

Music/Voice Separation using the Similarity Matrix

Zafar Rafii & Bryan Pardo



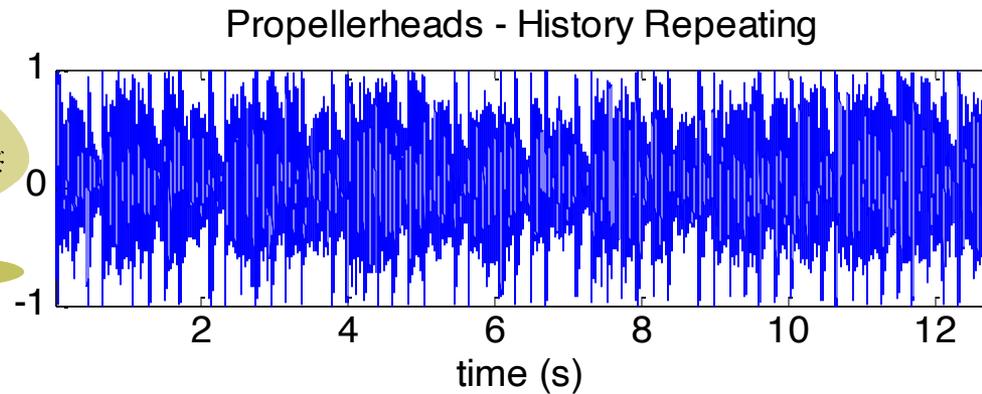
NORTHWESTERN
UNIVERSITY



**interactive
audio lab**

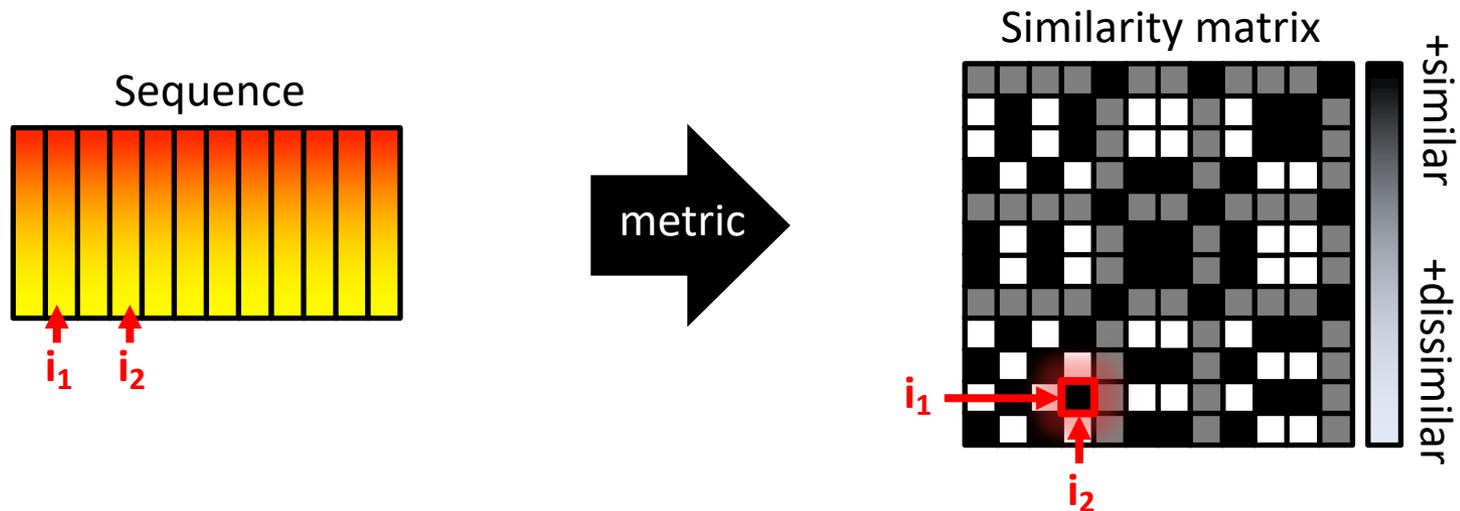
Introduction

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



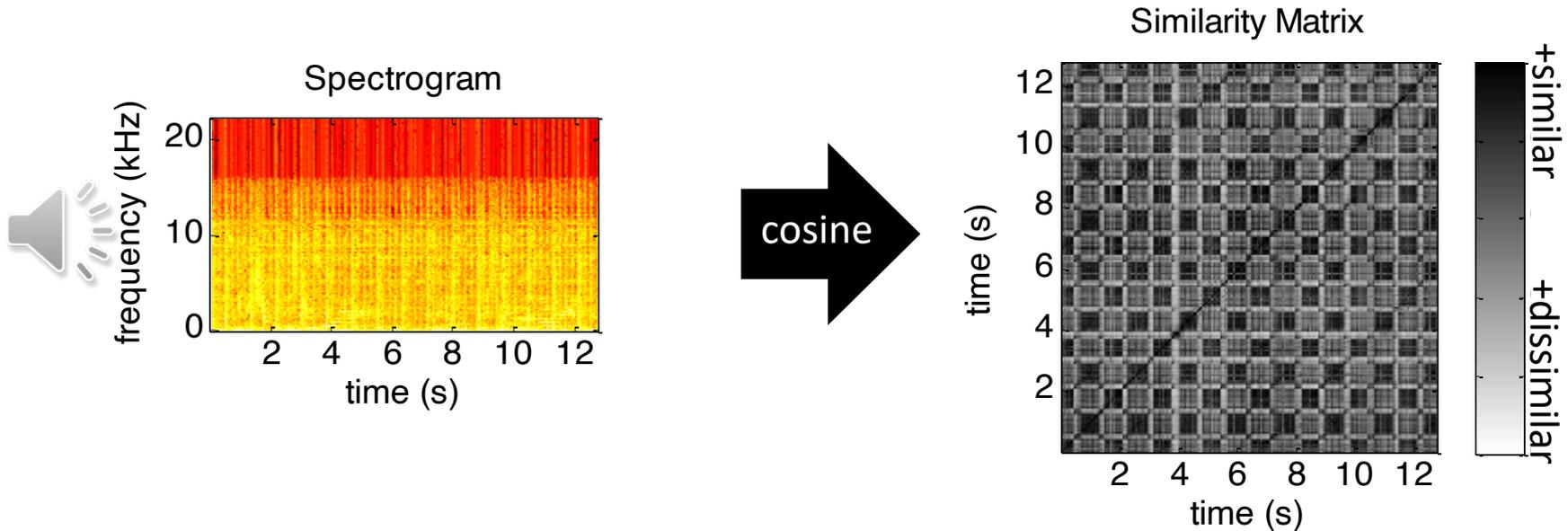
Similarity Matrix

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



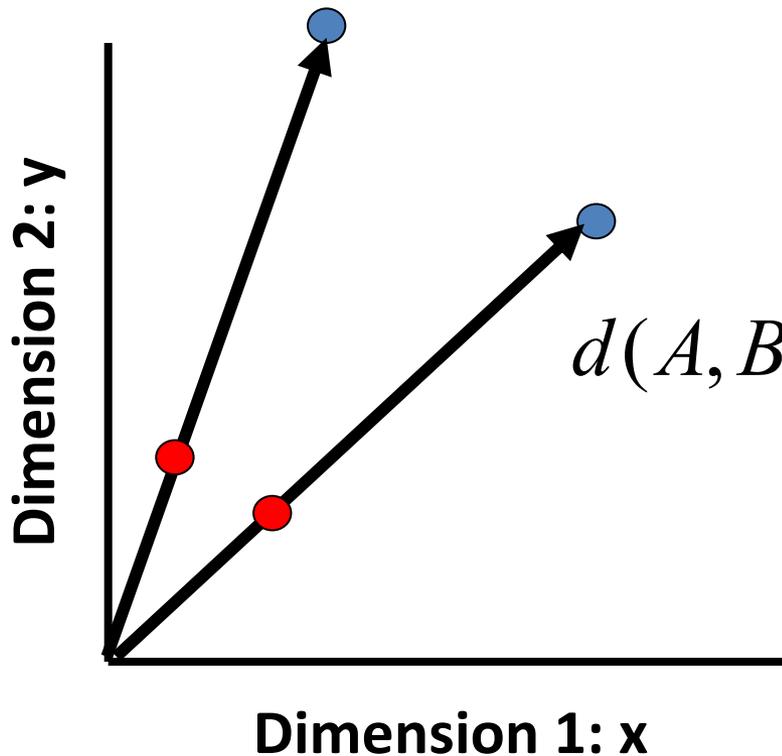
Similarity Matrix

- 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”



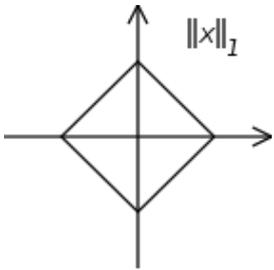
$$d(A, B) = \sqrt{(a_x - b_x)^2 + (a_y - b_y)^2}$$

L^p norms

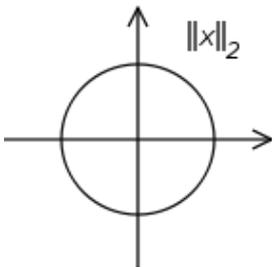
- L^p norms are all special cases of this function:

