Music/Voice Separation using the Similarity Matrix

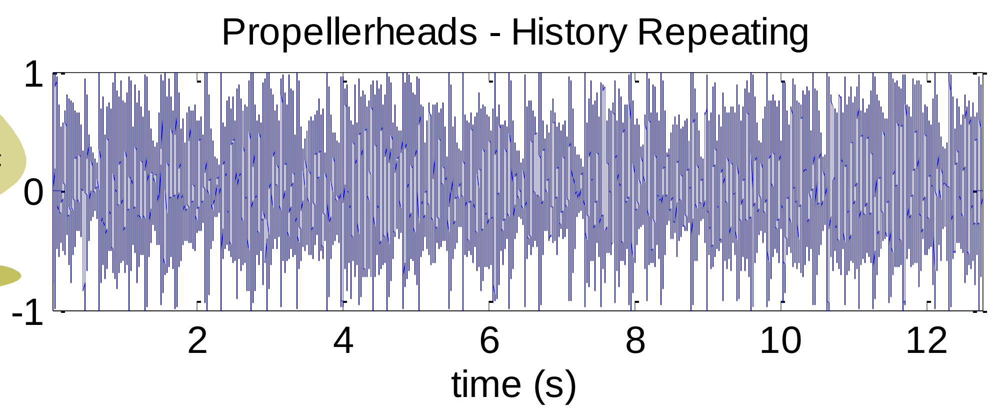
Zafar Rafii & Bryan Pardo

NORTHWESTERN UNIVERSITY

interactive audio lab

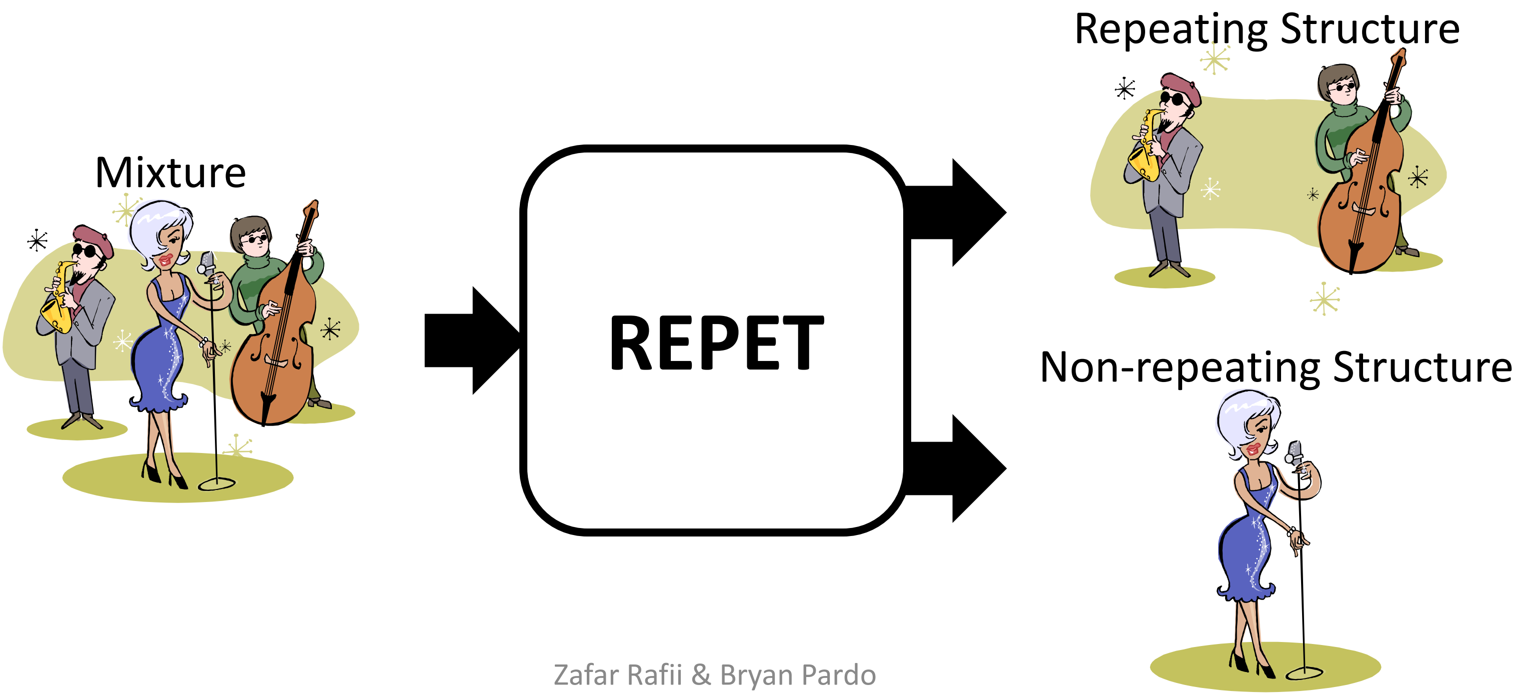
Introduction

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

