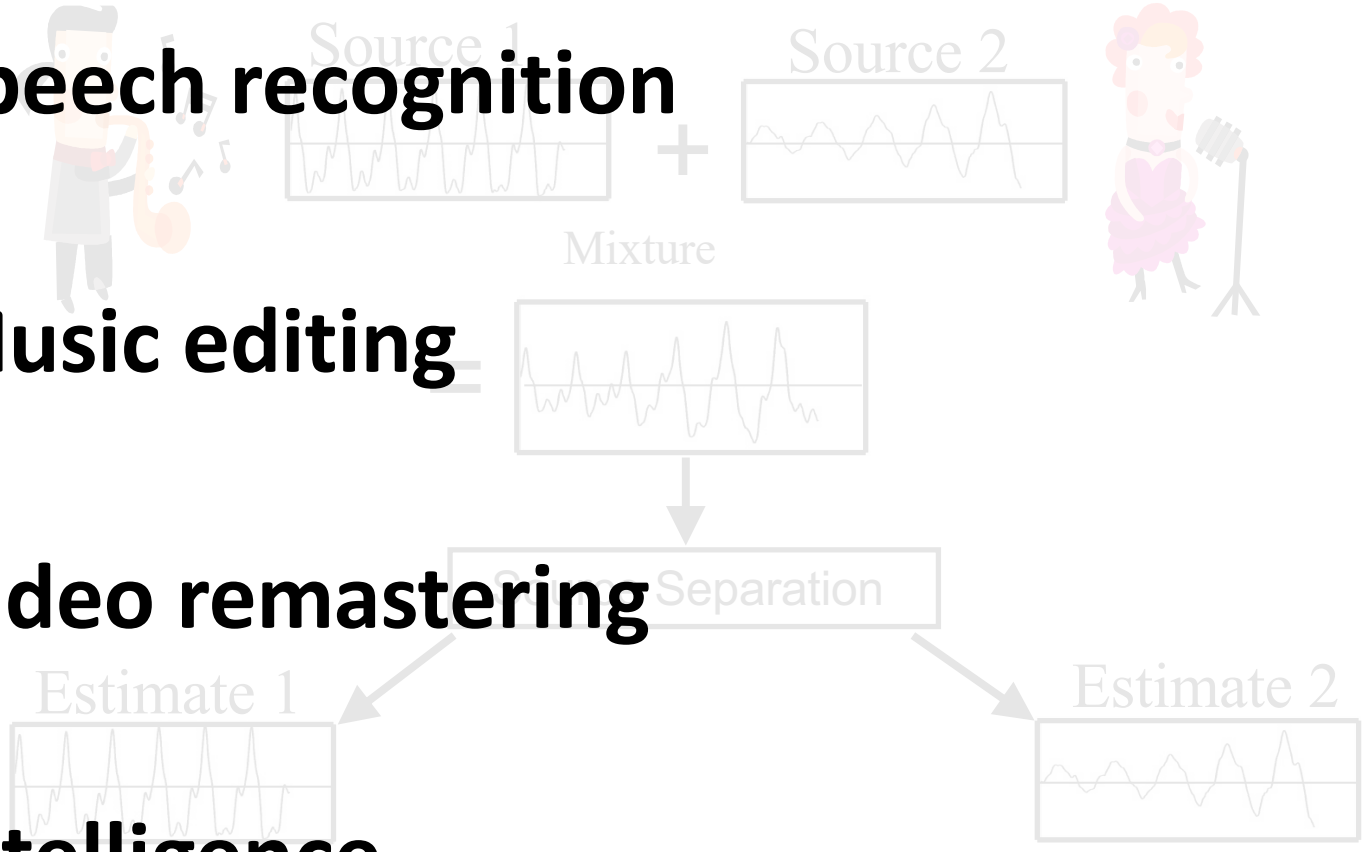
interactive
audio lab

Audio Source Separation

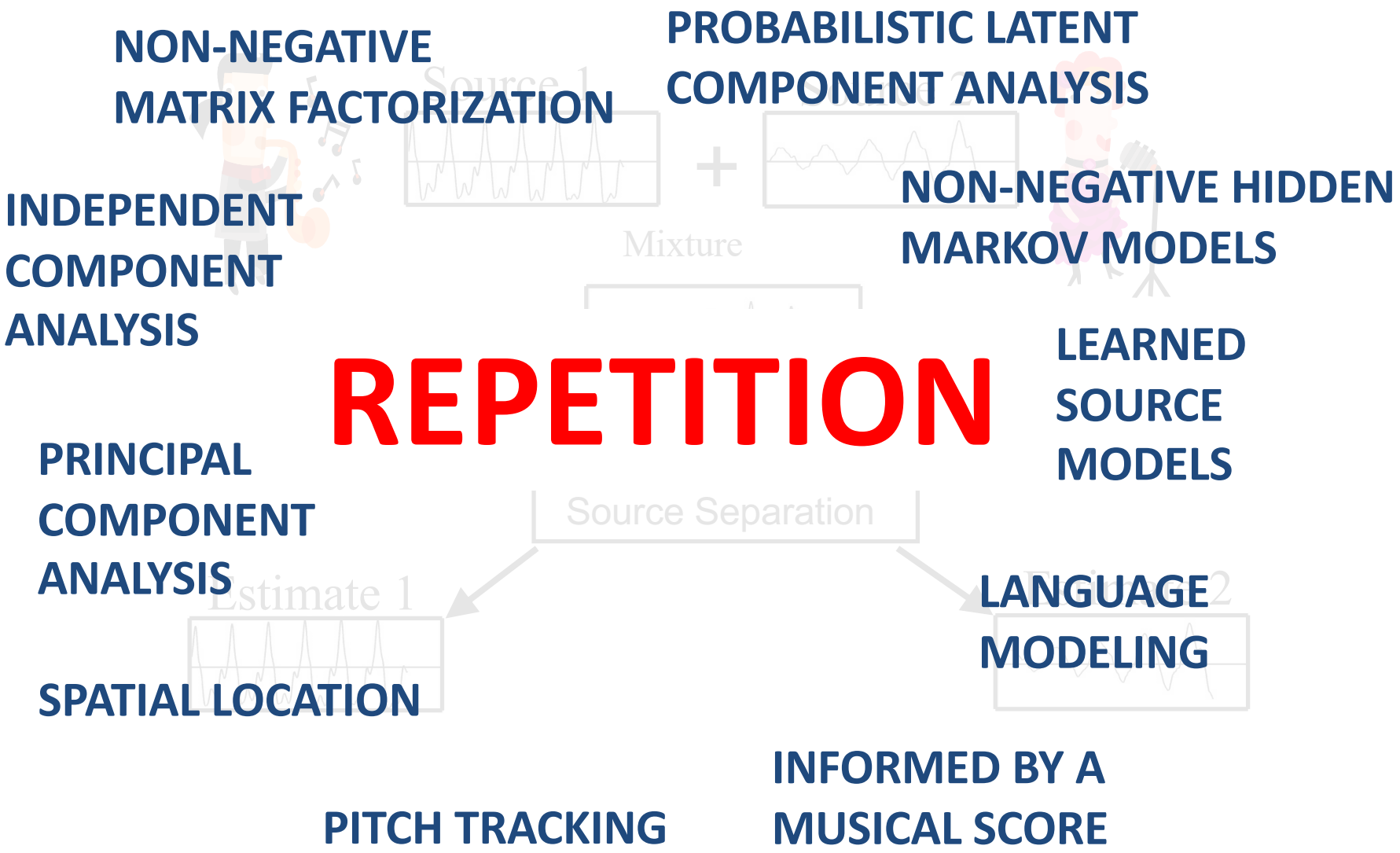


Audio Source Separation

- **Speech recognition**
- **Music editing**
- **Video remastering**
- **Intelligence**



Audio Source Separation



What should you get out of this?

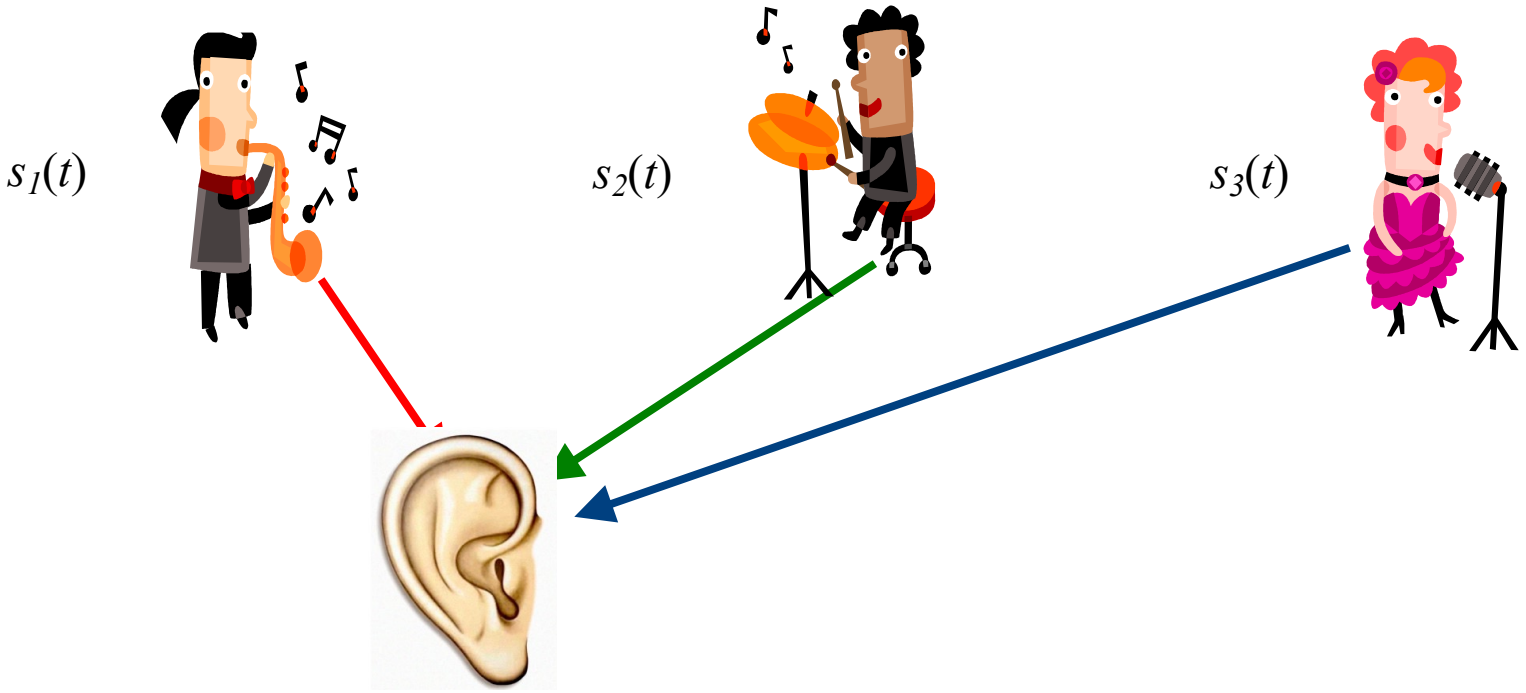
- Understand the psychological basis for the application of repetition to audio source separation
- Understand a new class of practical algorithms that perform repetition-based source separation

Working on a Single Mixture

Mix = Sound1 + Sound2 + Sound 3

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

Infinite number of solutions!