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

p changes the norm



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



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

Cosine Similarity

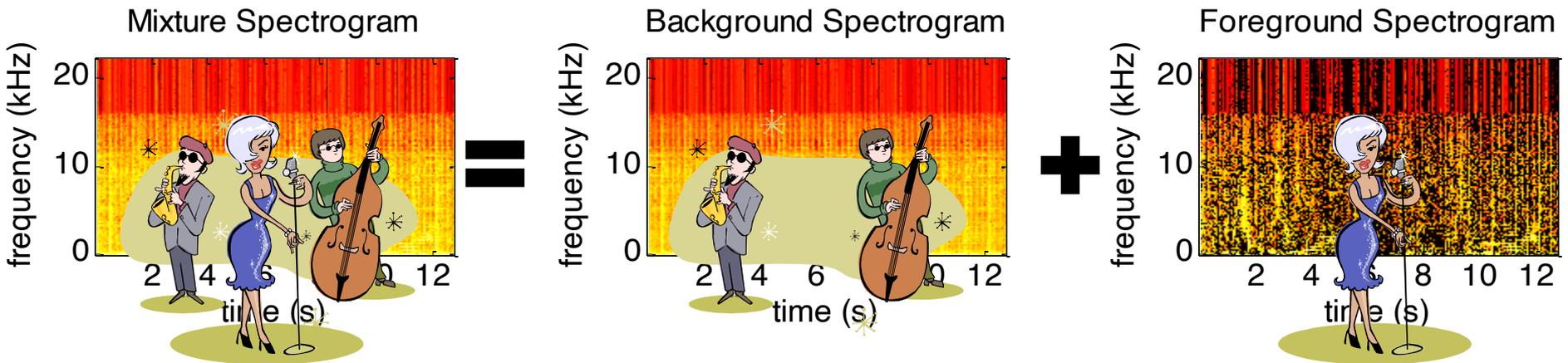
- Measure of similarity between two vectors
 - Range from -1 (opposite) to 1 (same)
 - Cosine distance = 1 – cosine similarity
- Cosine similarity between vector A and B :

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

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

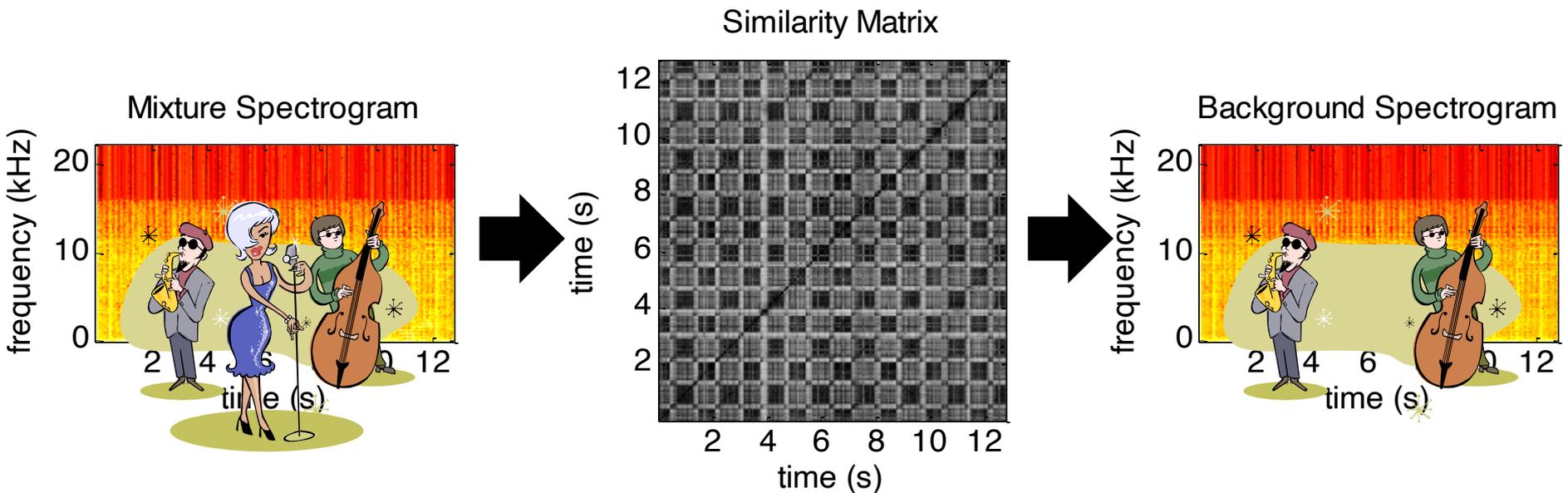
Assumptions

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



Assumptions

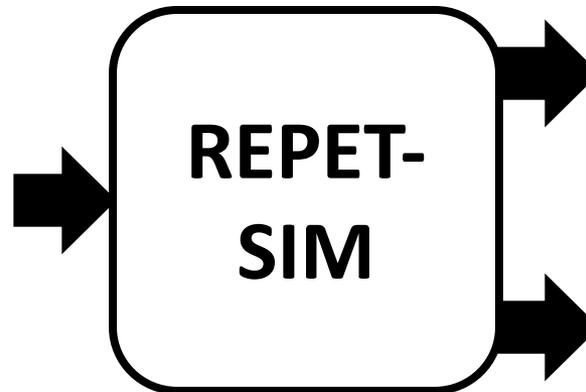
- The SM of a mixture is then likely to reveal the structure of the **repeating background**