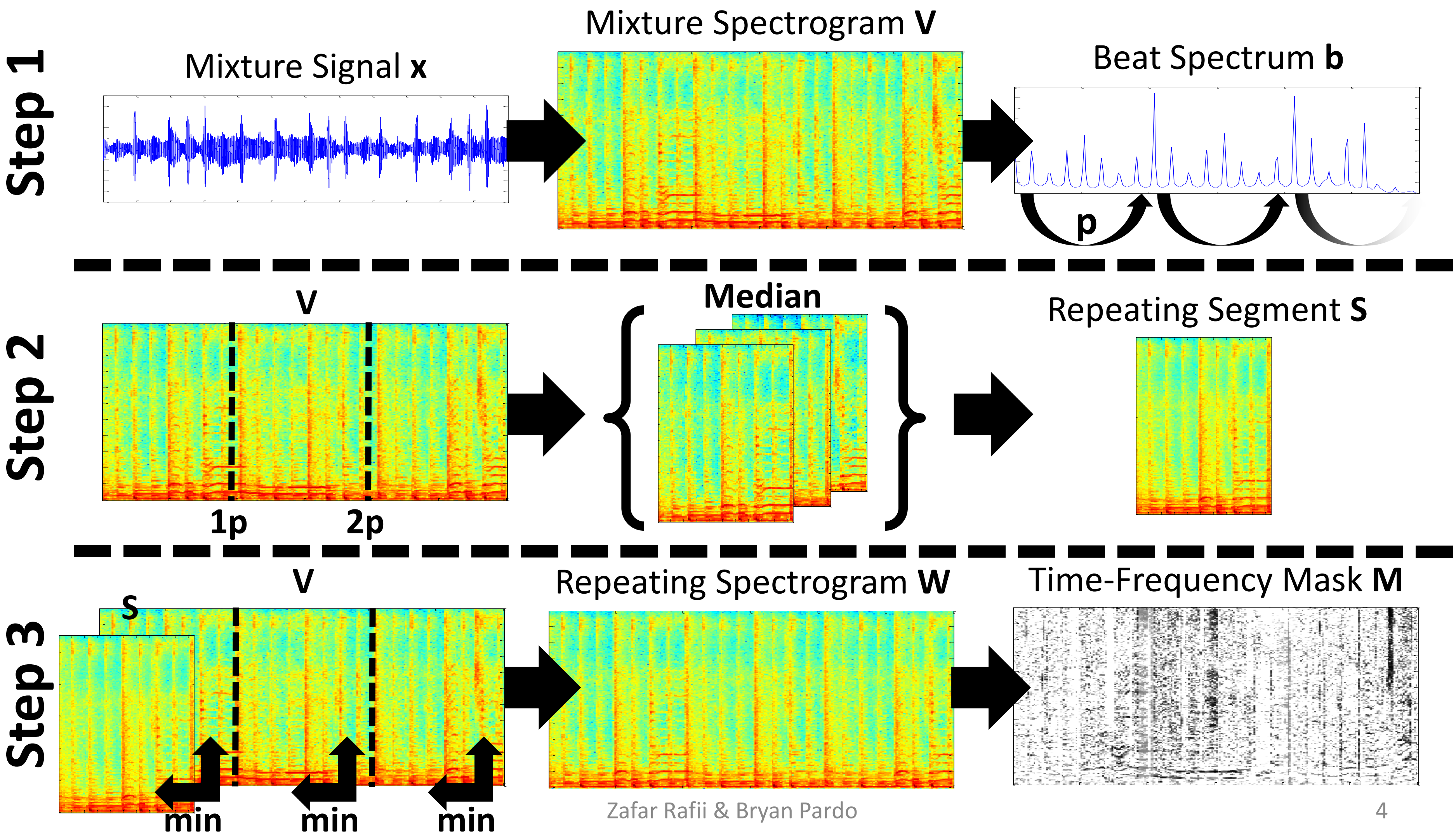


Introduction

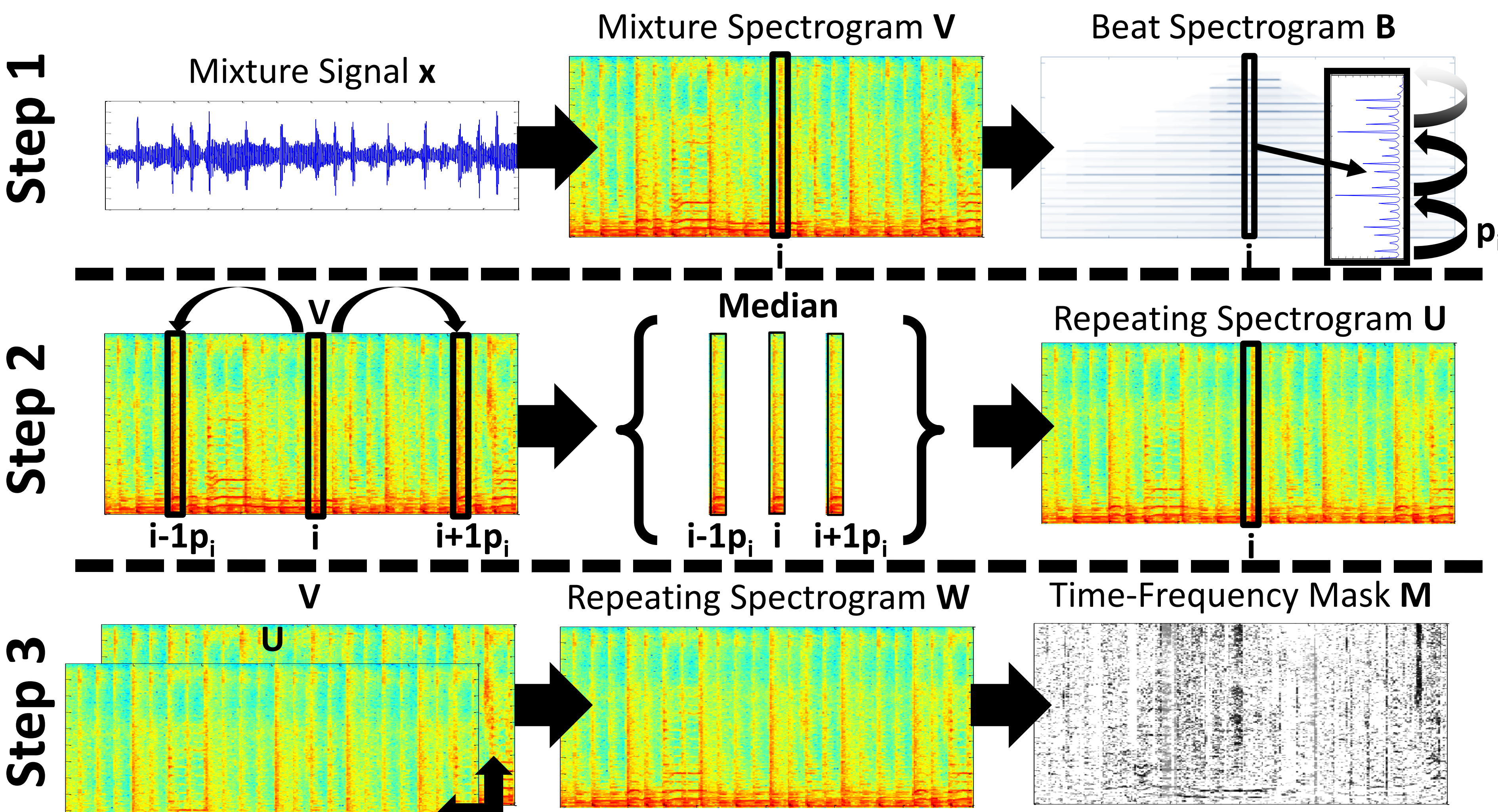
- The **REpeating Pattern Extraction Technique (REPET)** was proposed to extract the repeating structure from the non-repeating structure