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

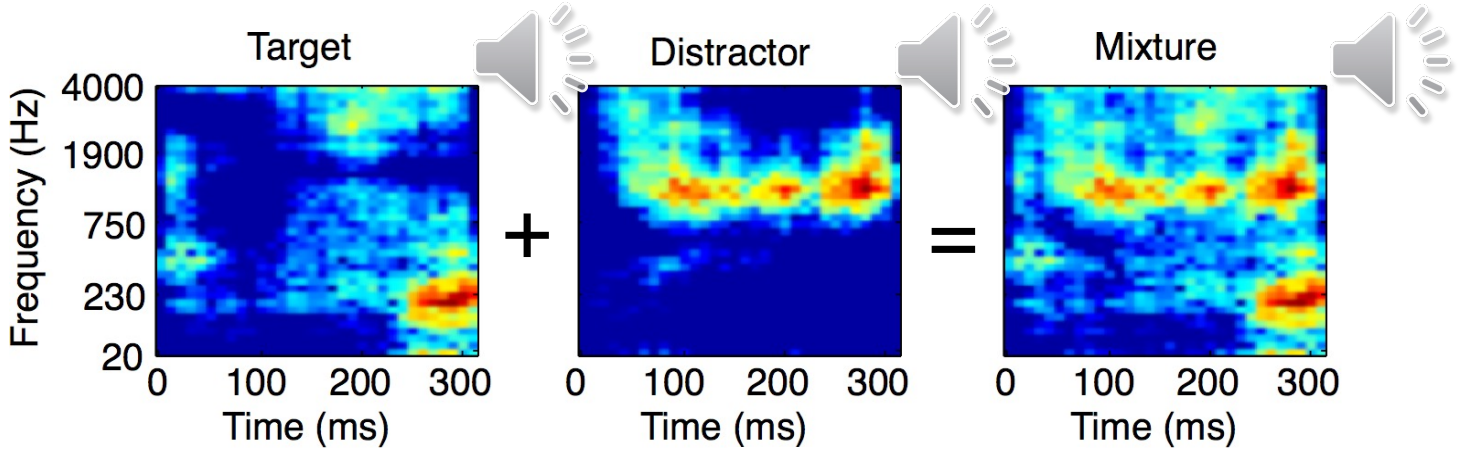
Repetition in Psychoacoustics

Q. What evidence is there that humans use repetition to parse an auditory scene?

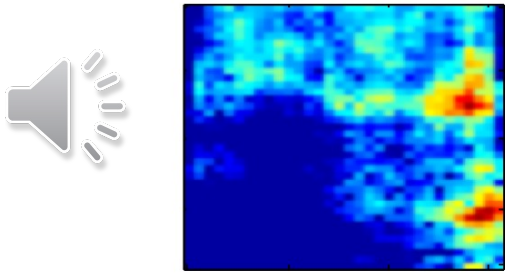
A. McDermott, J.H., Wrobleski, D. & Oxenham, A.J. (2011) Recovering sound sources from embedded repetition. *Proceedings of the National Academy of Sciences*, 108

The Experiment

- Synthesize discriminable sounds with no grouping cues
- Present one (or more) two-sound mixtures

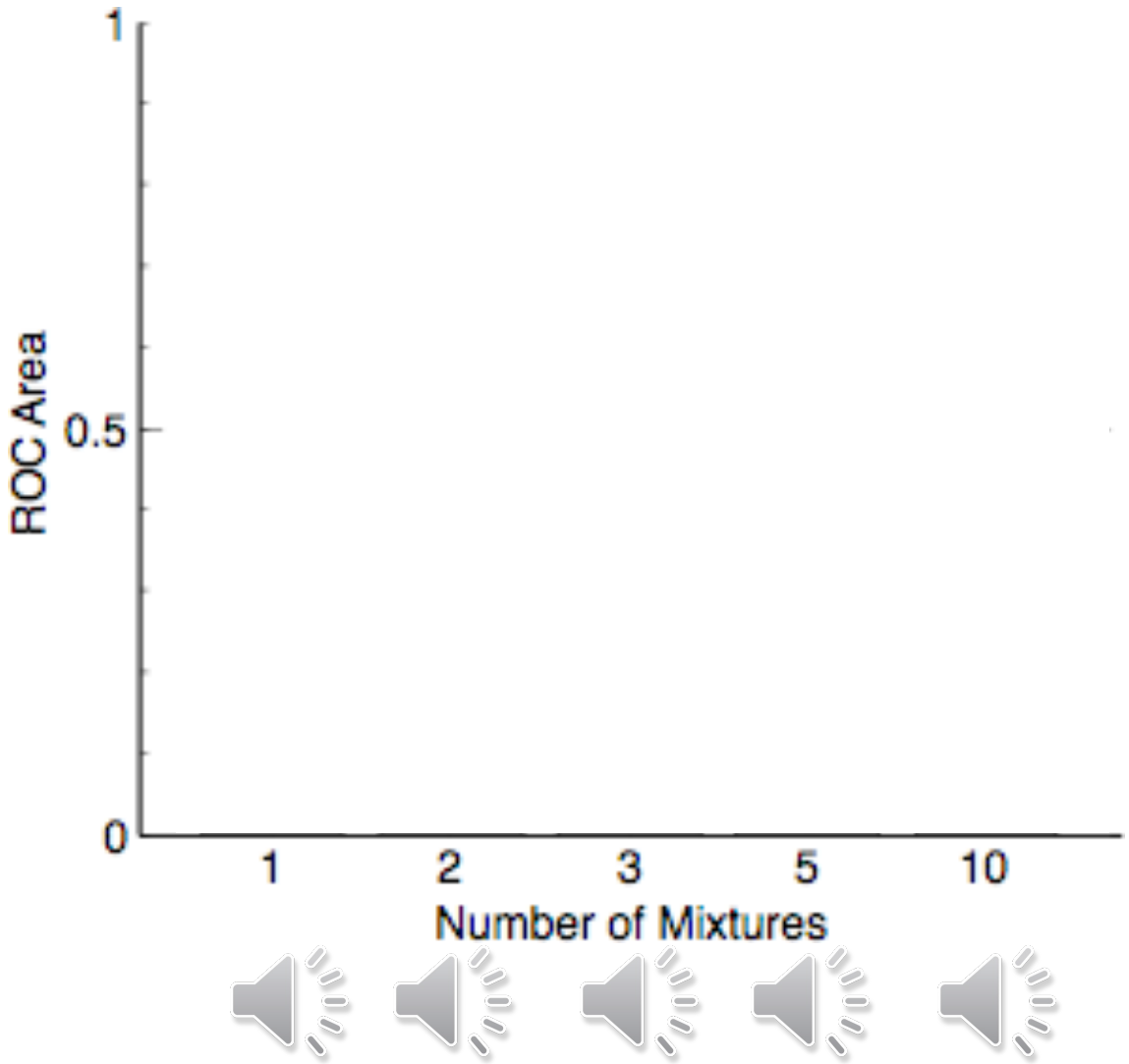


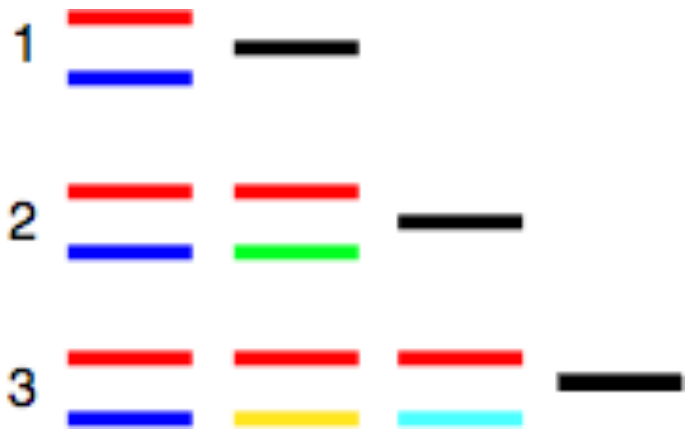
- Present a probe sound:



- Ask: Was the probe sounds in the mixture(s)?

The Results





McDermott et all showed..

- Repetition in varied contexts is sufficient to let us learn and discriminate new sounds.

- No other grouping cues needed
  - Harmonicity
  - Common event onset-offset
  - Prior-learned “sound templates”
  - Directionality

Repetition in Source Separation

Q. Can we build source separation algorithms based only on repetition cues?

A. REpeating Pattern Extraction Technique (REPET)

Z. Rafii and B. Pardo, "A Simple Music/Voice Separation Method Based on the Extraction of the Repeating Musical Structure," ICASSP 2011

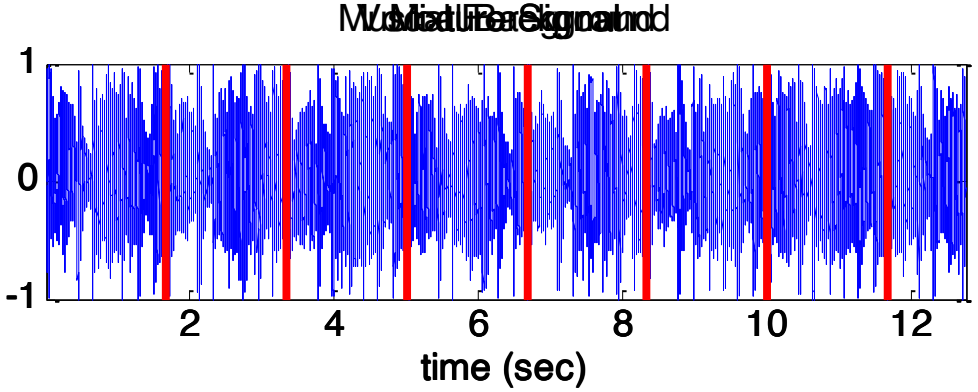
A. Liutkus, Z. Rafii, R. Badeau, B. Pardo and G. Richard, "Adaptive Filtering for Music/Voice Separation Exploiting the Repeating Musical Structure," ICASSP 2012

Z. Rafii and B. Pardo, "Music/Voice Separation Using the Similarity Matrix," ISMIR 2012

Z. Rafii and B. Pardo, "REpeating Pattern Extraction Technique (REPET): A Simple Method for Music/Voice Separation ," TASLP 2013