REPET-SIM

- **REPET with Similarity Matrix!**

1. Identify the repeating/similar elements
2. Derive a repeating model
3. Extract the repeating structure



Repeating Structure

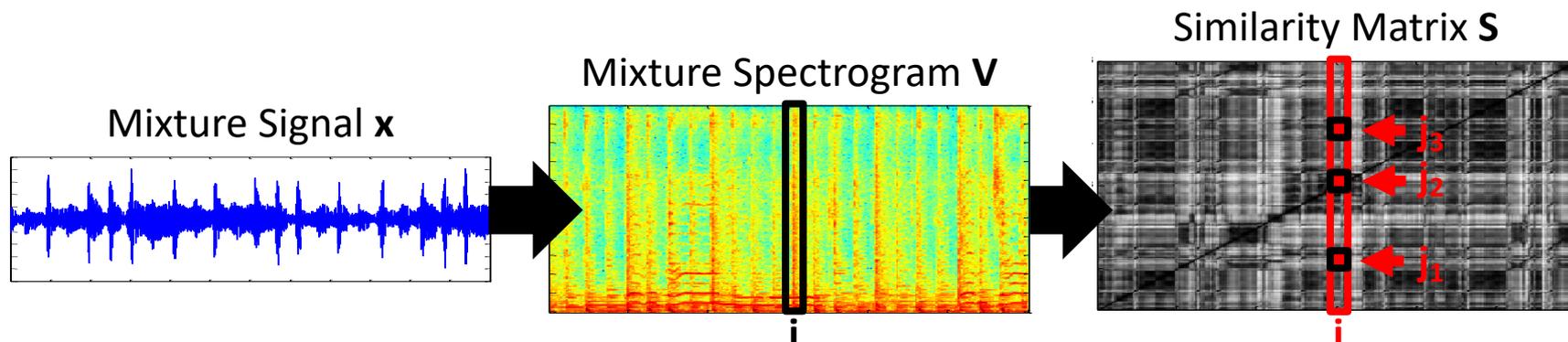


Non-repeating Structure

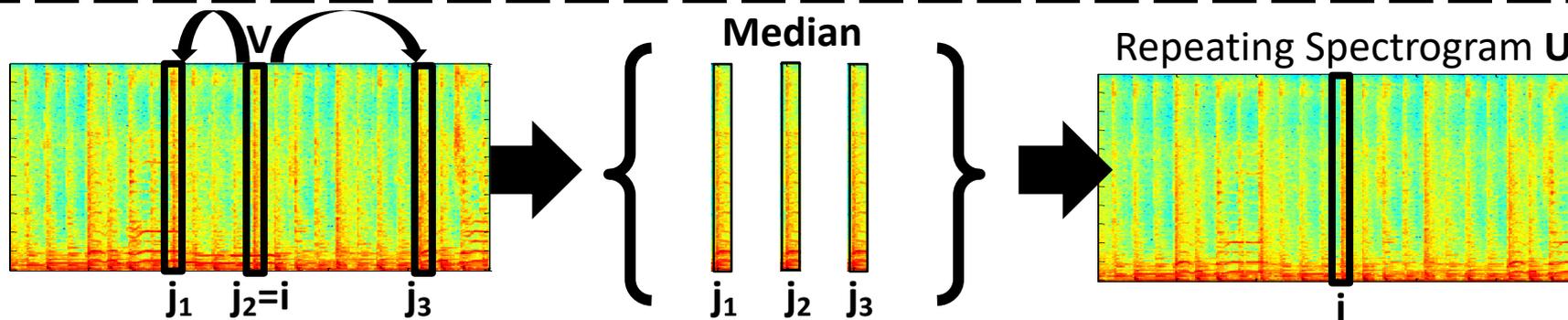


REPET-SIM

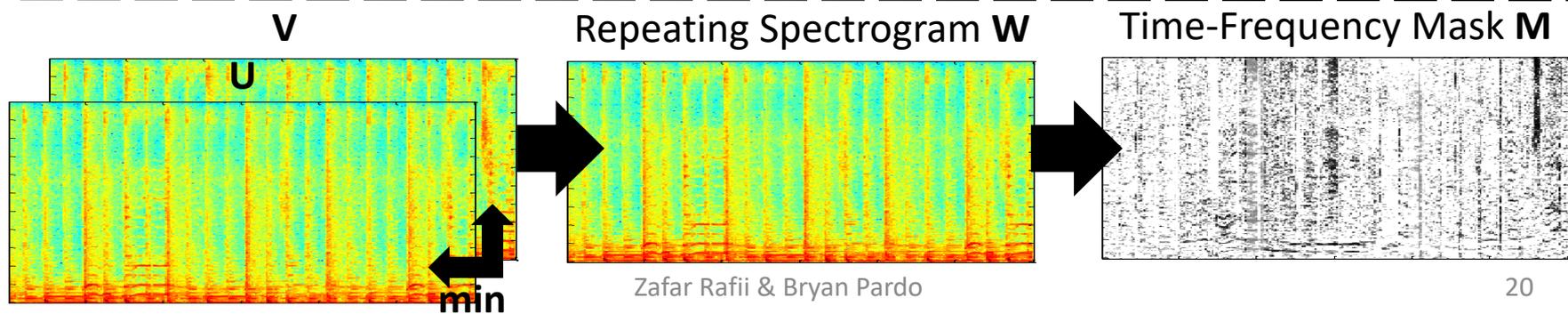
Step 1



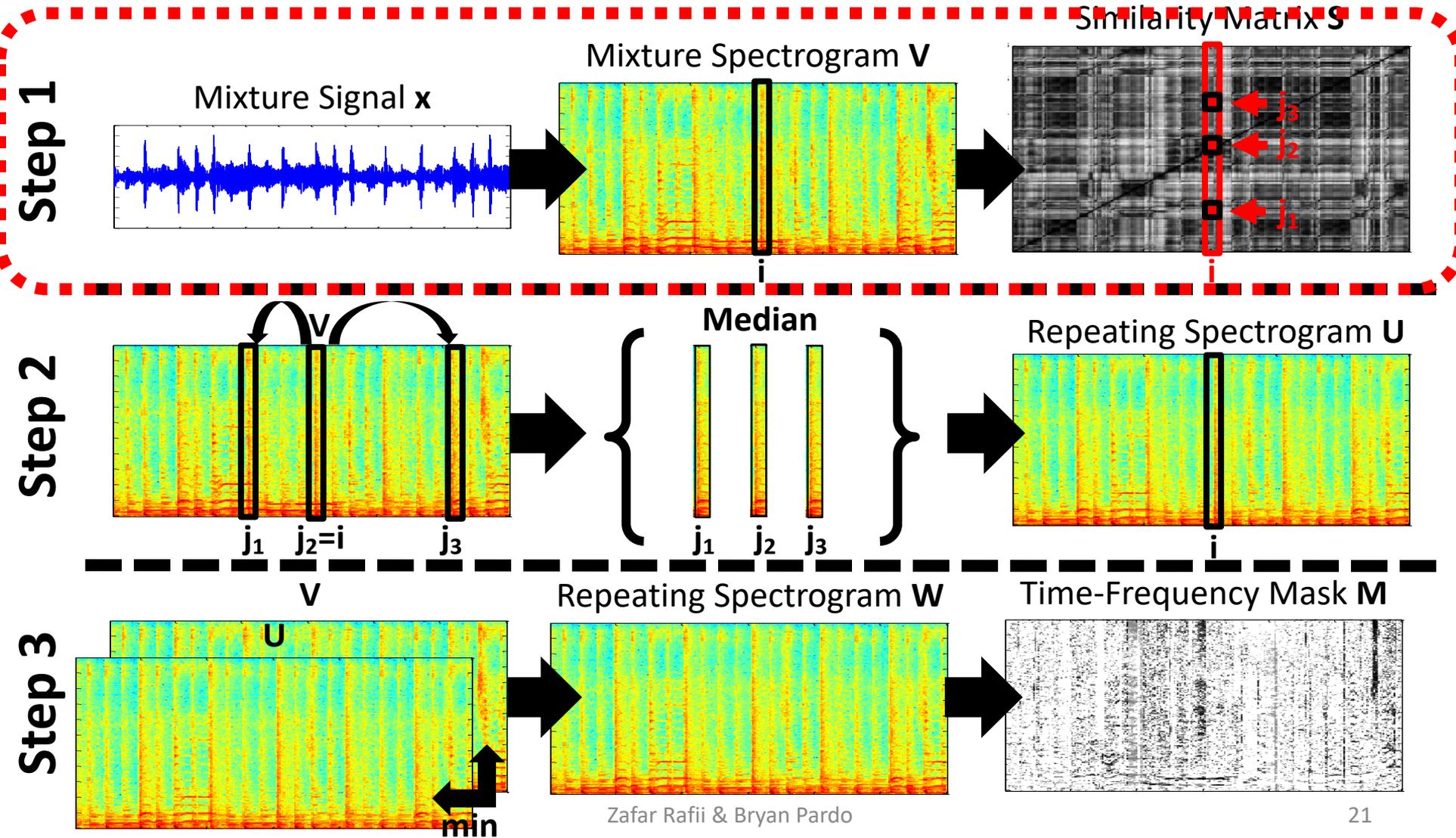
Step 2