REPET

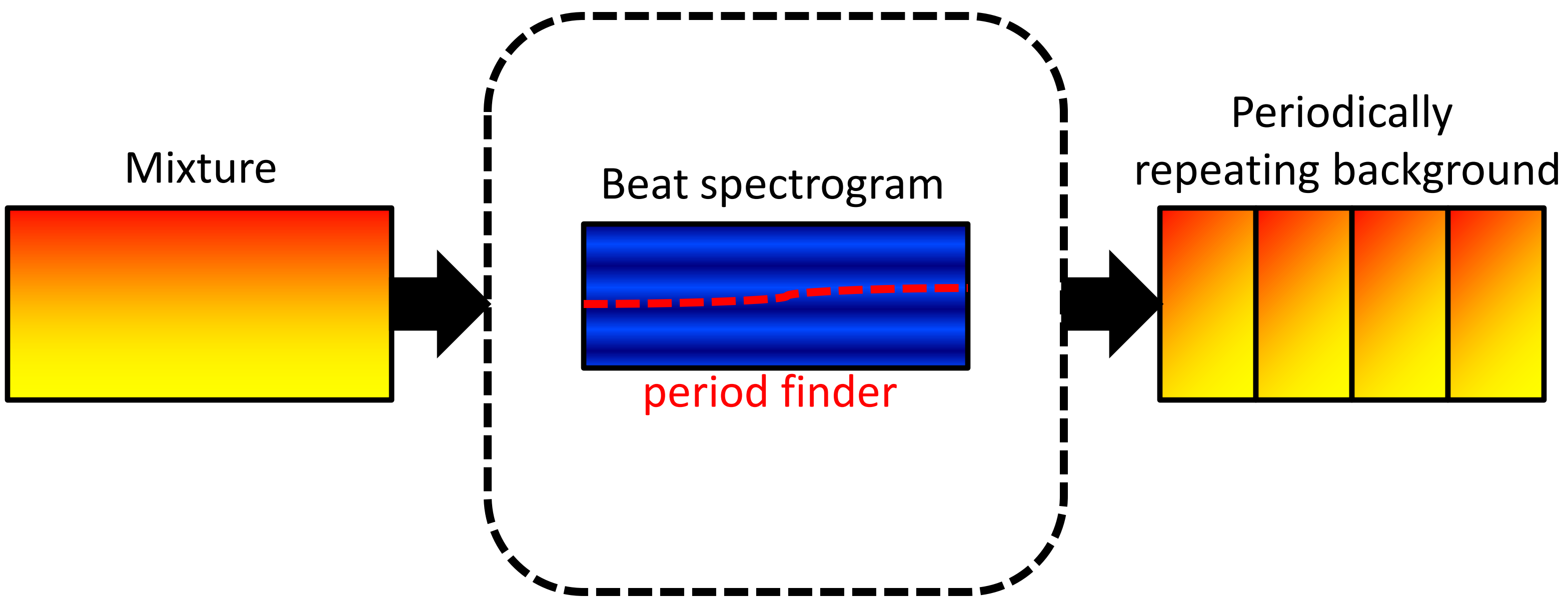

Step 1
Mixture Signal **x**
Mixture Spectrogram **V**
Beat Spectrum **b**
p
Step 2
V
1p
2p
Median
Repeating Segment **S**
Step 3
V
S
min
min
min
Repeating Spectrogram **W**
Time-Frequency Mask **M**

Adaptive REPET



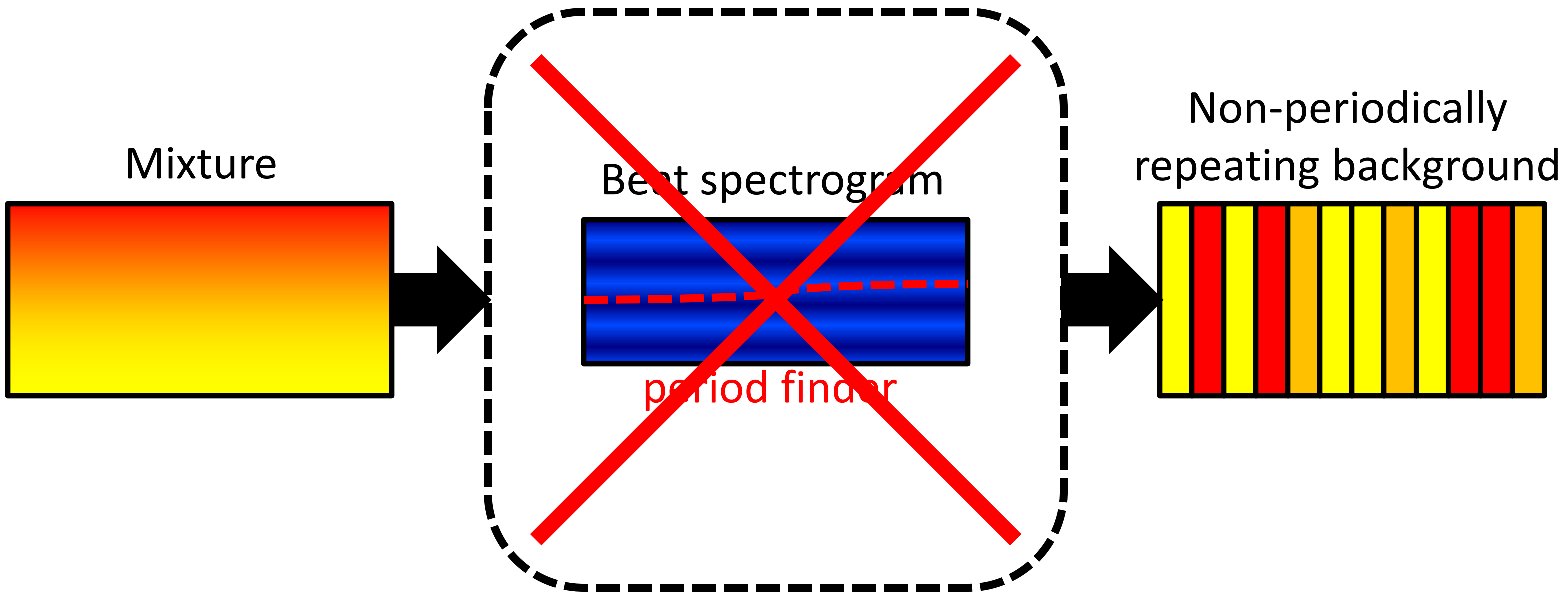
Limitations

- Both the original and the adaptive REPET assume **periodically repeating patterns**