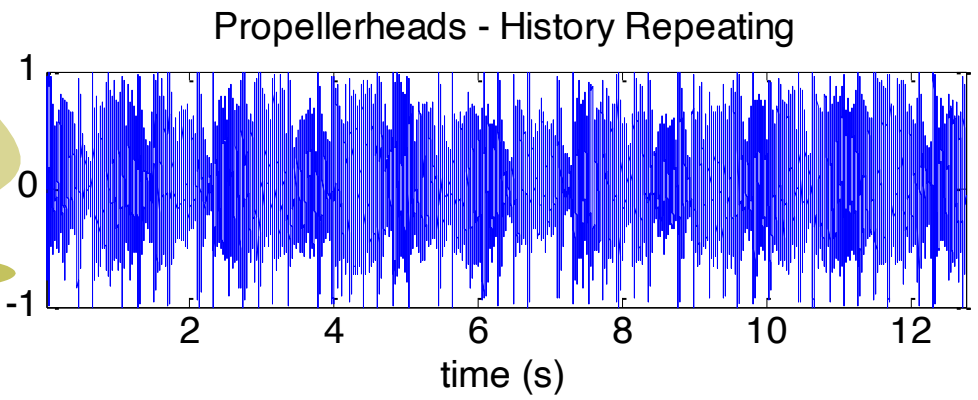
The BIG IDEA

1. Identify the repeating elements in the sound
2. Split sound into repeating and non-repeating parts



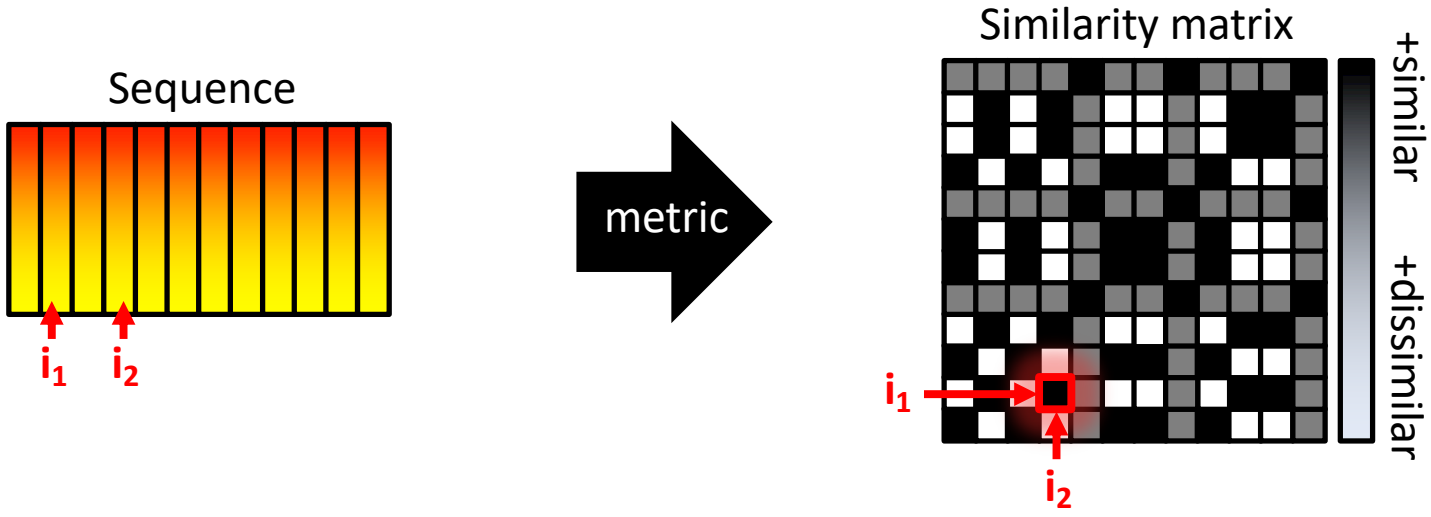
Introduction

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



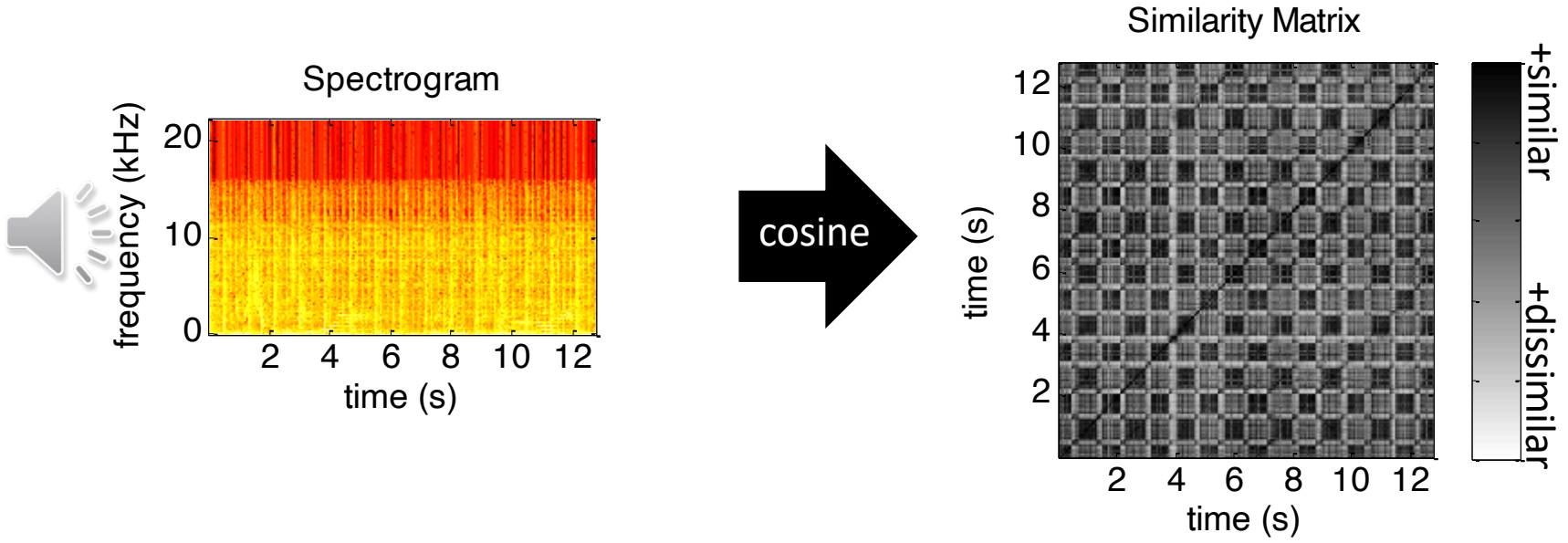
Similarity Matrix

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



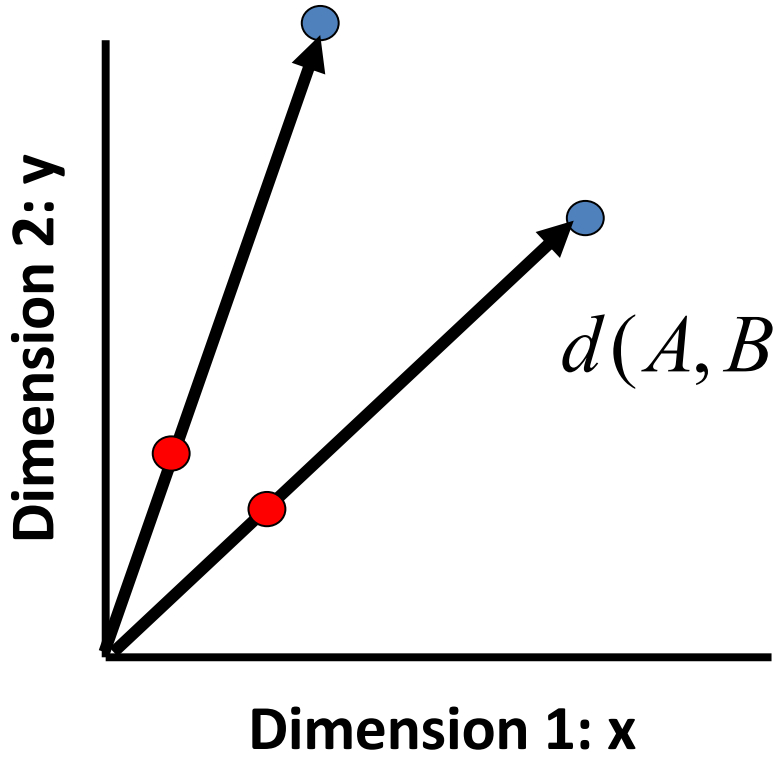
Similarity Matrix

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



How do we measure similarity?

- Euclidian distance
  - what people intuitively think of as "distance"



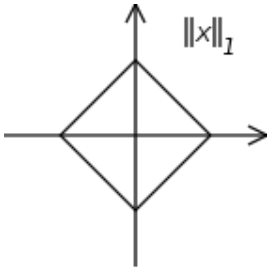
$$d(A,B) = \sqrt{(a_x - b_x)^2 + (a_y - b_y)^2}$$

L^p norms

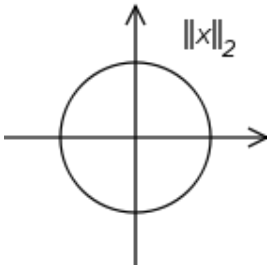
- L^p norms are all special cases of this function:

$$d(\vec{x},\vec{y}) = \left[\sum_{i=1}^{n} | x_i - y_i |^p\right]^{1/p}$$

p changes the norm



L^1 norms = Manhattan Distance: *p=1*



L^2 norms = Euclidean Distance: *p=2*

Cosine Similarity

- Measure of similarity between two vectors
  - Range from -1 (opposite) to 1 (same)
  - Cosine distance = 1 – cosine similarity

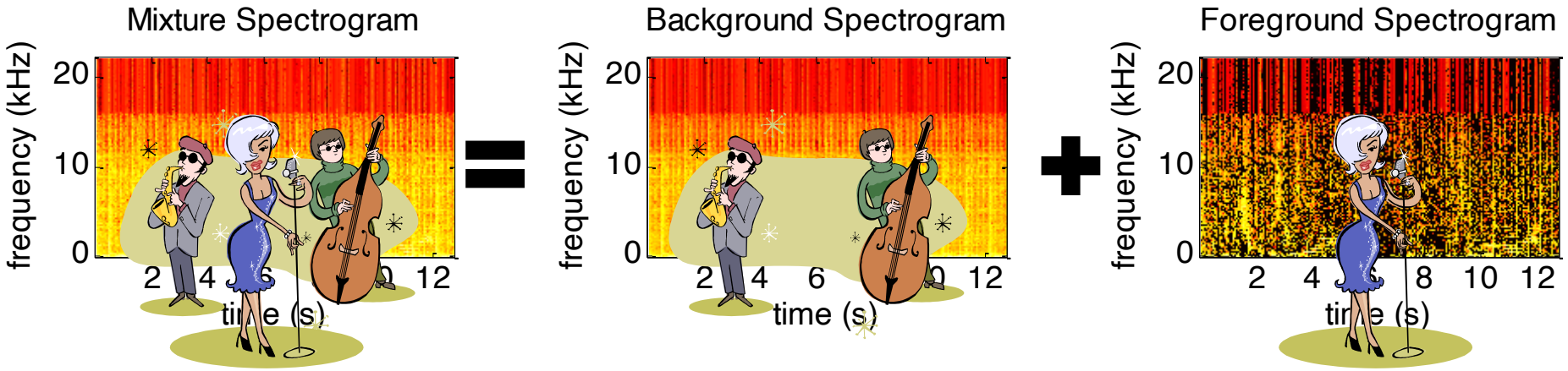
- Cosine similarity between vector *A* and *B:*

$$sim(A, B) = \frac{A \cdot B}{\|A\|\|B\|}$$

$$A \cdot B = \sum_{i=1}^{n} A_i B_i \qquad \|A\|\|B\| = \sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}$$