Step 3

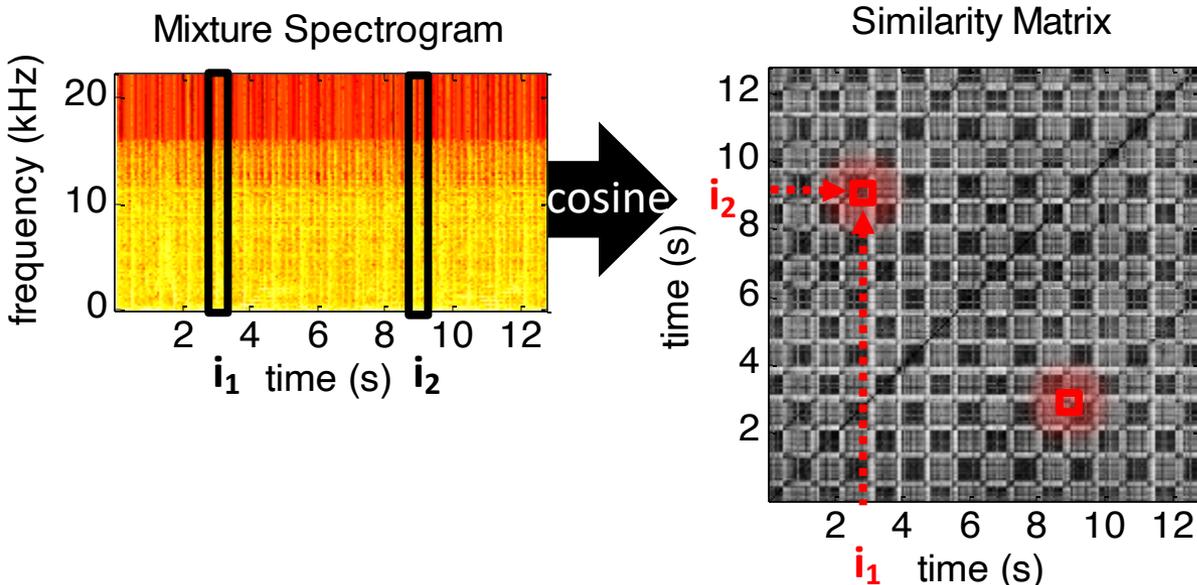


1. Repeating Elements



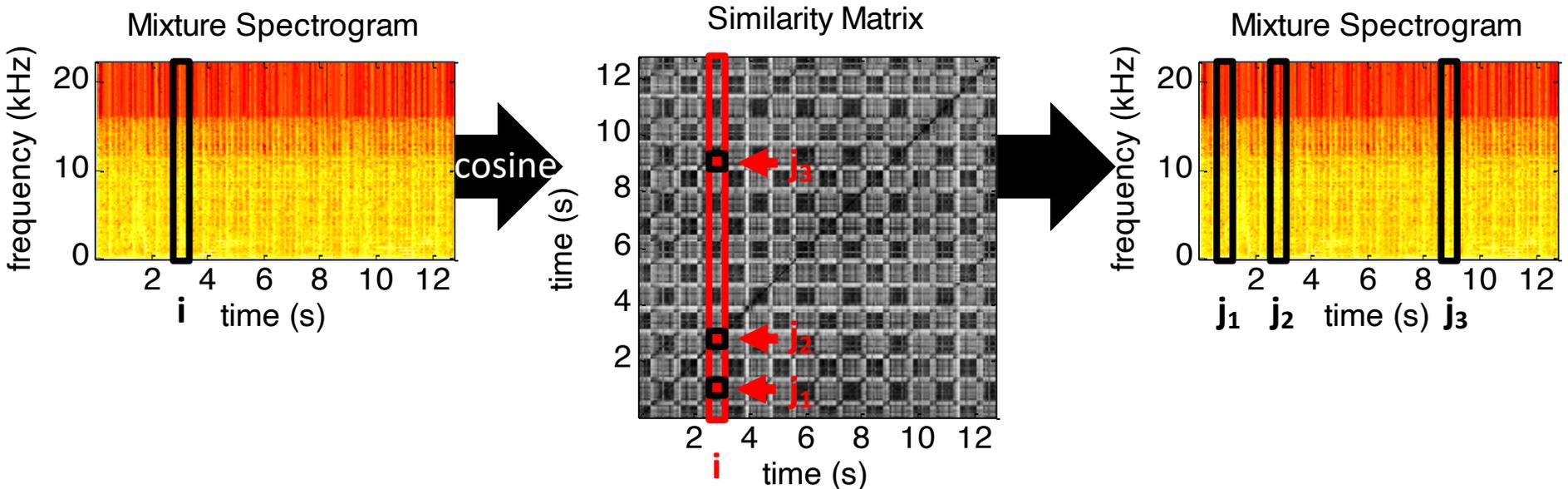
1. Repeating Elements

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

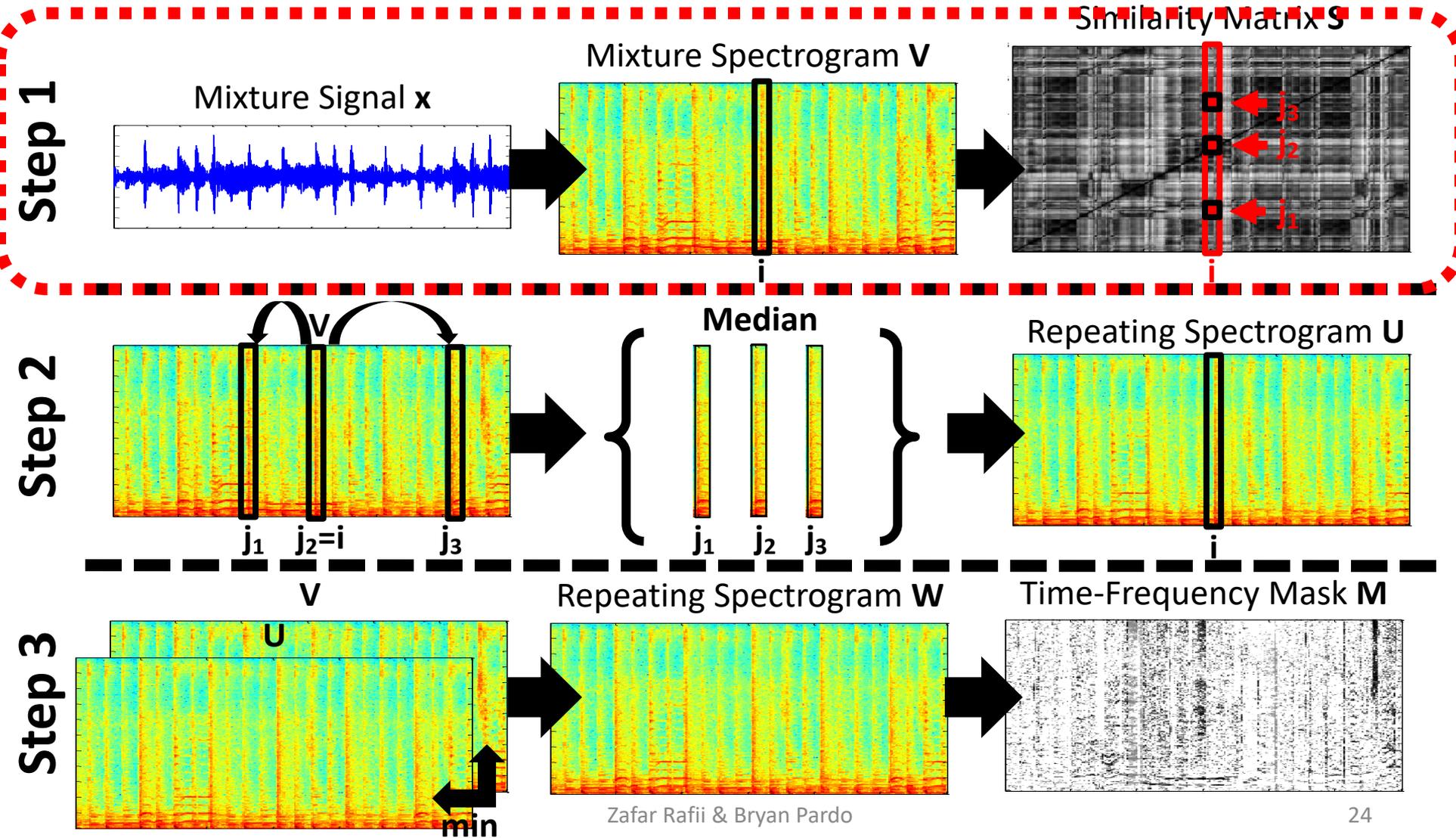


1. Repeating Elements

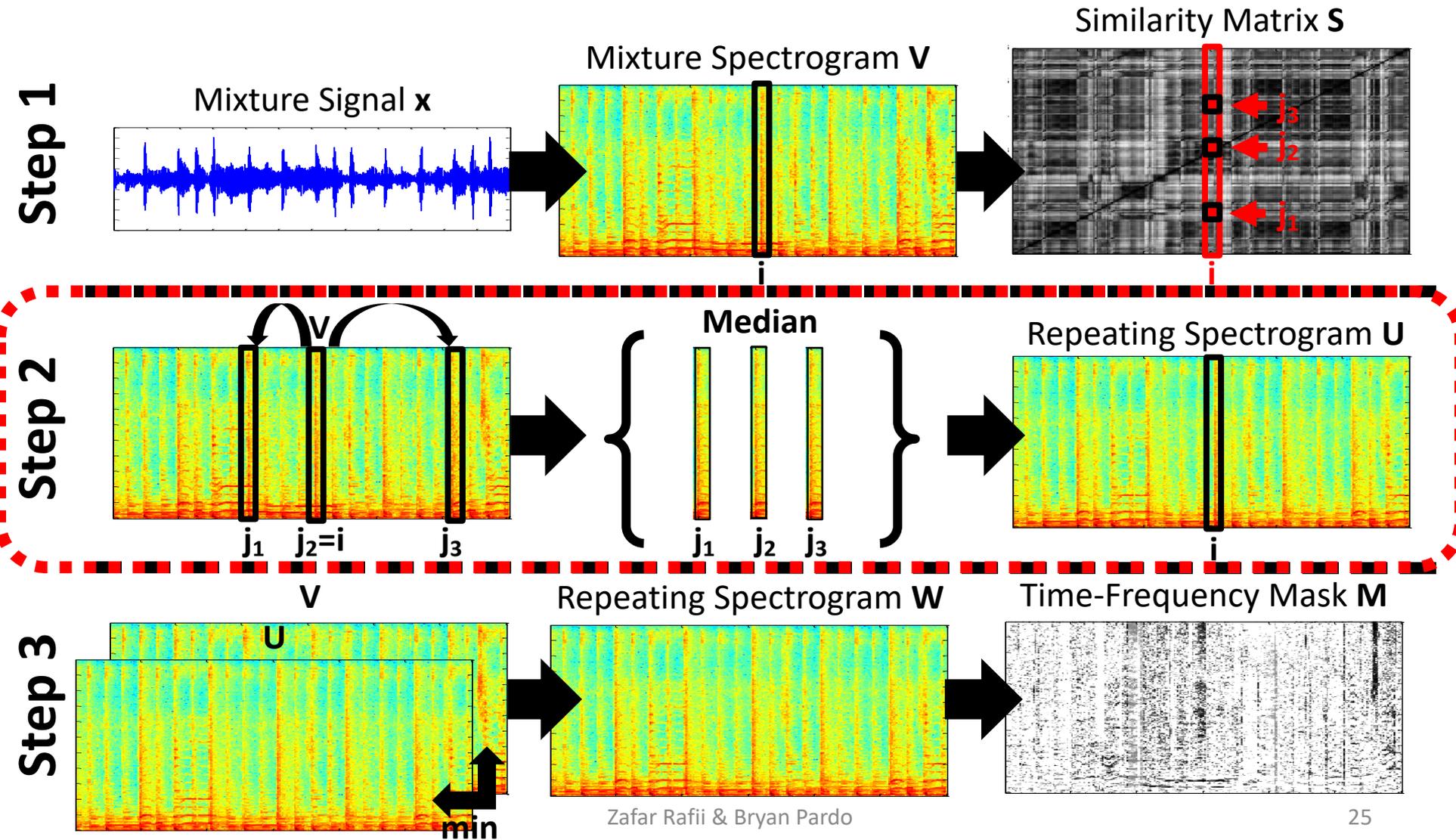
- The SM reveals for every frame i , the frames j_k that are **the most similar** to frame i