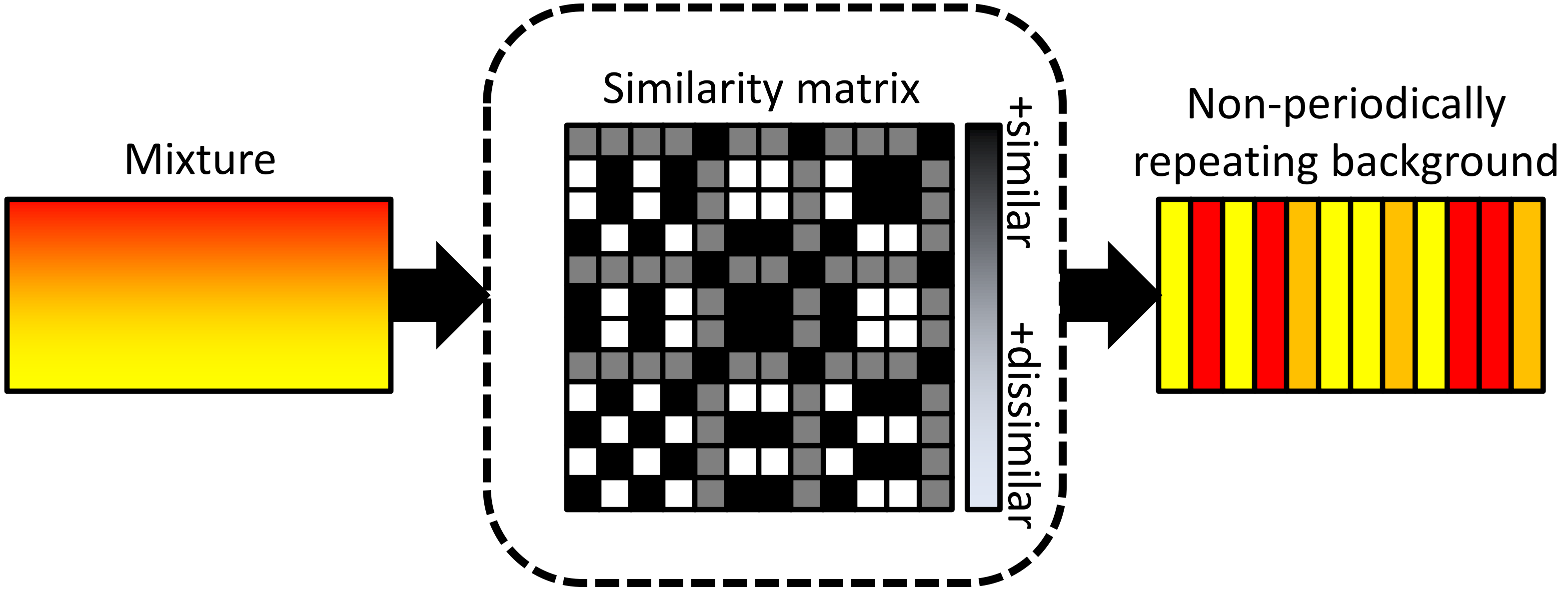
Limitations

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



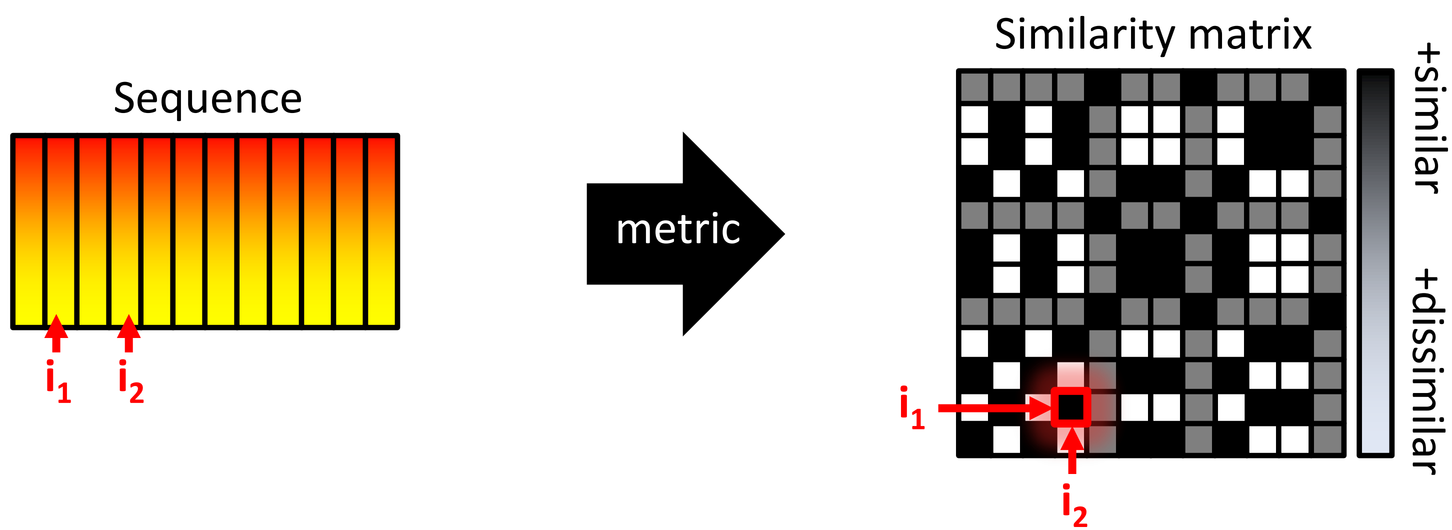
Limitations

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



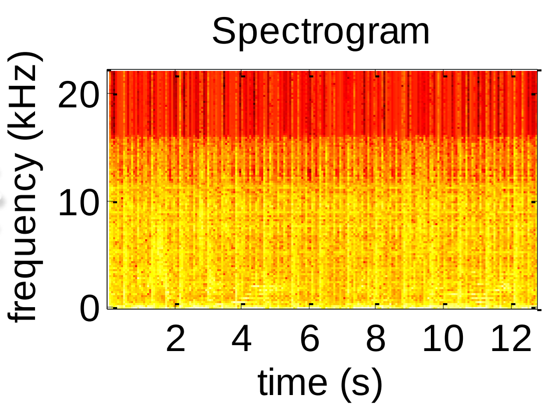
Similarity Matrix

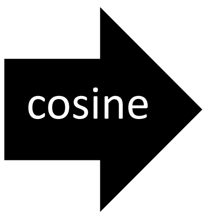
- The **similarity matrix** is a matrix where each bin measures the (dis)similarity between any two elements of a sequence given a metric