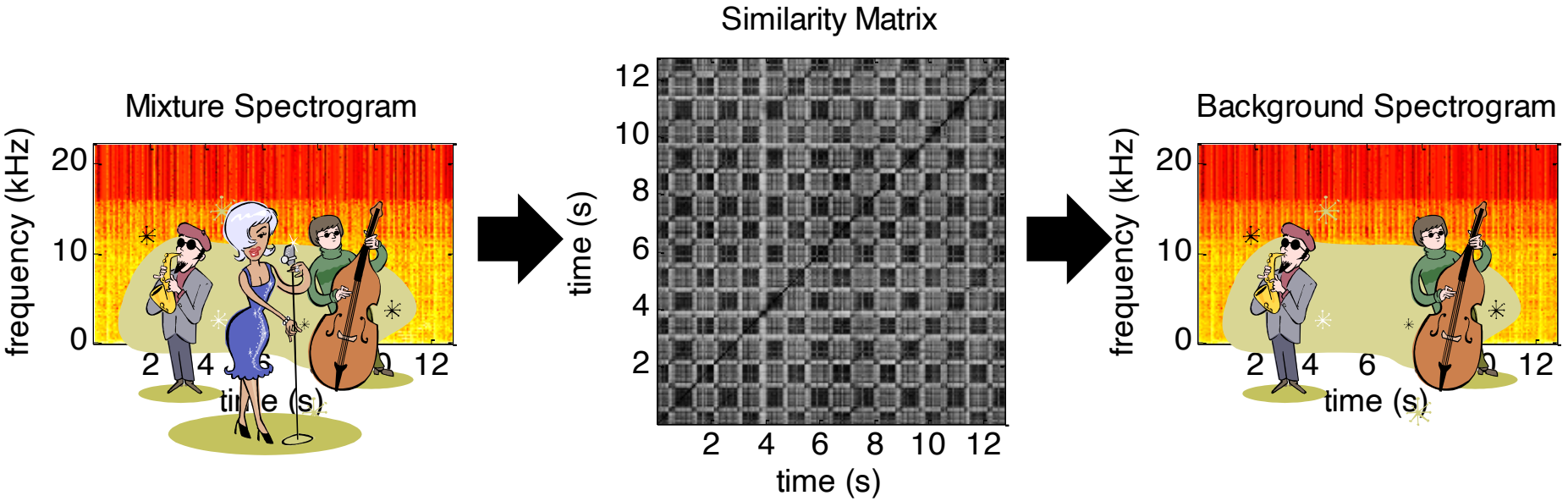
Assumptions

- Given a mixture of music + voice:
  - The repeating background is **dense & low-ranked**
  - The non-repeating foreground is **sparse & varied**



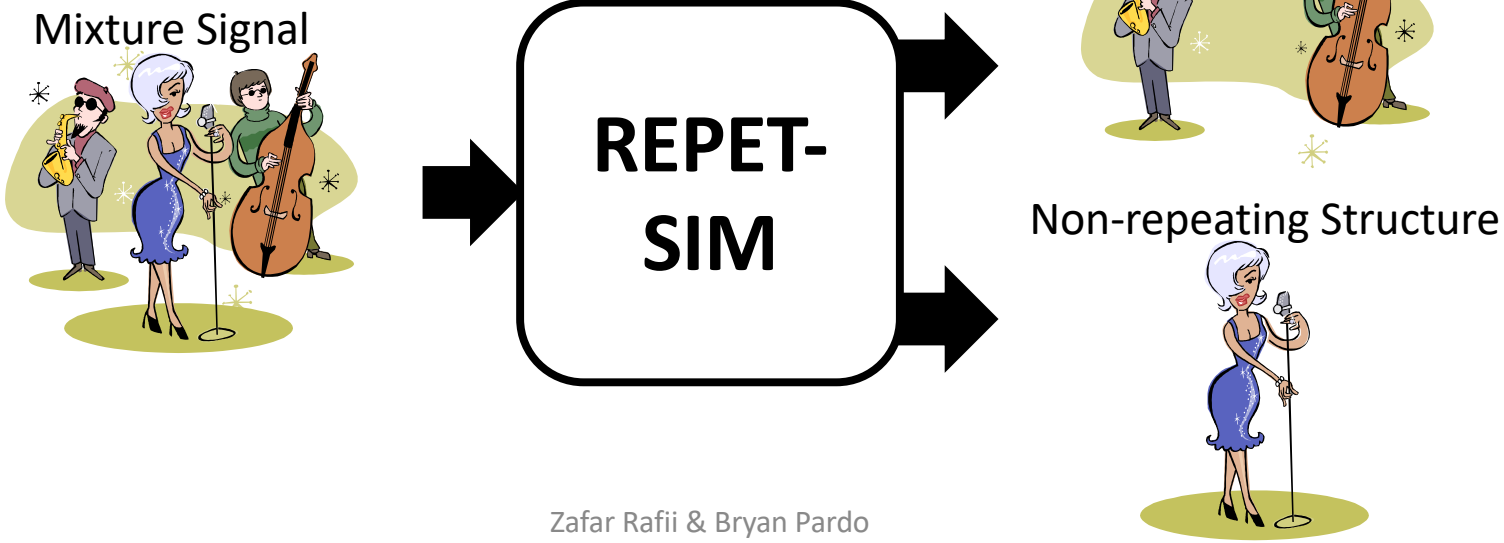
Assumptions

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

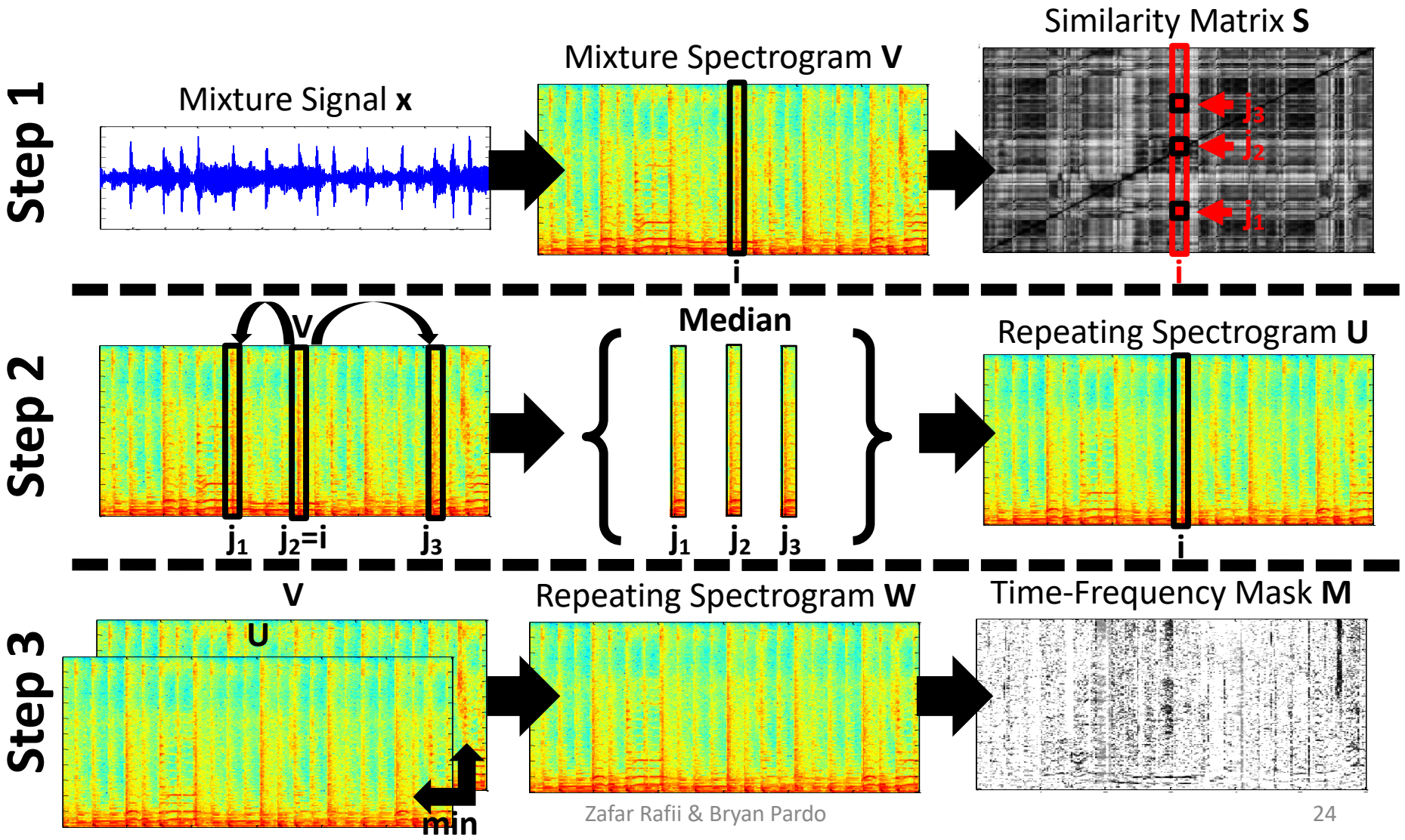


REPET-SIM

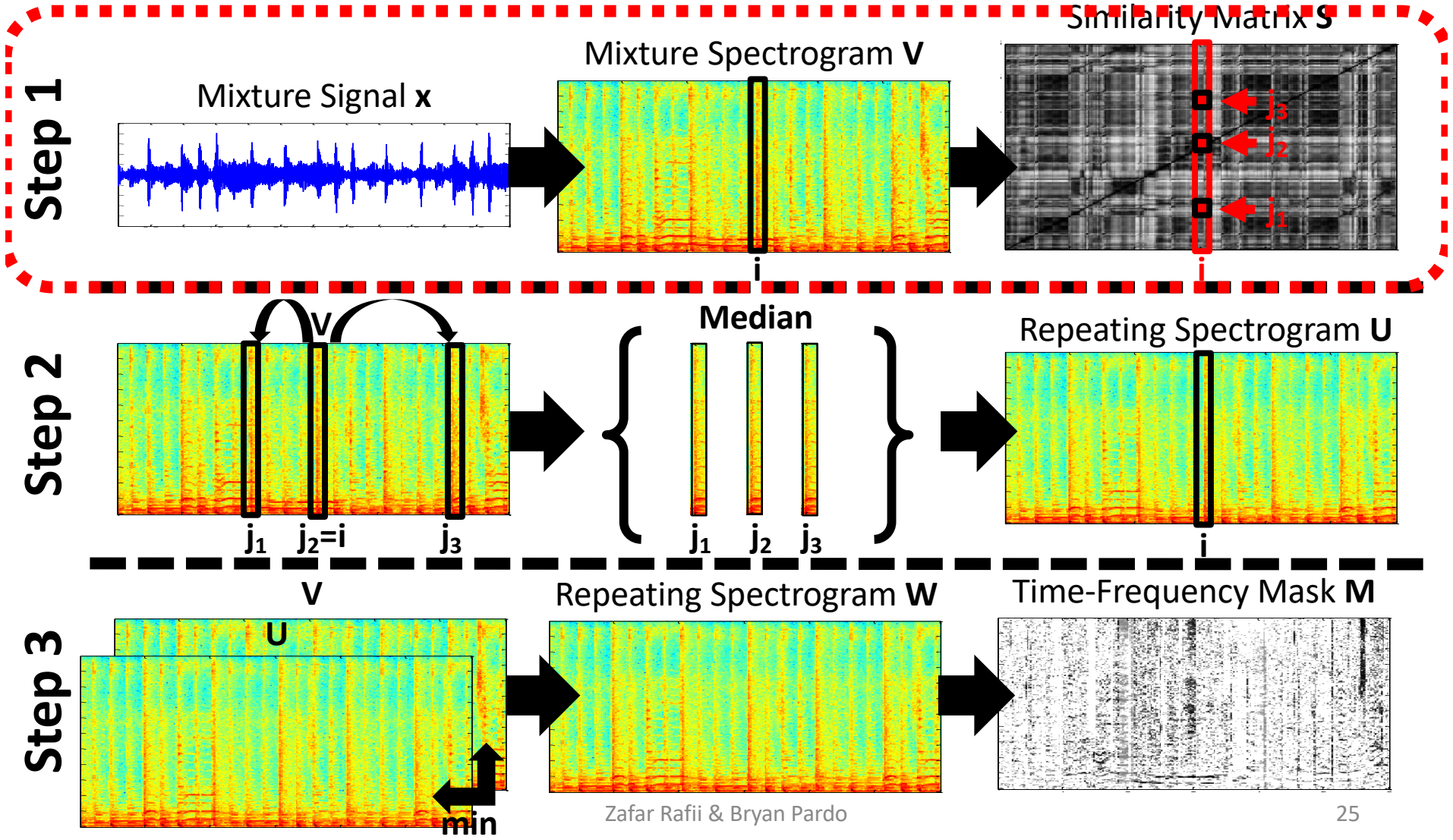
- **REPET with Similarity Matrix!**
  1. Identify the repeating/similar elements
  2. Derive a repeating model
  3. Extract the repeating structure