1. Repeating Elements

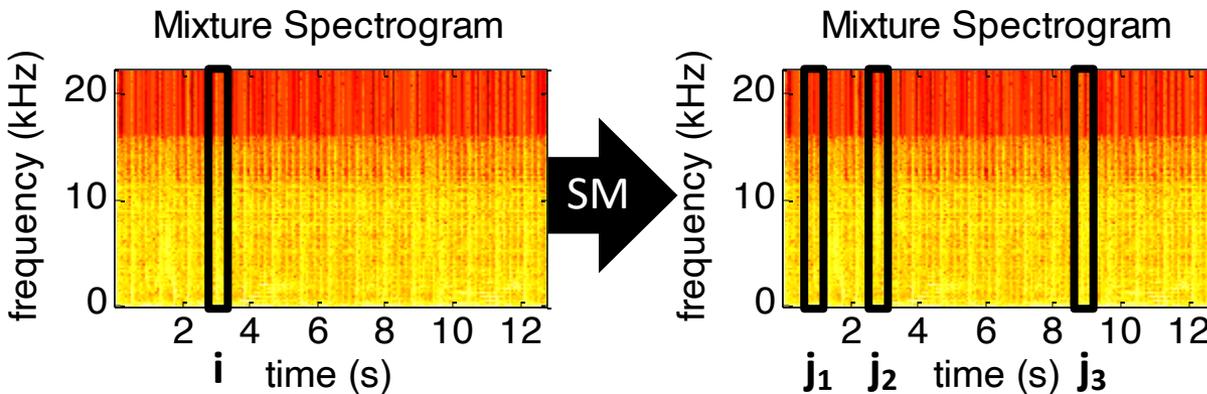


2. Repeating Model



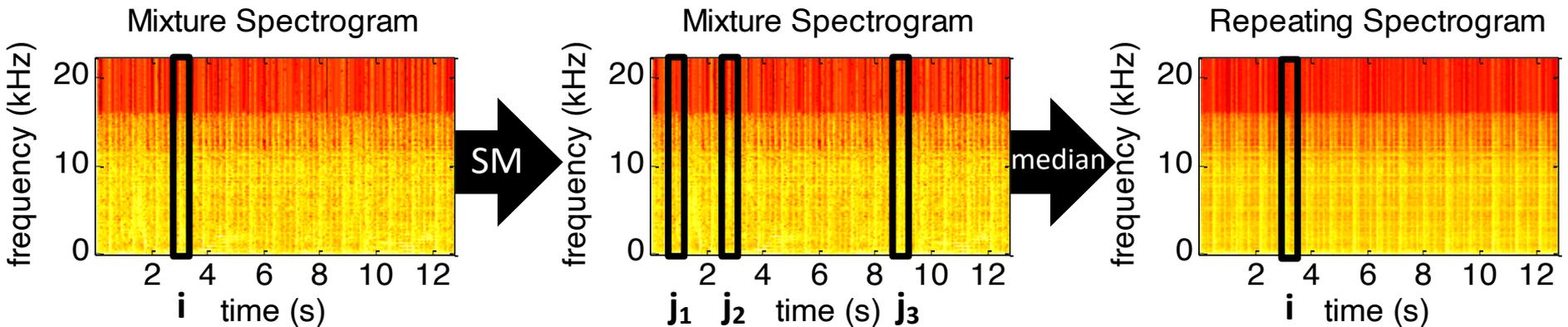
2. Repeating Model

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

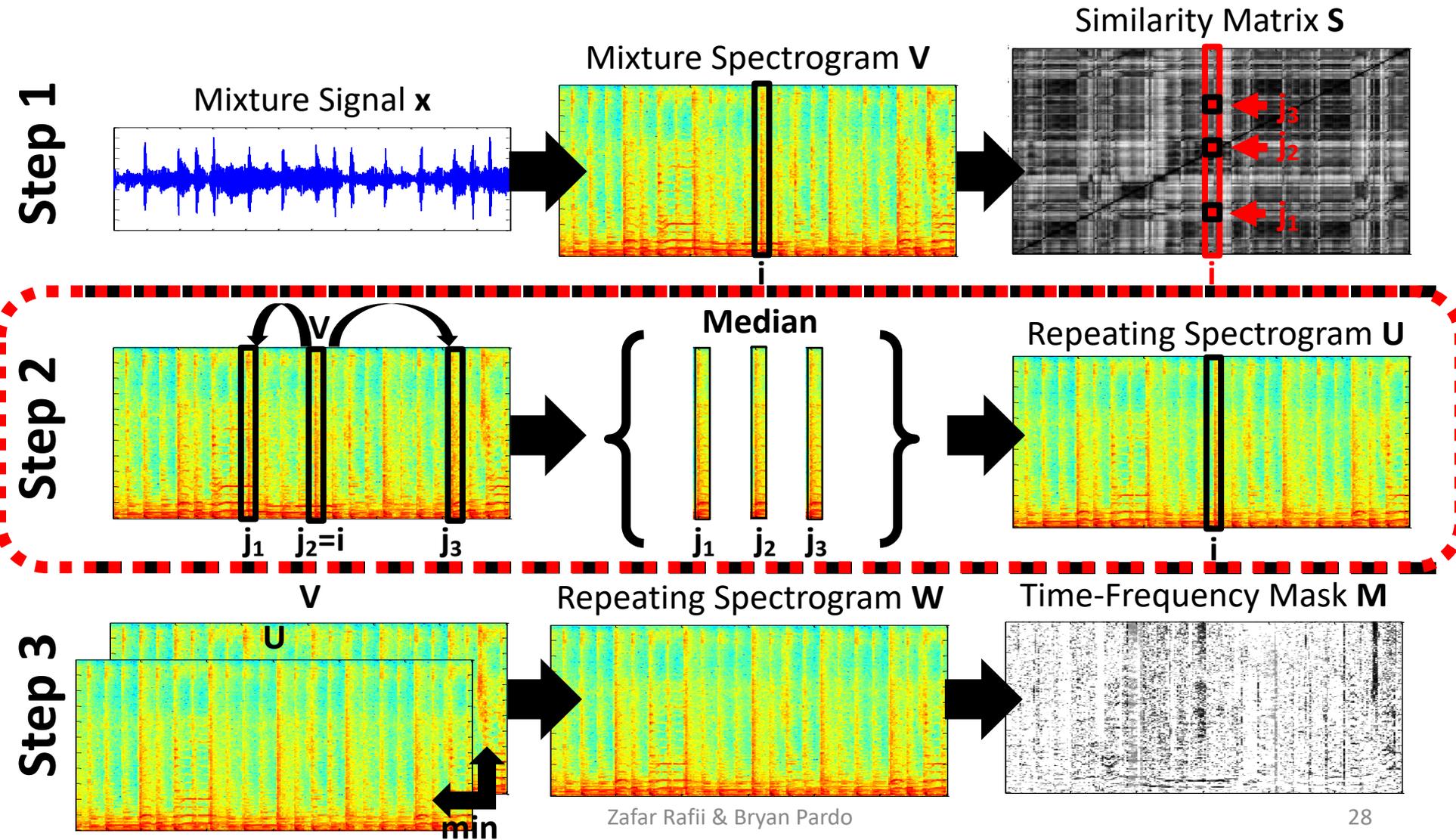


2. Repeating Model

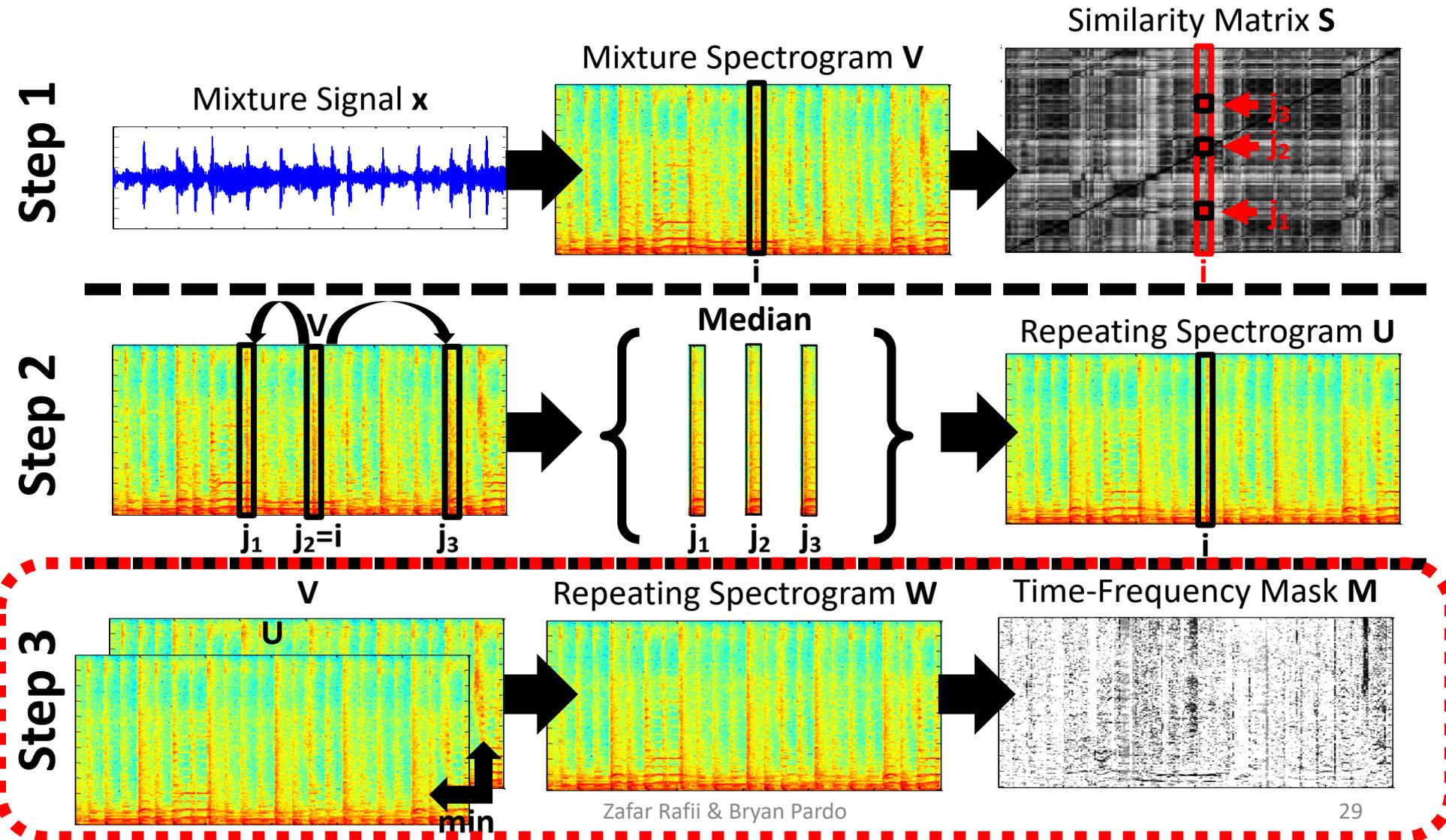
- We obtain an initial **repeating spectrogram model**