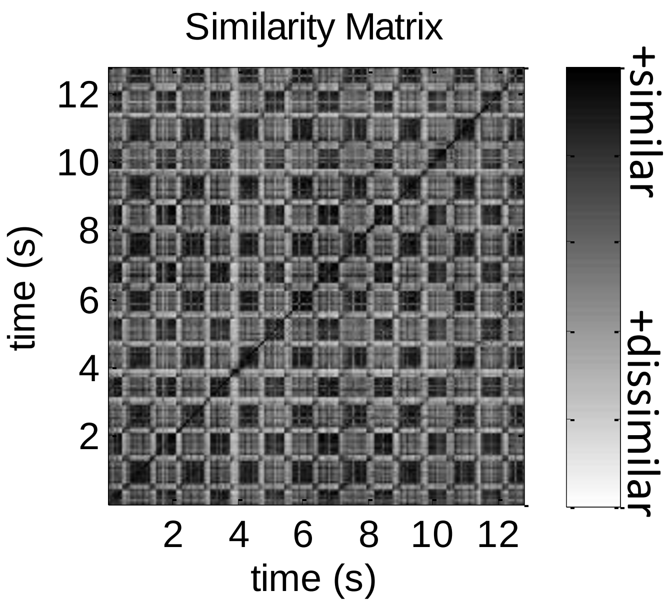


Similarity Matrix

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

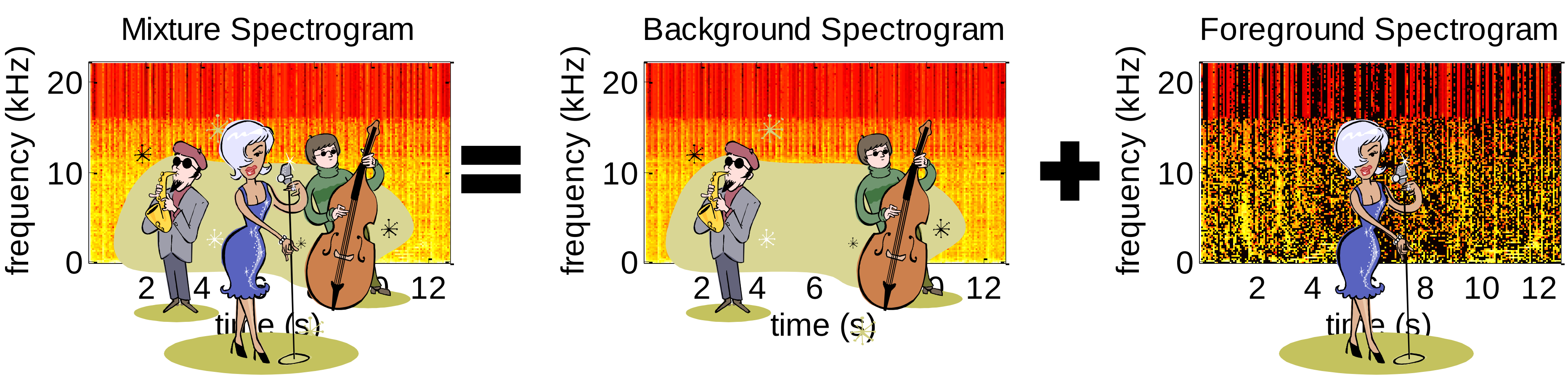






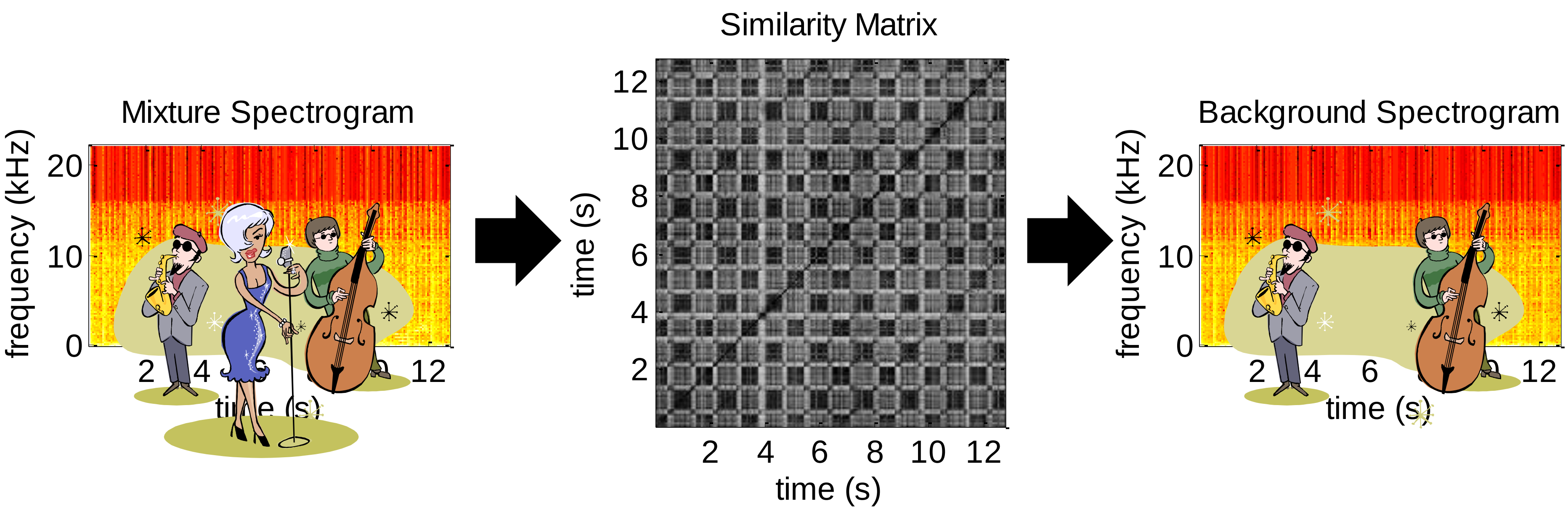
Assumptions

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