REPET-SIM

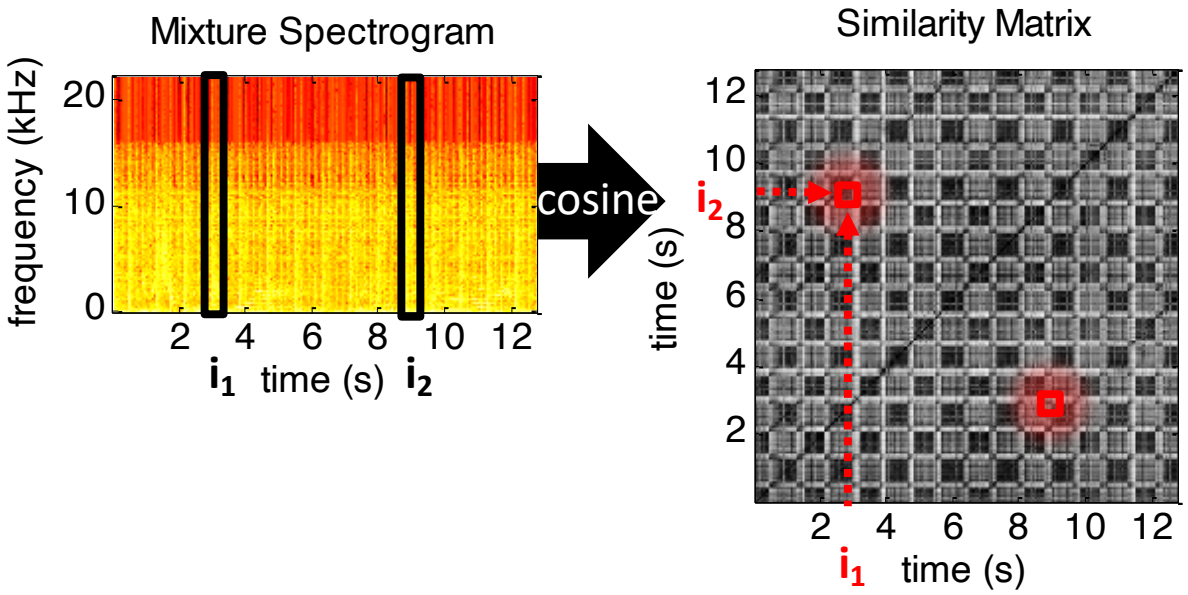

Step 1
Mixture Signal x
Mixture Spectrogram V
Similarity Matrix S
j3
j2
j1
i
i
Step 2
V
Median
Repeating Spectrogram U
j1
j2=i
j3
j1
j2
j3
i
Step 3
V
U
Repeating Spectrogram W
Time-Frequency Mask M
min