2. Repeating Model

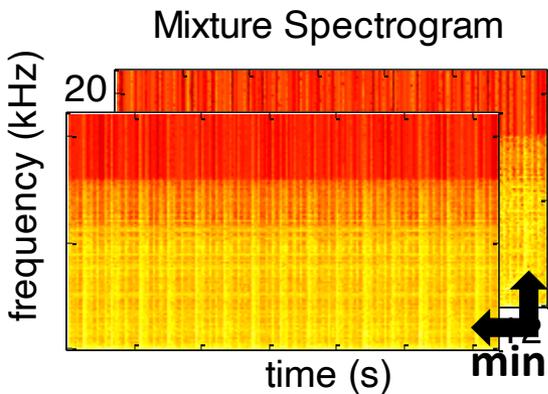


3. Repeating Structure



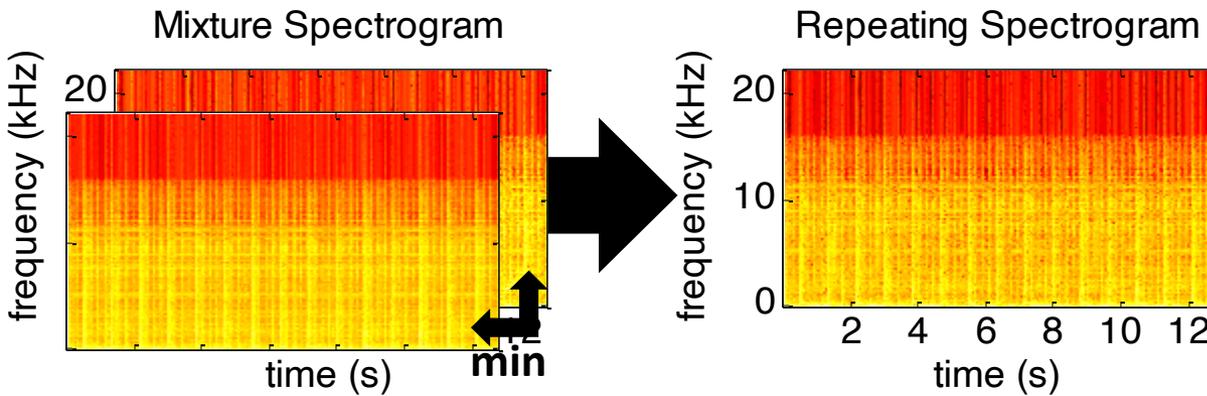
3. Repeating Structure

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



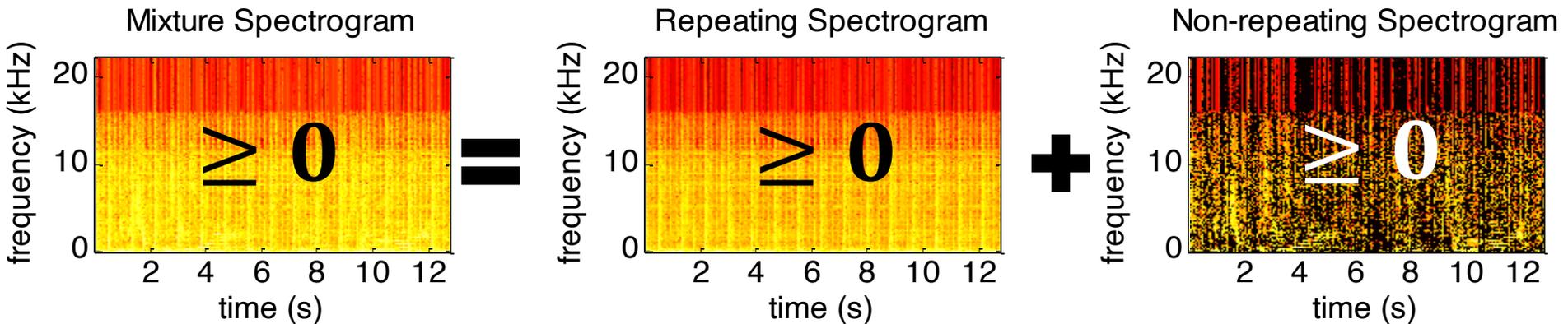
3. Repeating Structure

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



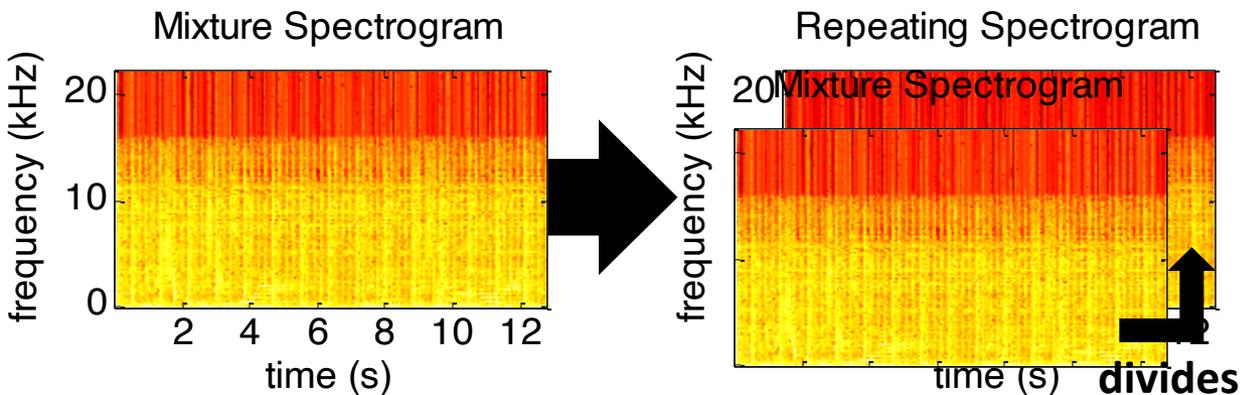
3. Repeating Structure

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



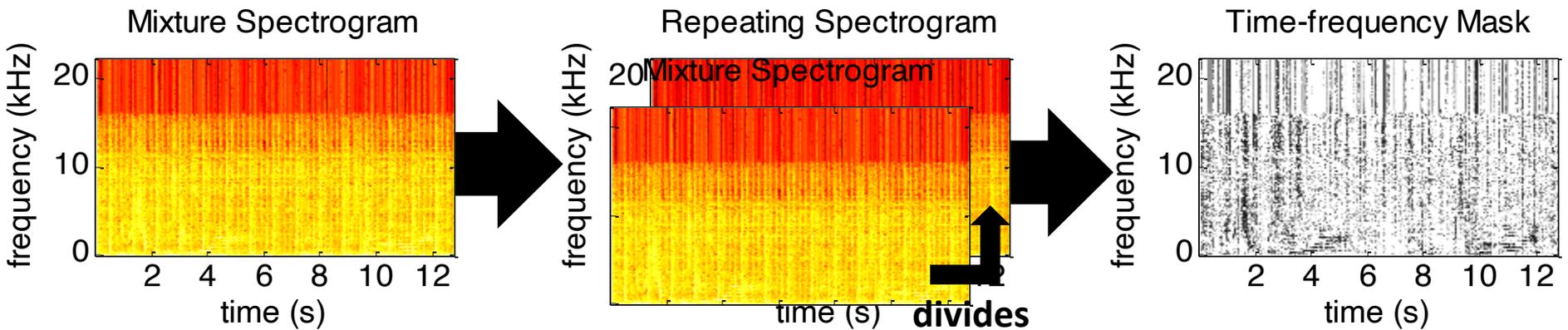
3. Repeating Structure

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



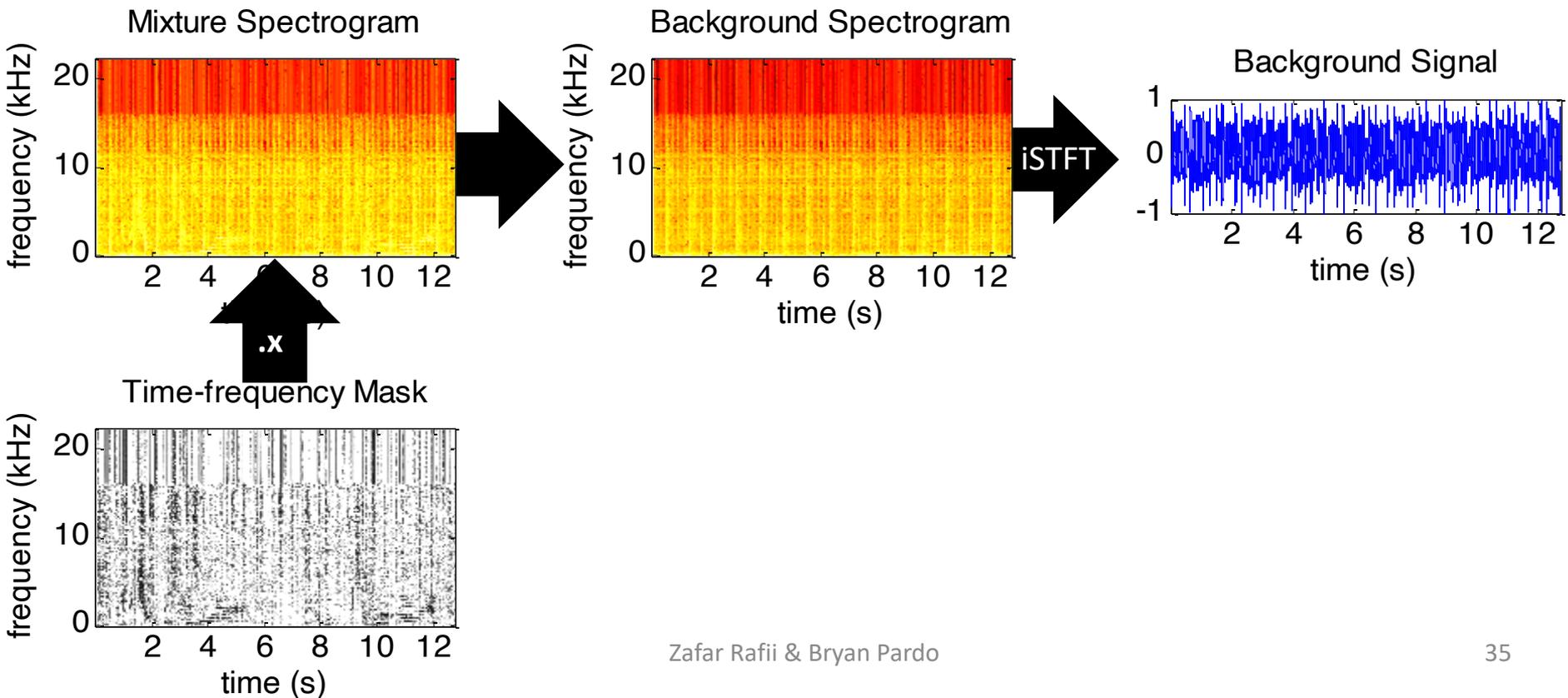
3. Repeating Structure

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



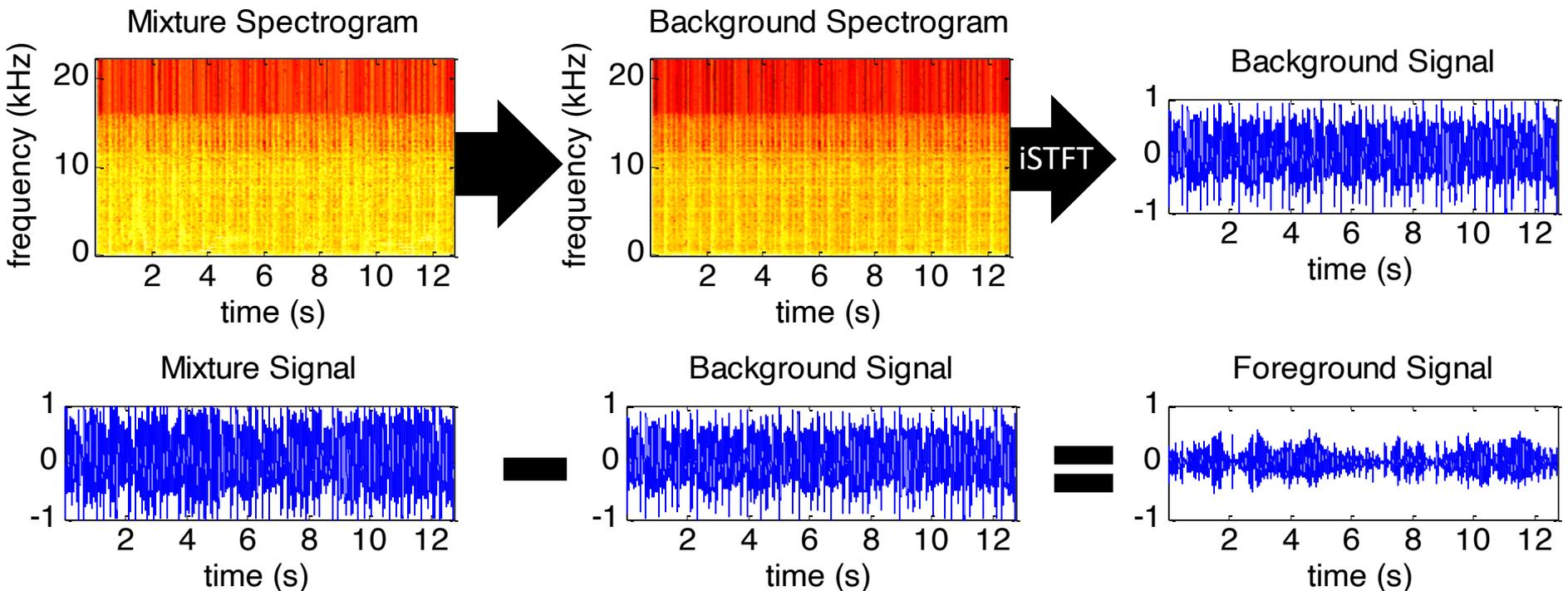
3. Repeating Structure