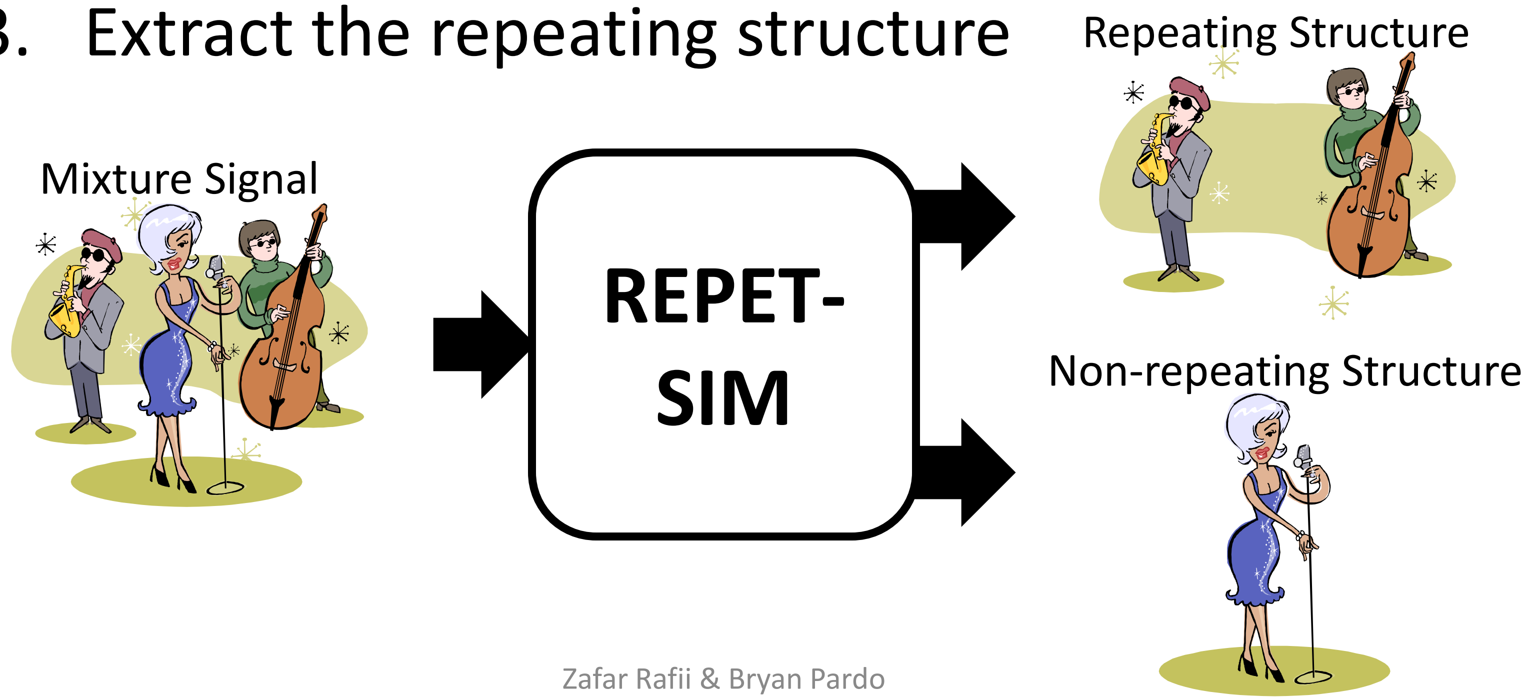
Assumptions

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



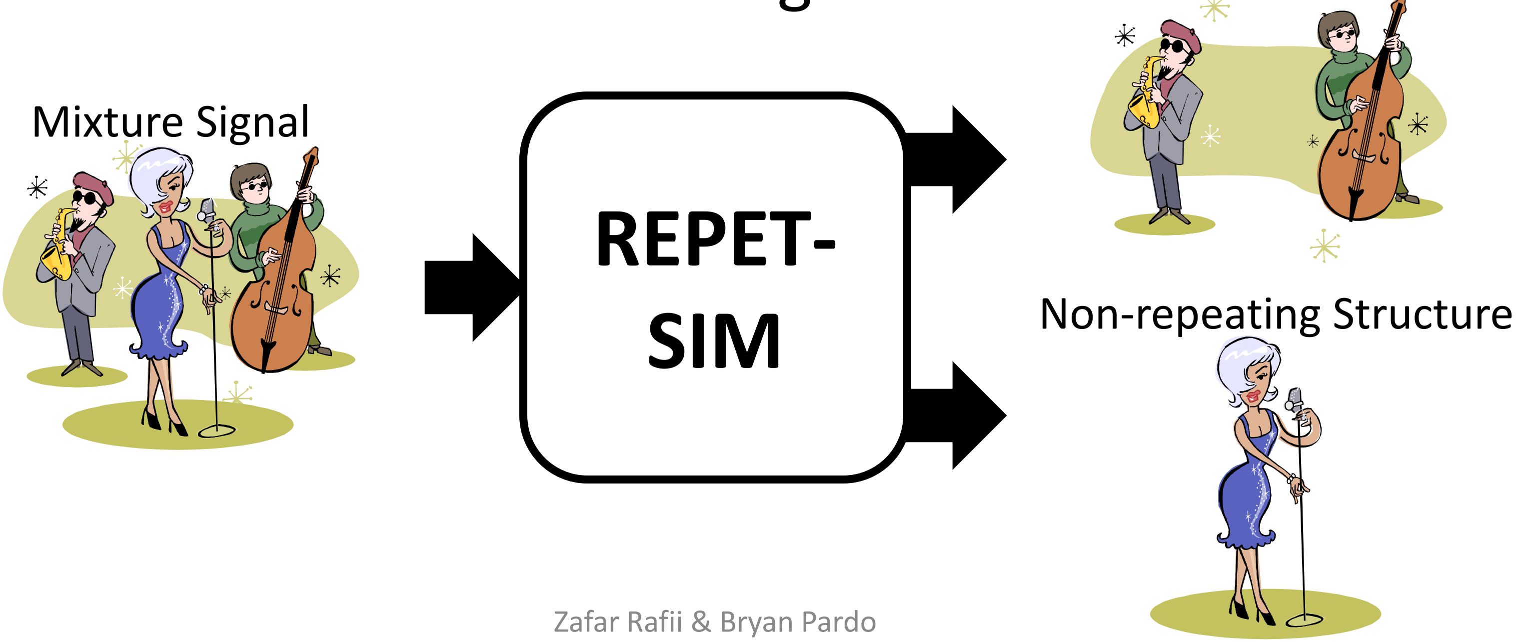
REPET-SIM

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