1. Repeating Elements



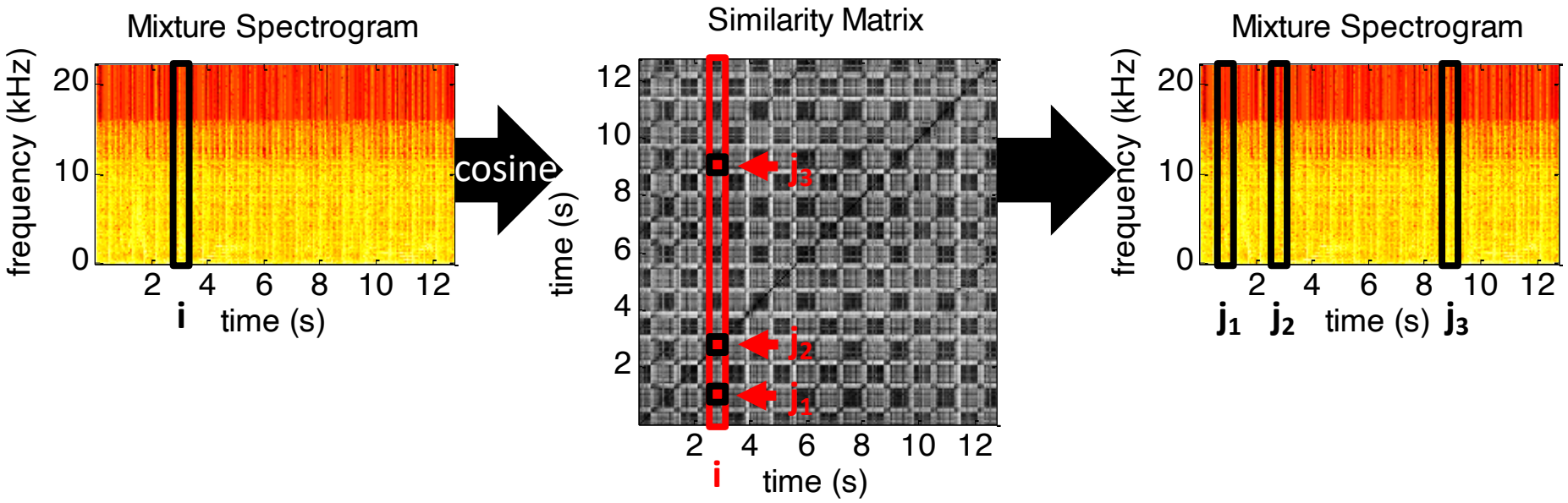
1. Repeating Elements

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

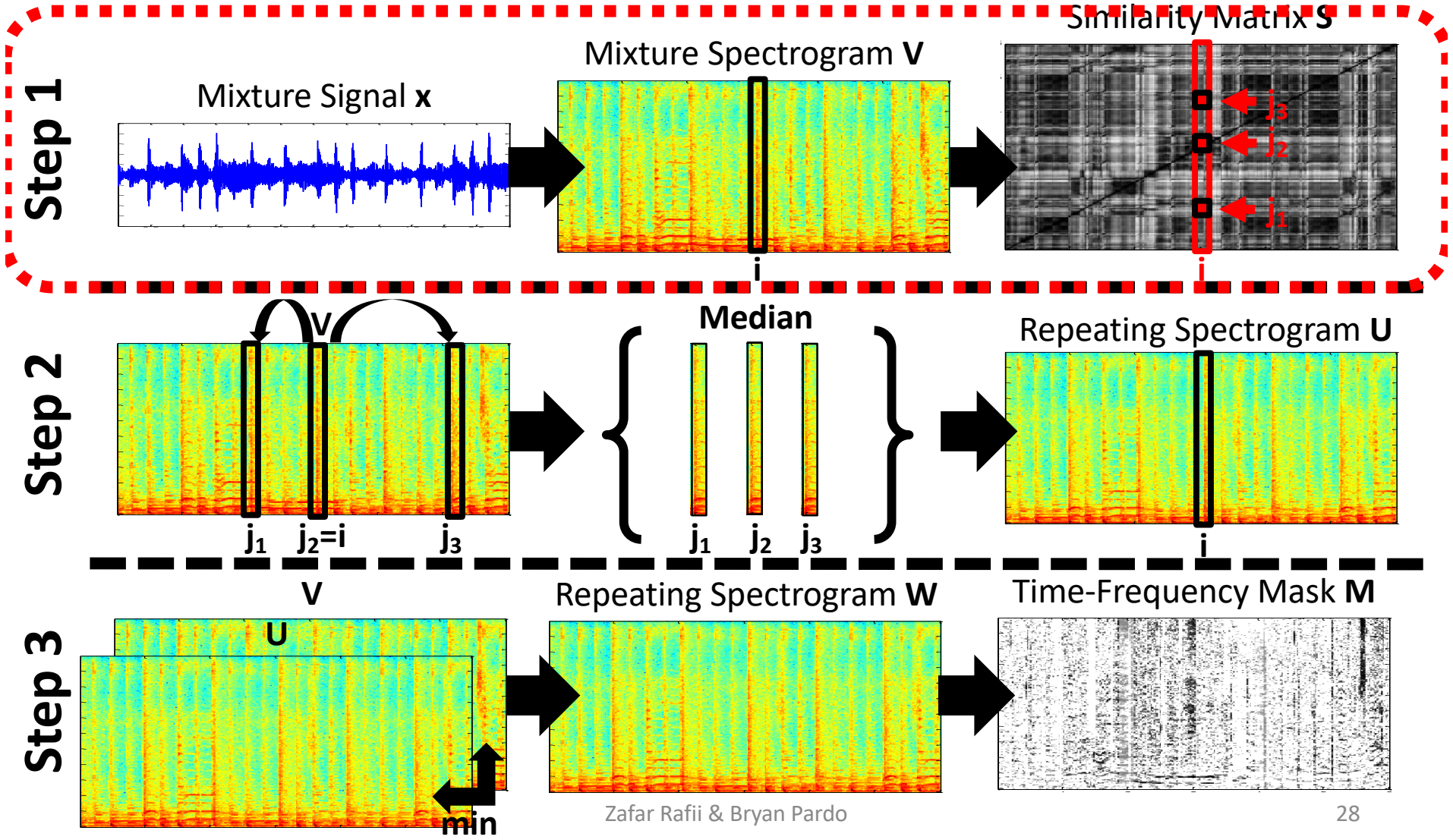


1. Repeating Elements

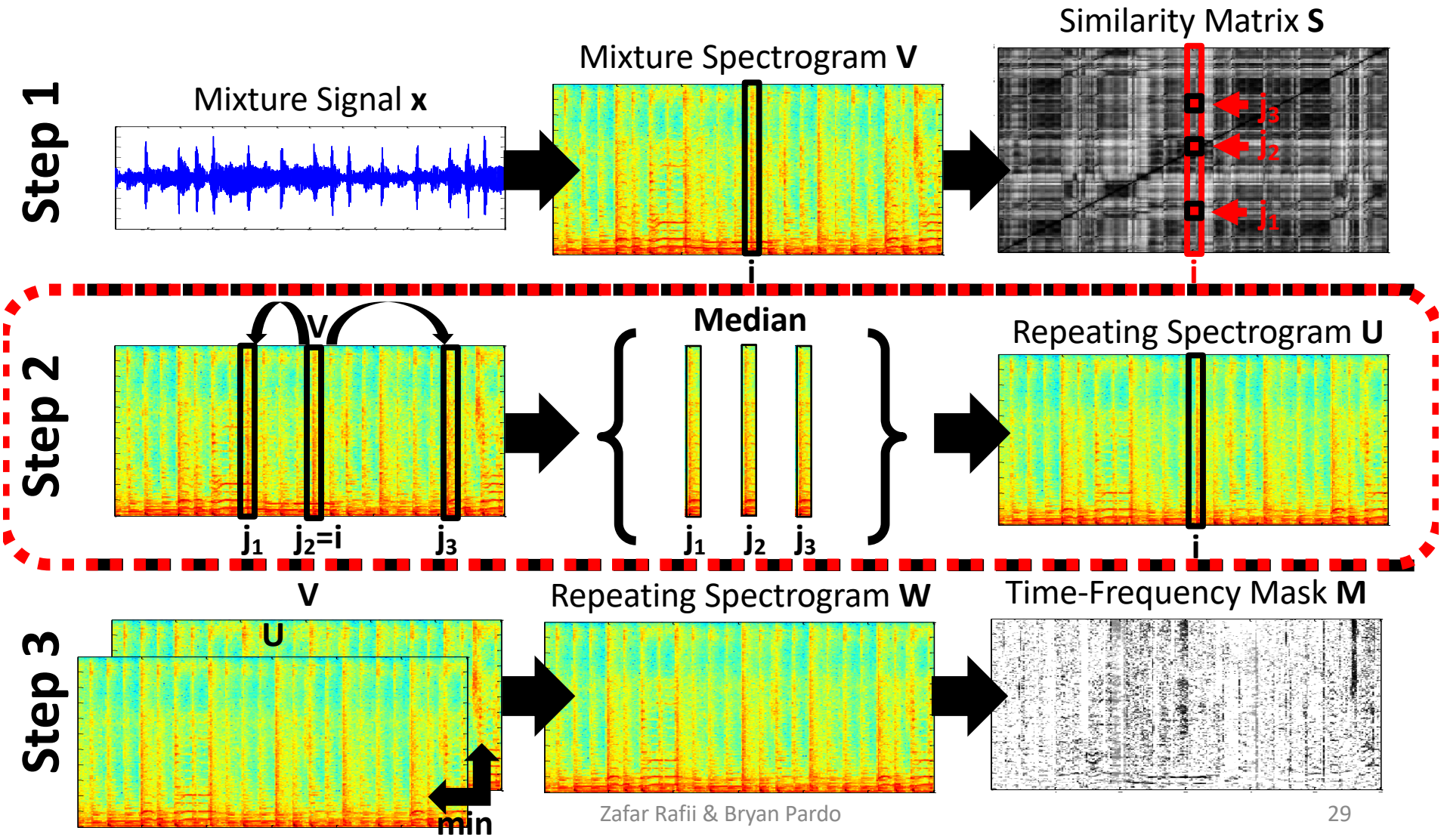
- The SM reveals for every frame i, the frames j_k that are **the most similar** to frame i



1. Repeating Elements