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



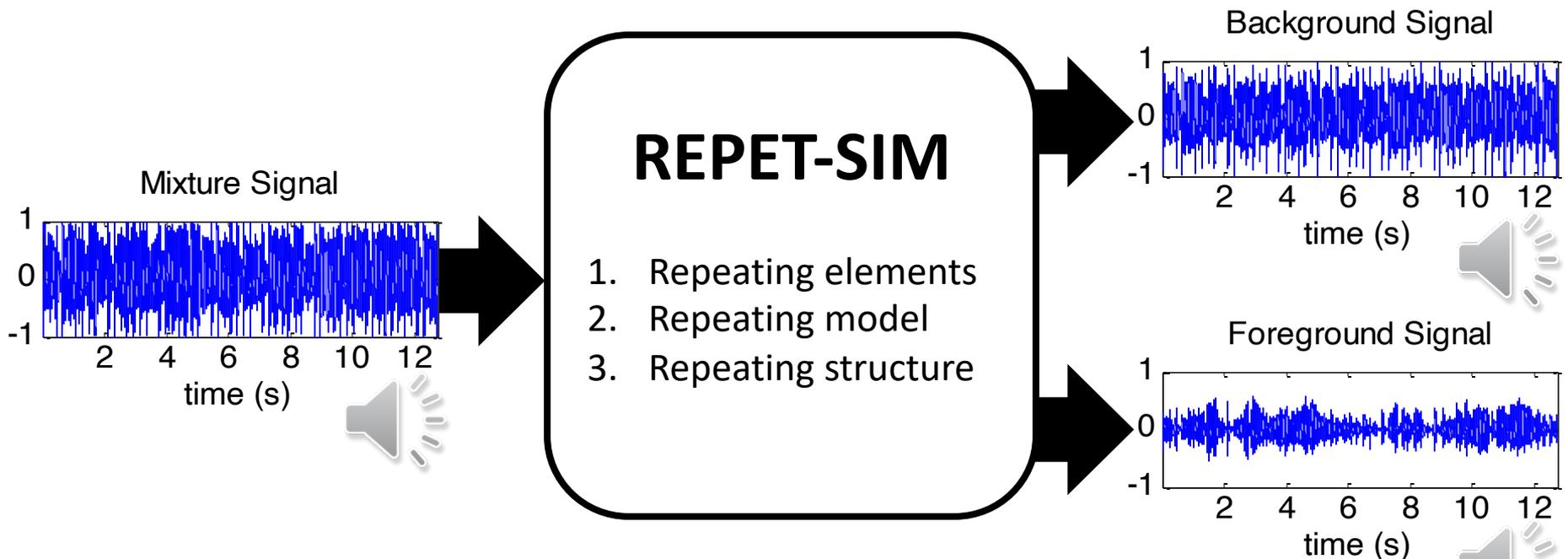
3. Repeating Structure

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



Music/Voice Separation

- Repeating background \approx **music component**
- Non-repeating foreground \approx **voice component**

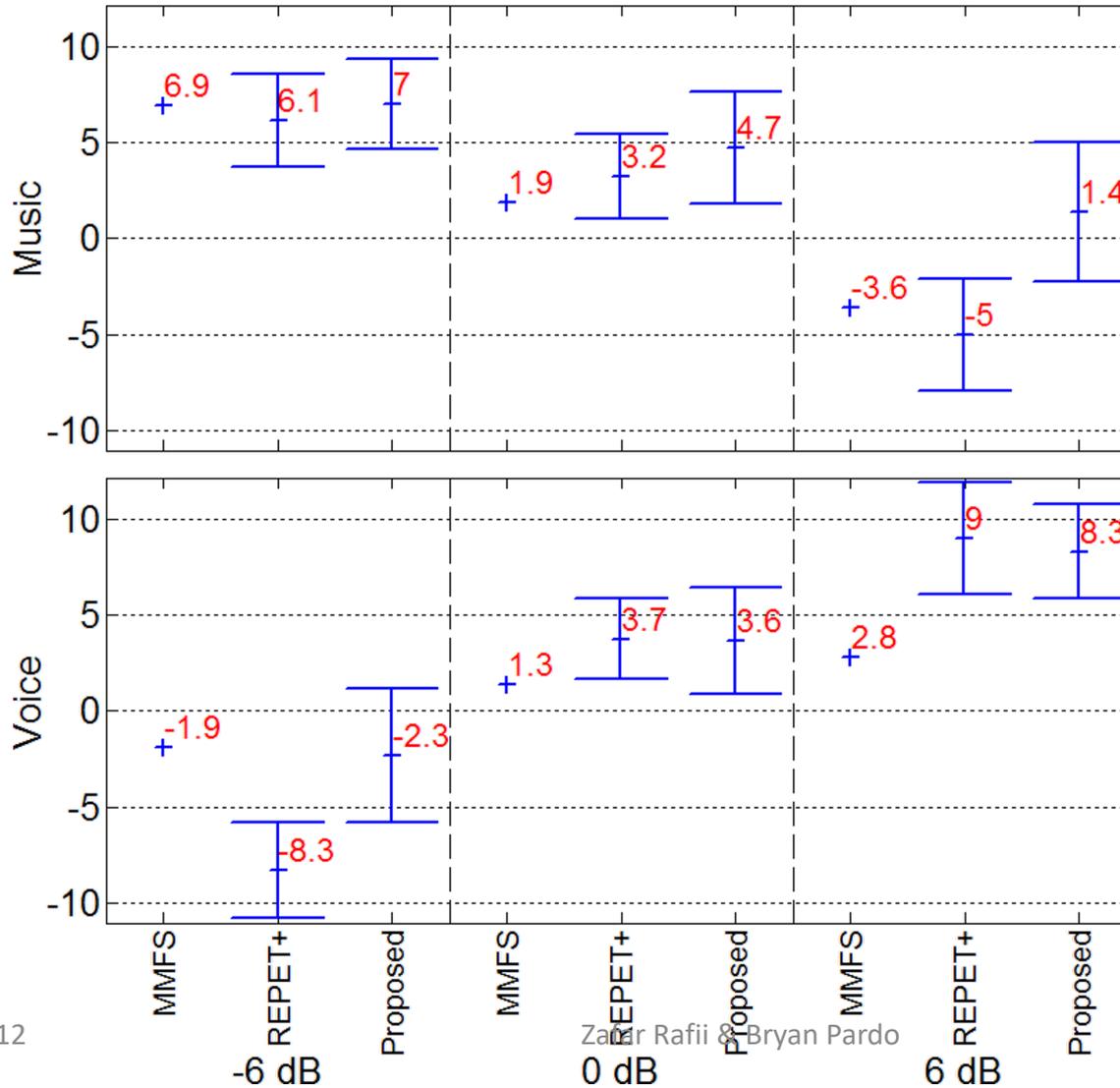


Evaluation

- **Competitive method 1** [Liutkus et al., 2012]
 - Adaptive REPET with automatic periods finder and soft time-frequency masking
- **Competitive method 2** [FitzGerald et al., 2010]
 - Median filtering of the spectrogram at different frequency resolutions to extract the vocals
- **Data set**
 - 14 full-track real-world songs (Beach Boys)
 - 3 voice-to-music mixing ratios (-6, 0, and 6 dB)