REPET-SIM

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

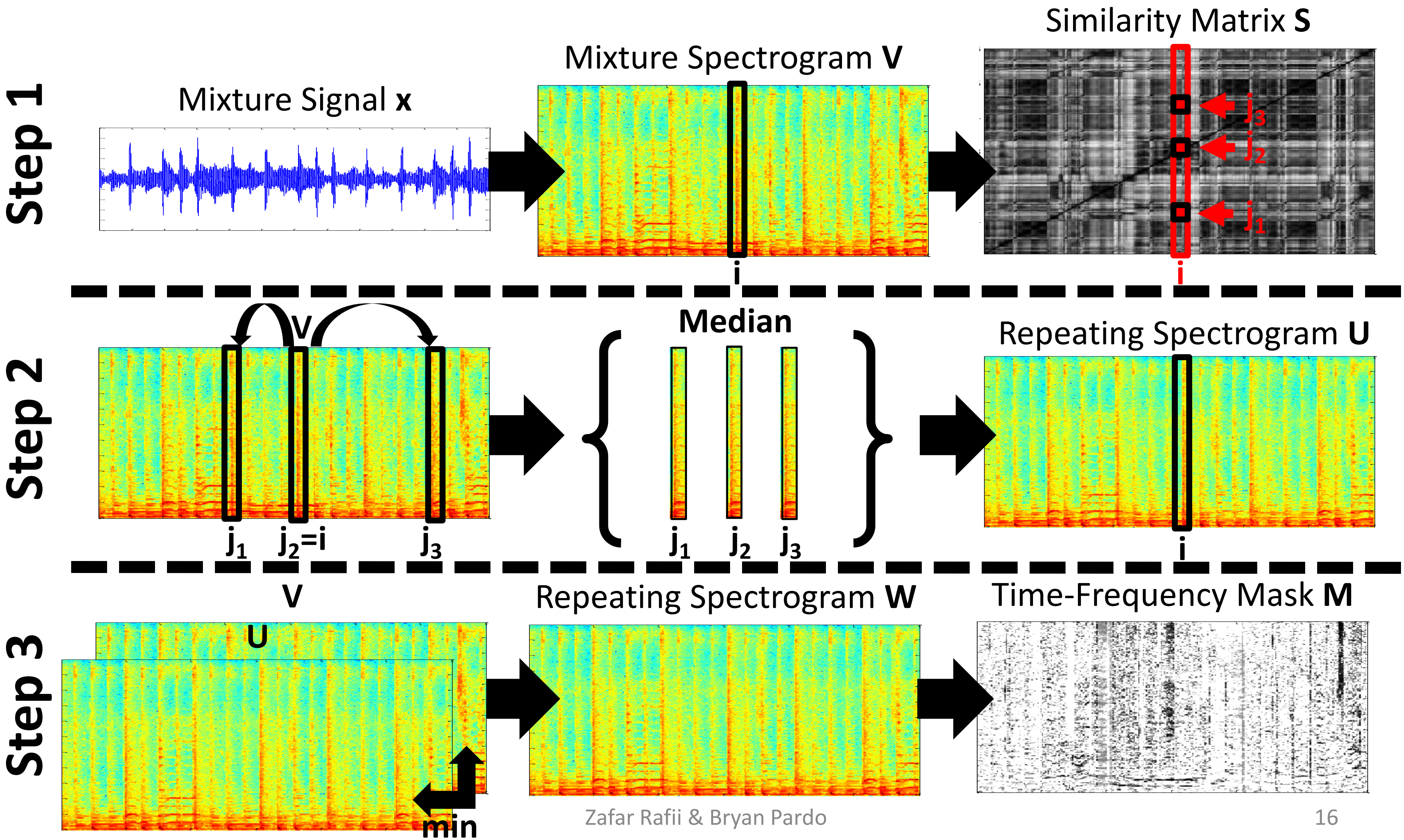


Interests

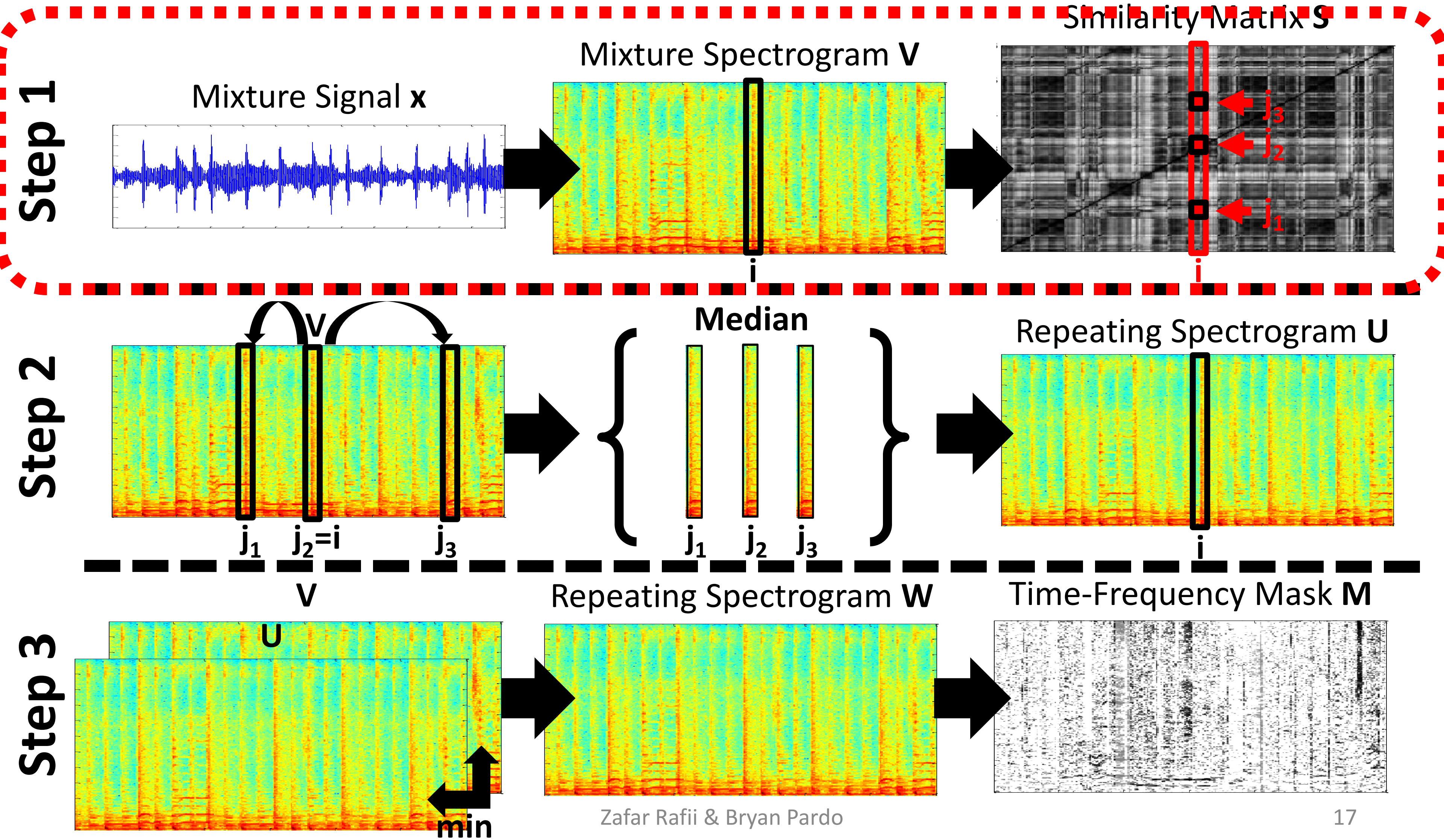
- **Practical Interests**
  - Audio post processing
  - Melody extraction
  - Karaoke gaming