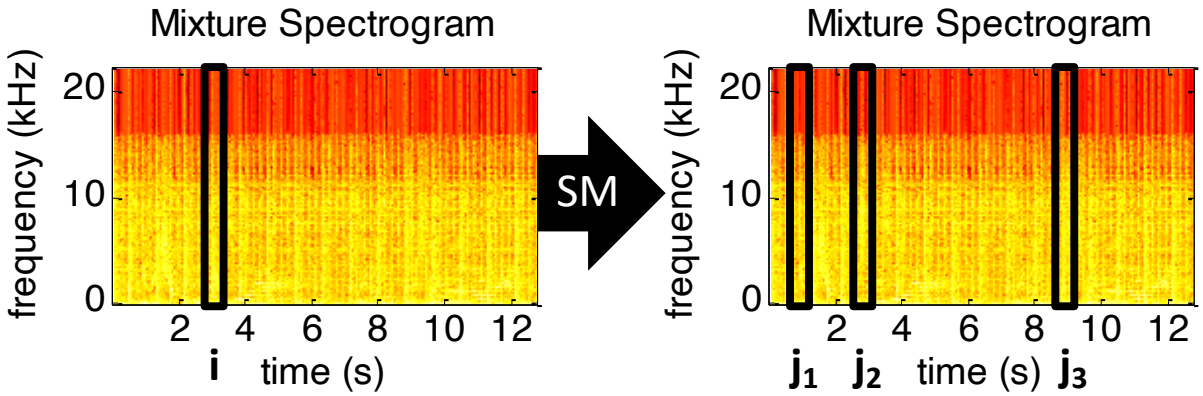


2. Repeating Model



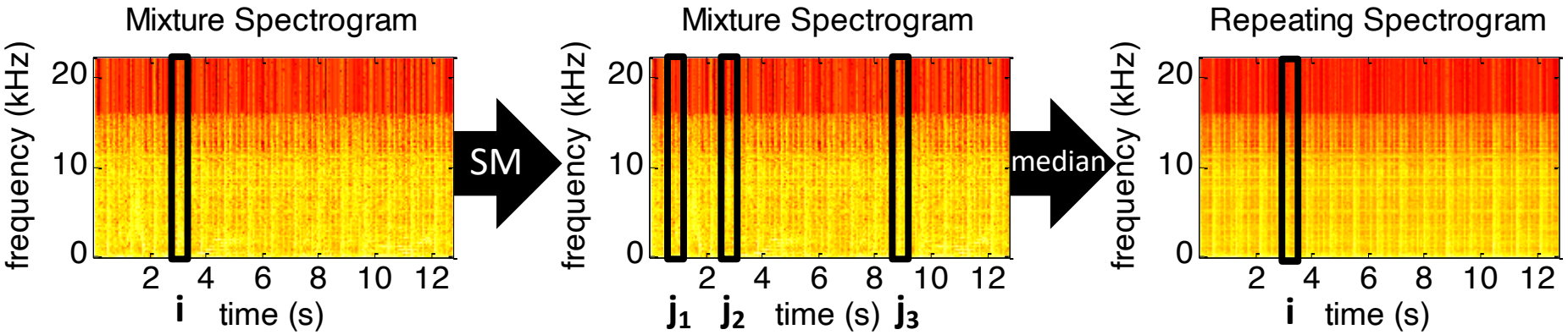
2. Repeating Model

- For every frame i, we take the **median** of its most similar frames j_k found using the SM

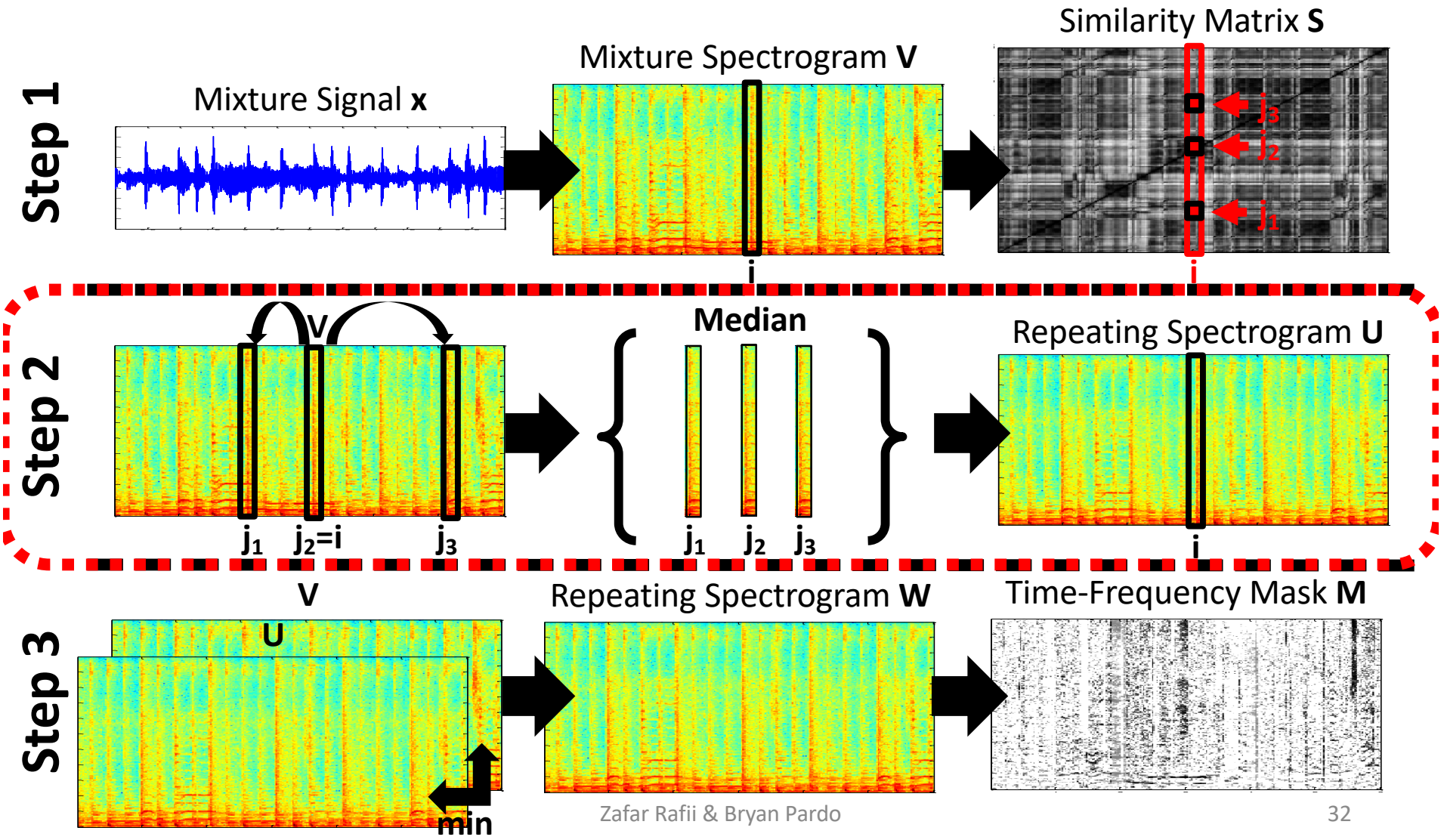


2. Repeating Model

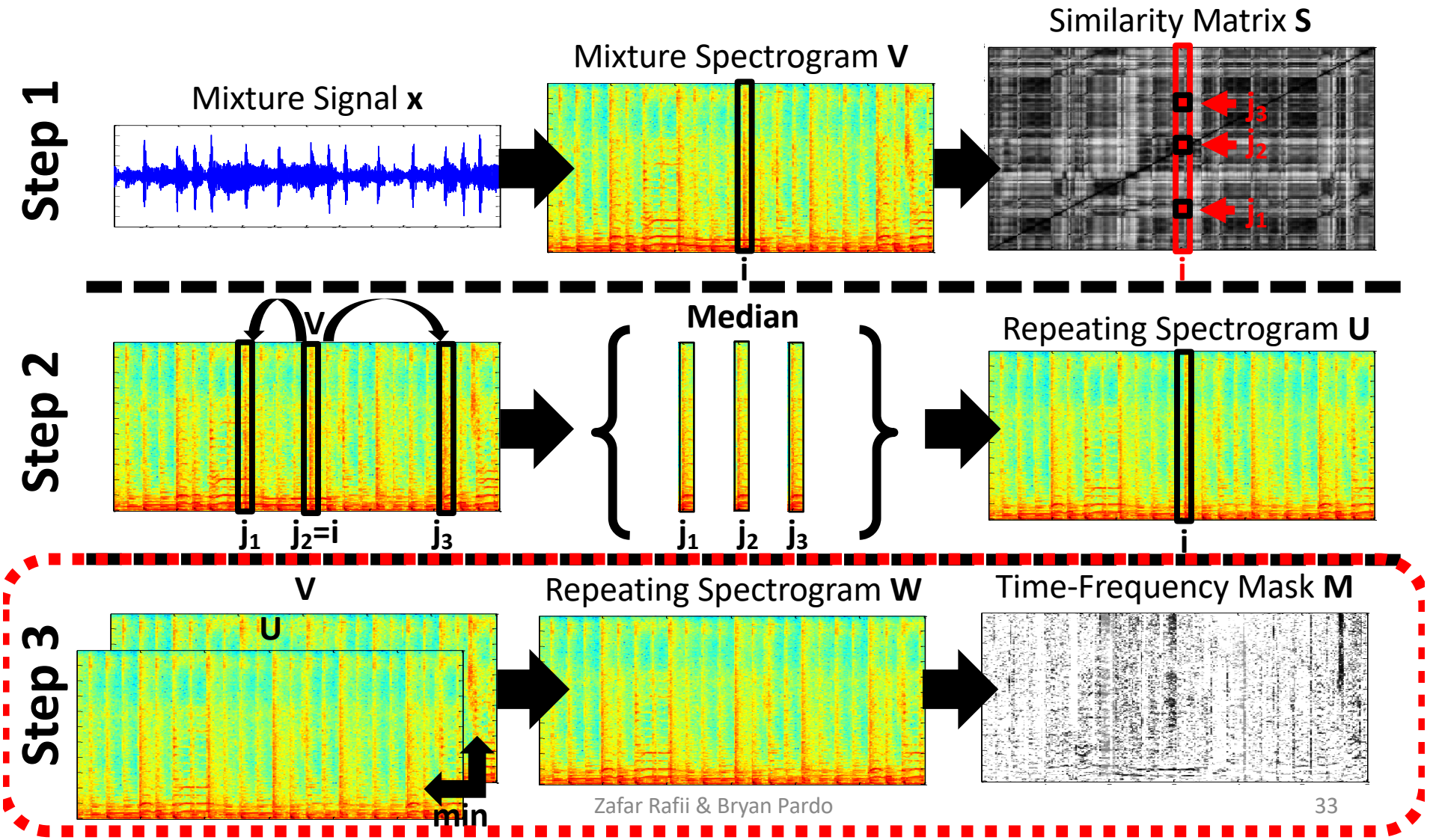
- We obtain an initial **repeating spectrogram model**