Evaluation

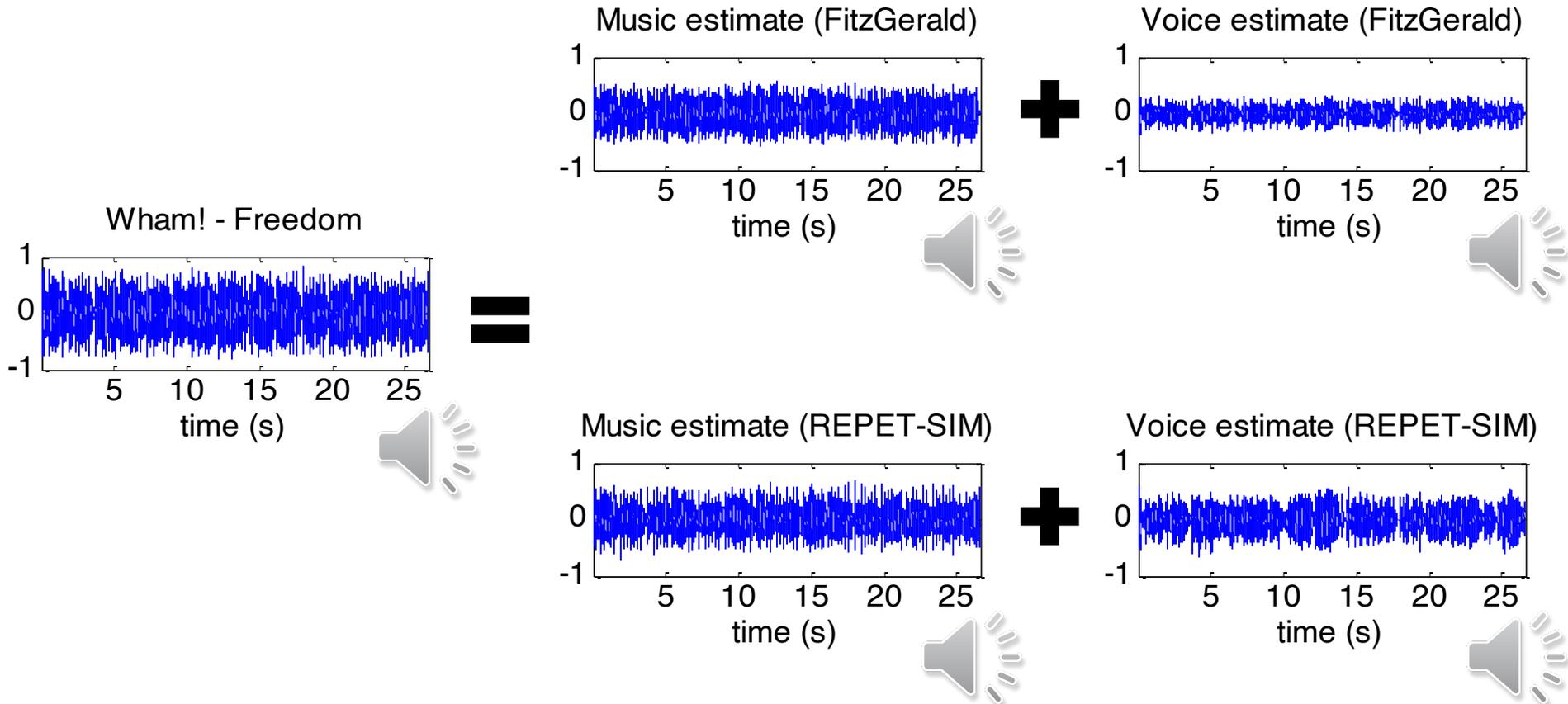
SDR (dB)



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

Examples

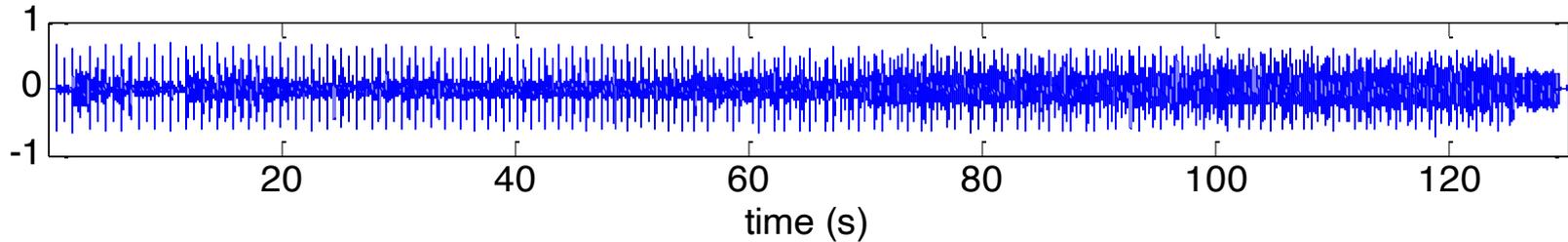
- REPET-SIM vs. FitzGerald et al.



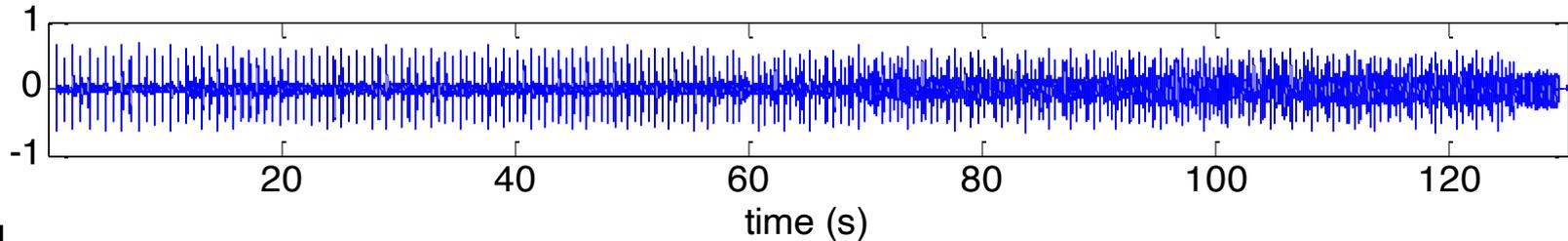
Examples

- REPET-SIM

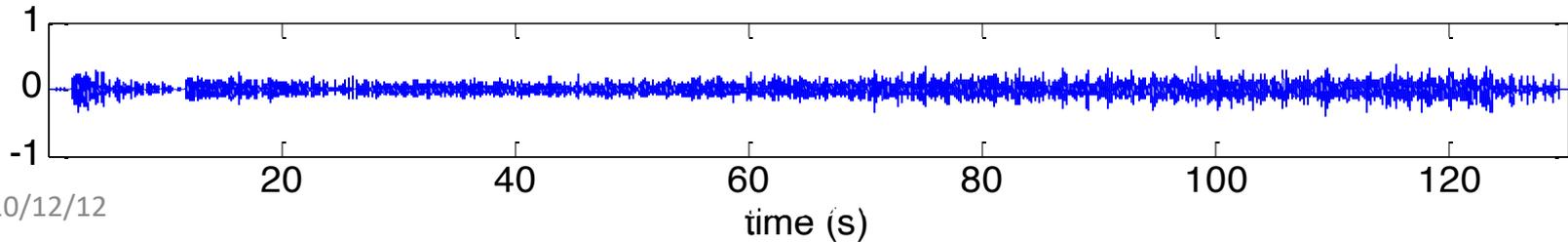
Blackalicious - Alphabet Aerobics



Music estimate

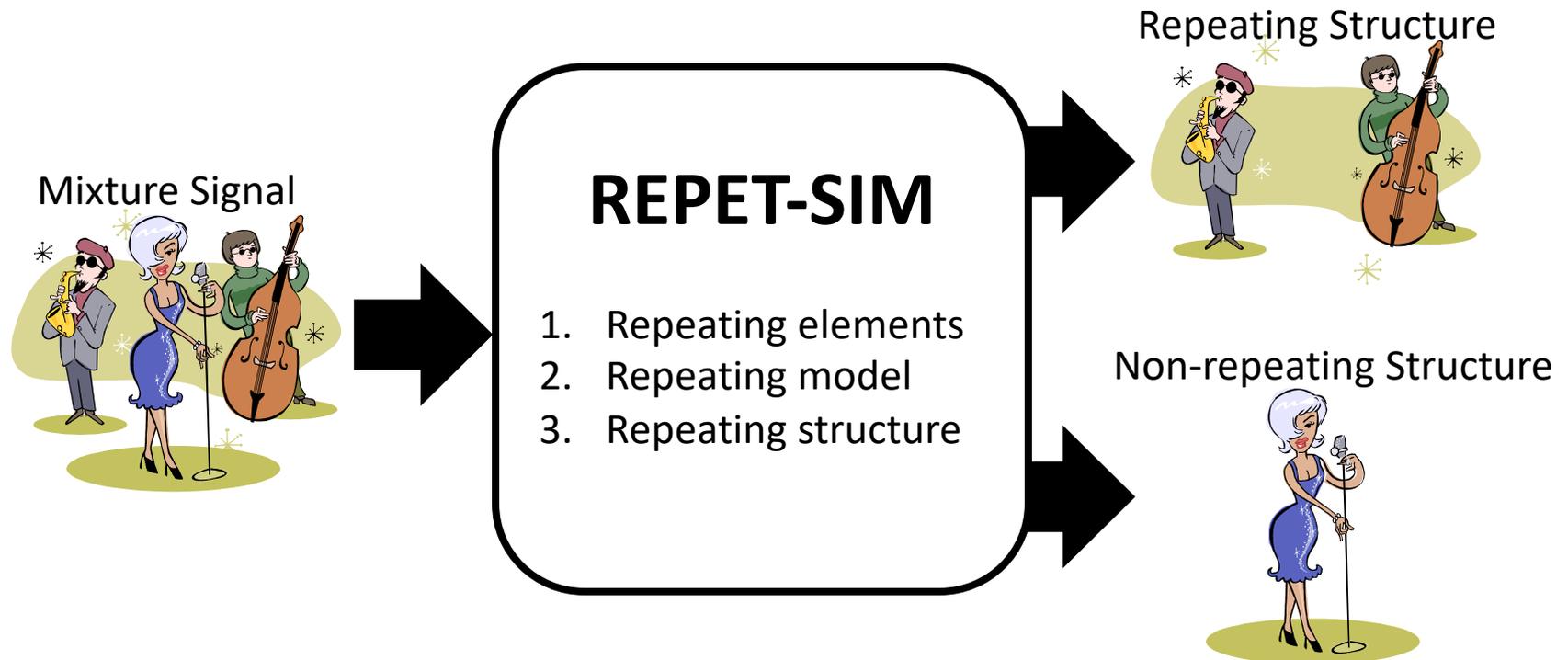


Voice estimate



Conclusion

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



Questions?

- 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.
- **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.**
- **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.**