2. Repeating Model

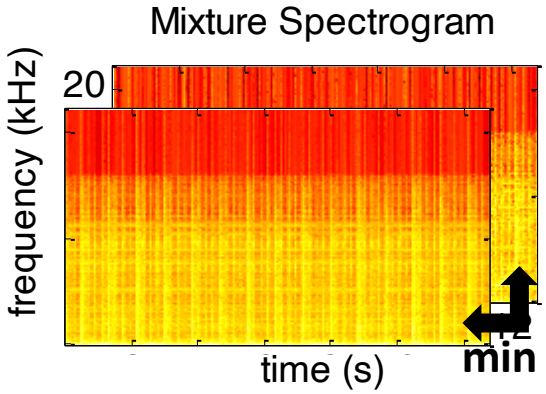


3. Repeating Structure



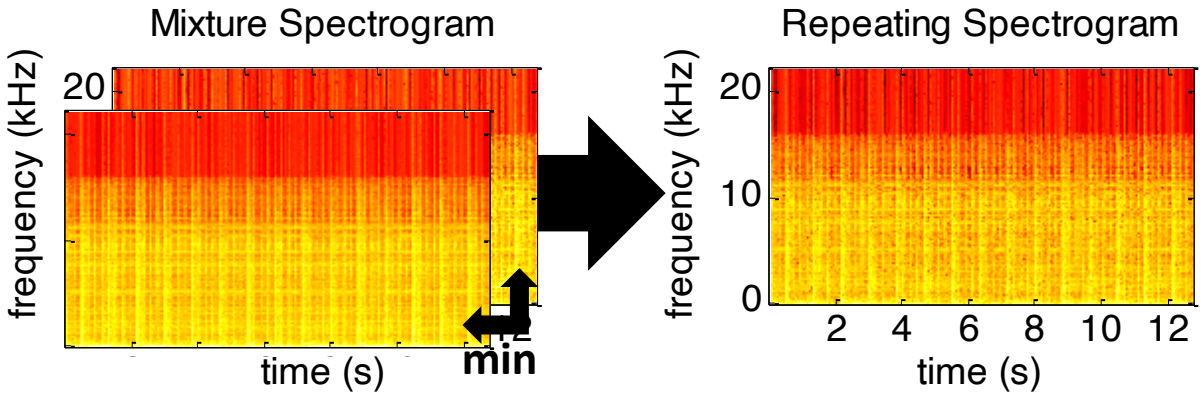
3. Repeating Structure

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



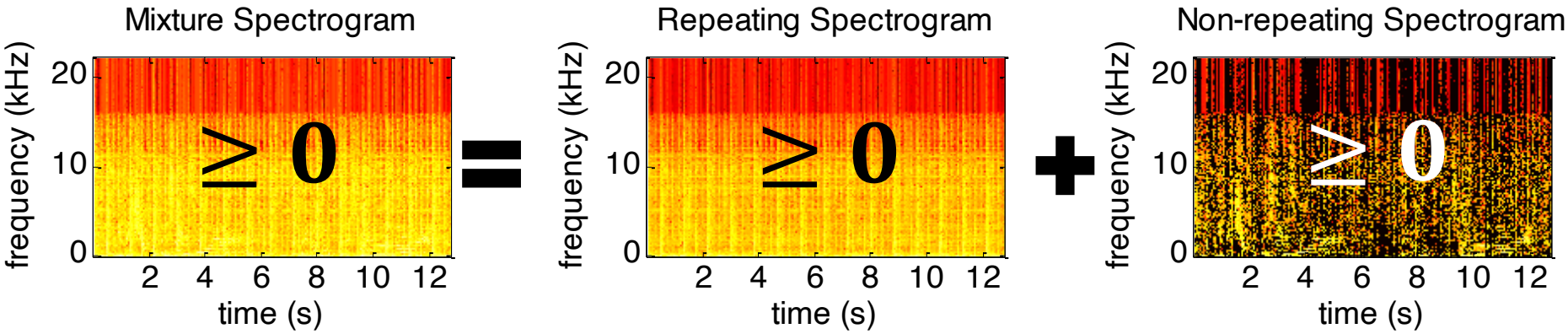
3. Repeating Structure

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



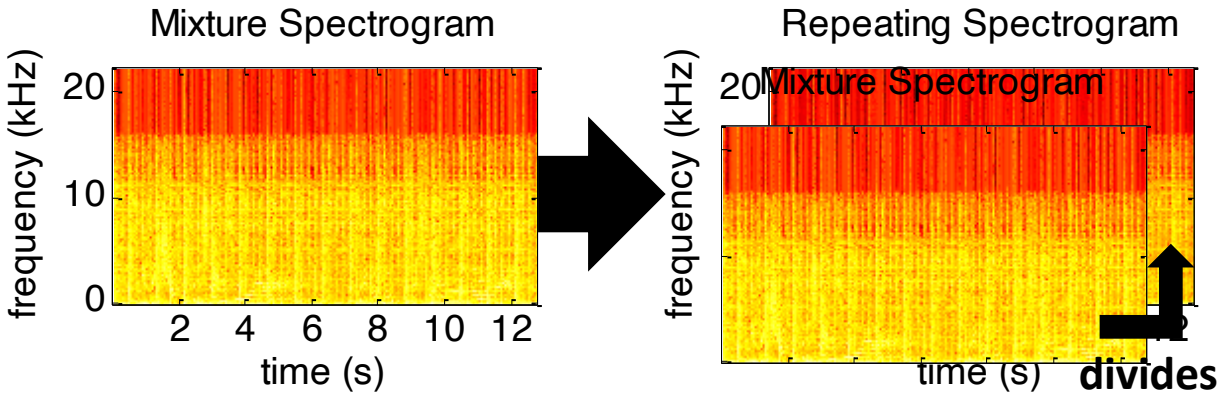
3. Repeating Structure

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



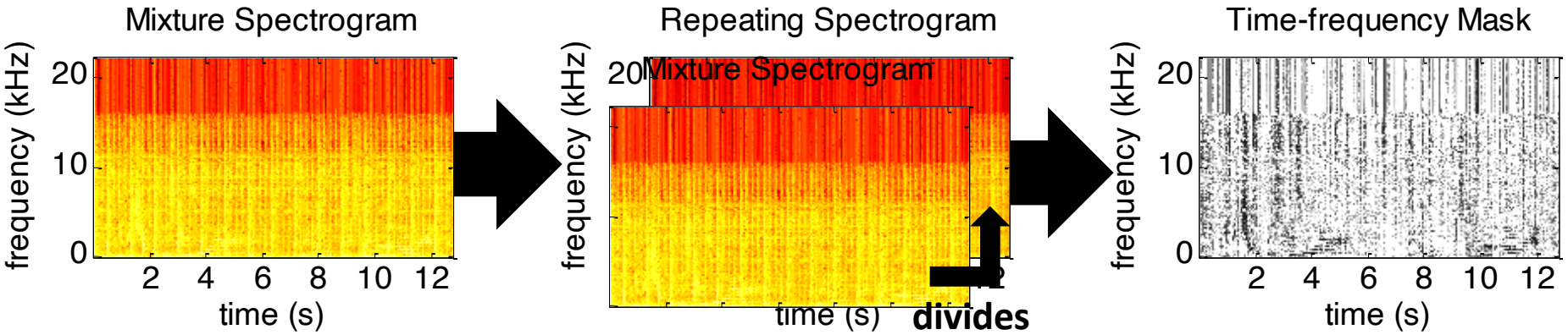
3. Repeating Structure

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



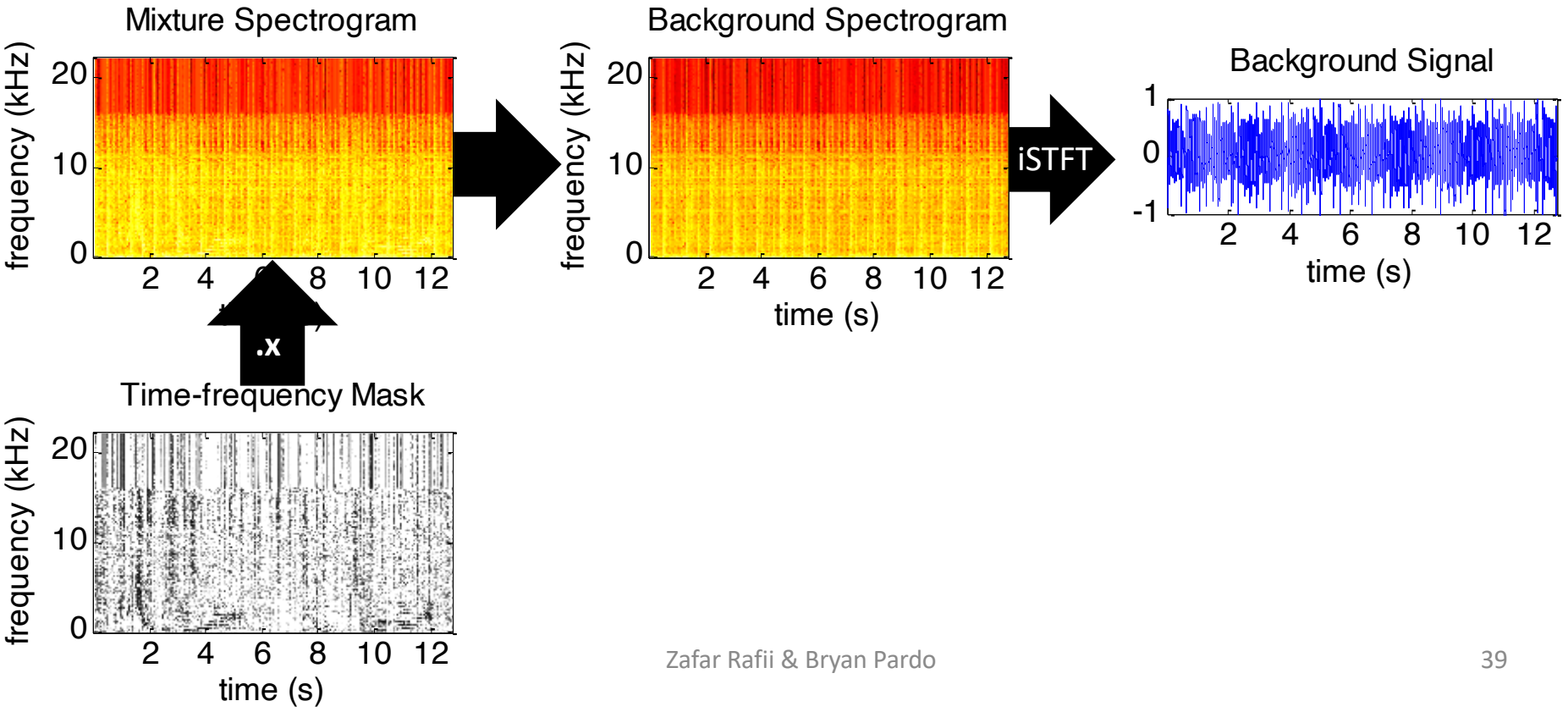
3. Repeating Structure

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



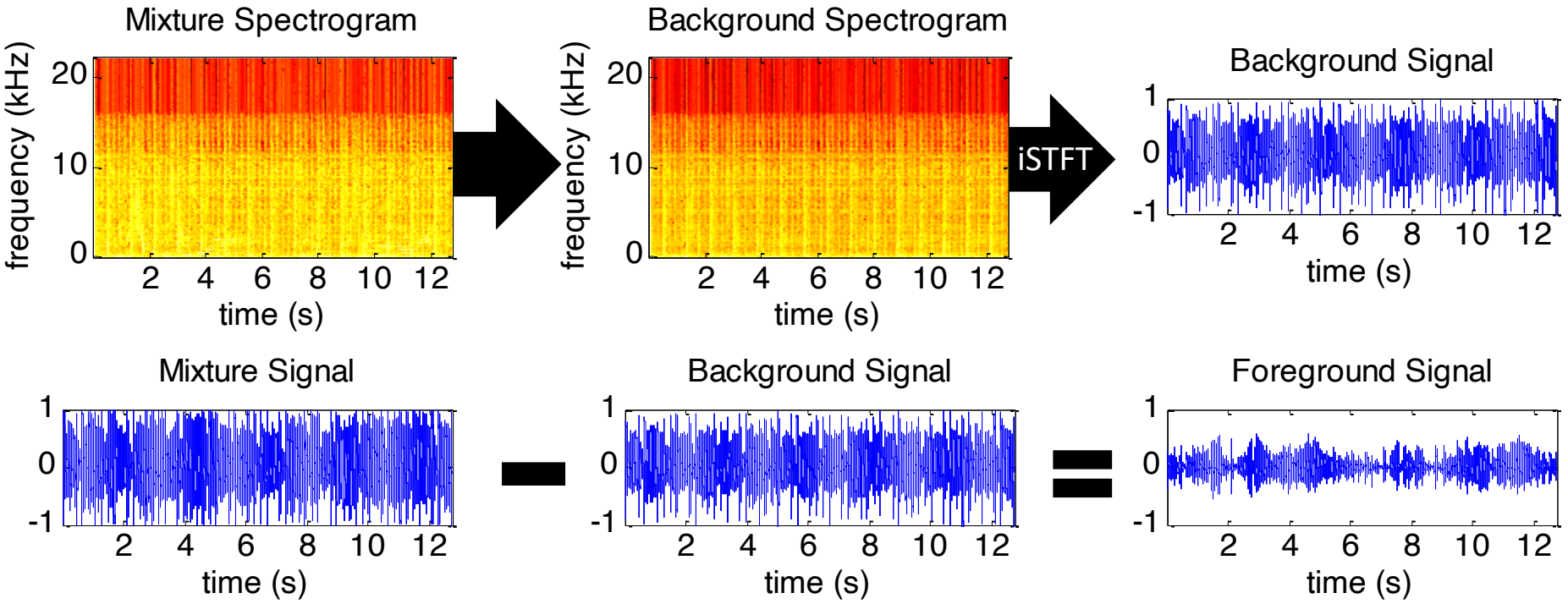
3. Repeating Structure

- We apply the t-f mask to the mixture STFT and obtain the **repeating background**



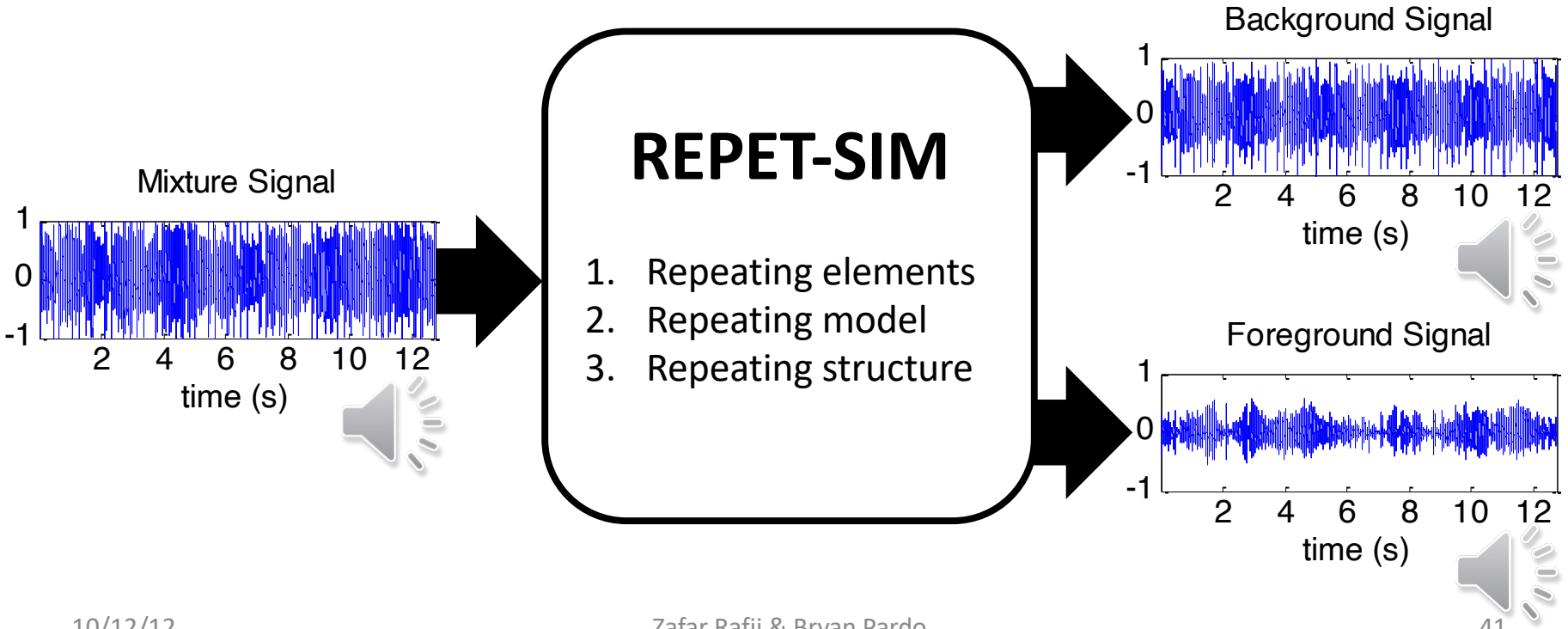
3. Repeating Structure

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



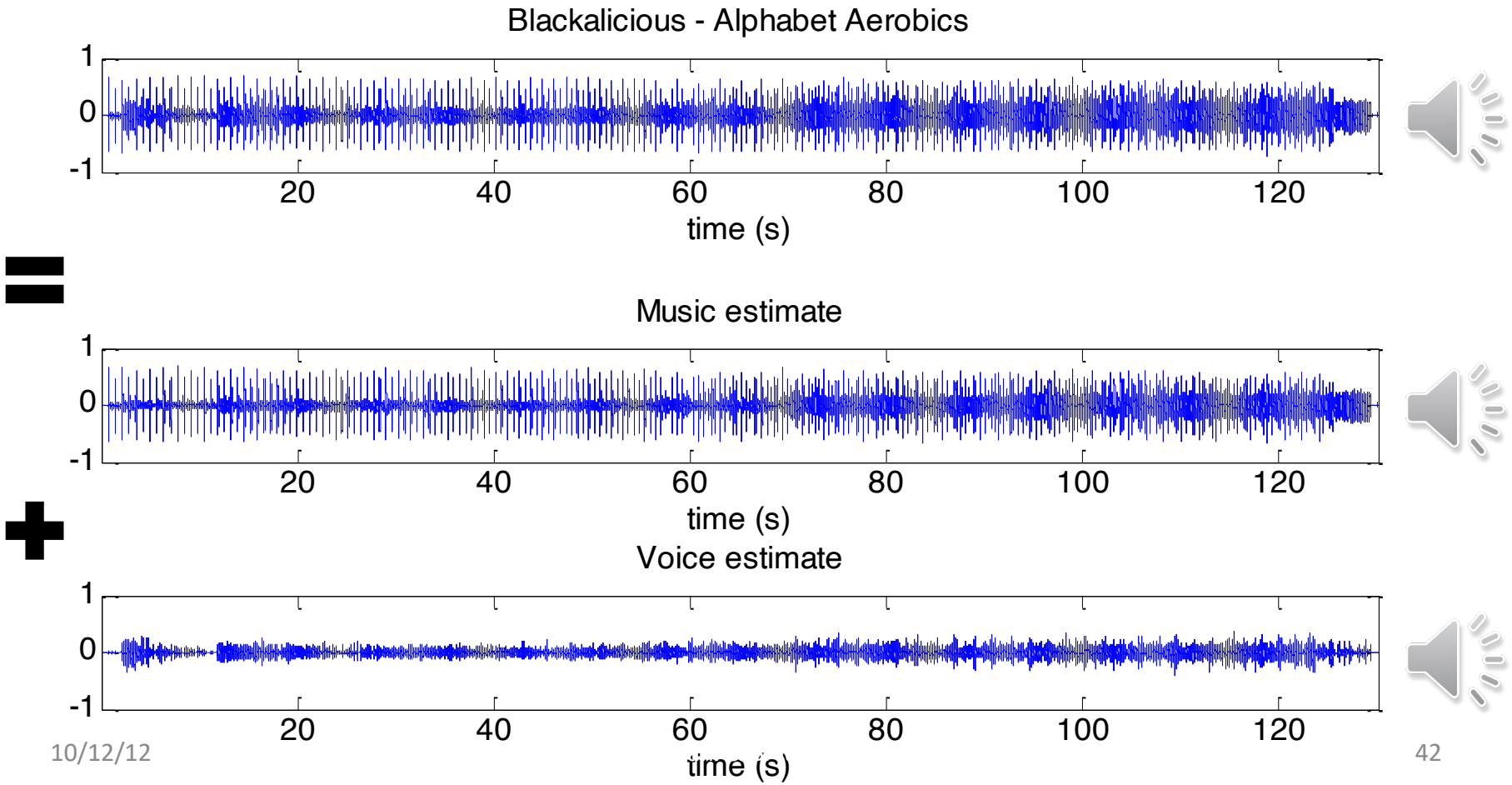
Music/Voice Separation

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



Examples

- REPET-SIM



Conclusion

- The analysis of the repetitions/similarities in music can be used for **source separation**