- **Intellectual Interests**
  - Music perception
  - Music understanding
  - Simply based on self-similarity!

REPET-SIM

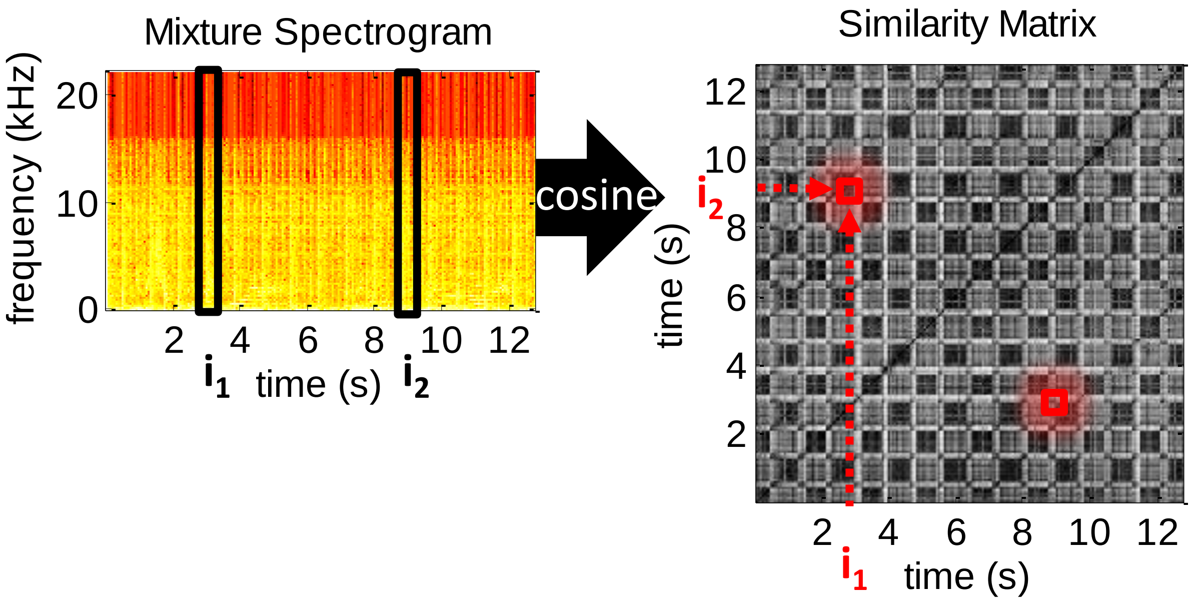

Step 1
Mixture Signal x
Mixture Spectrogram V
Similarity Matrix S
j3
j2
j1
i
i
Step 2
V
Median
Repeating Spectrogram U
j1
j2=i
j3
j1
j2
j3
i
Step 3
V
U
min
Repeating Spectrogram W
Time-Frequency Mask M

1. Repeating Elements



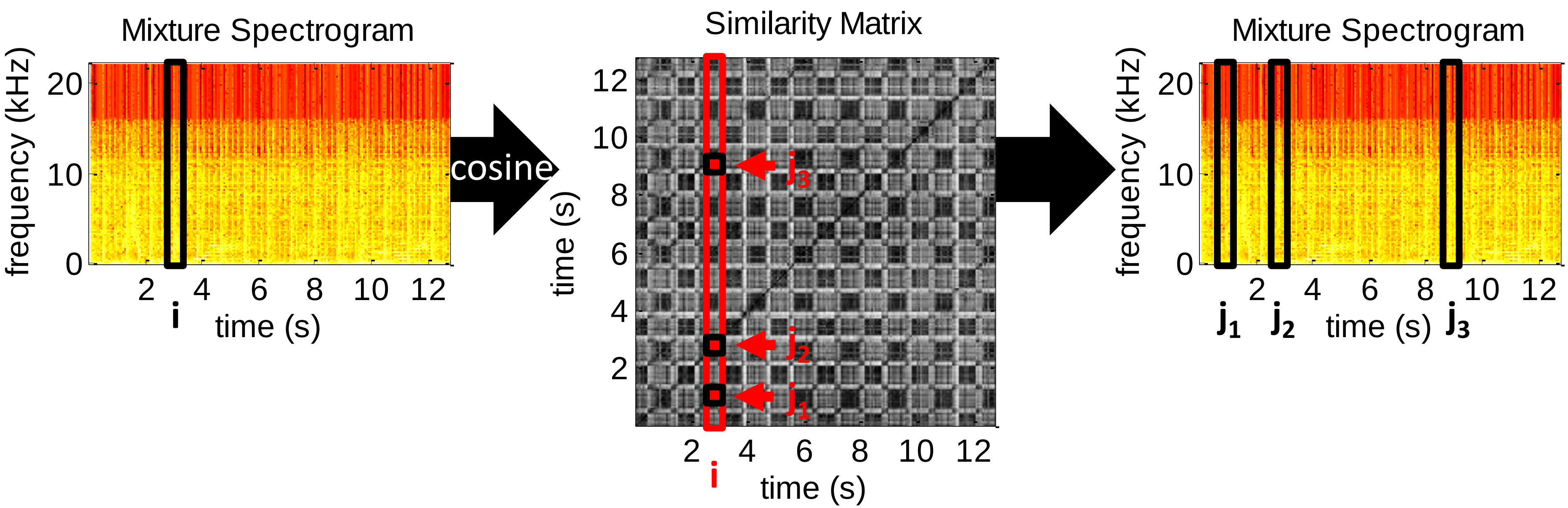
1. Repeating Elements

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

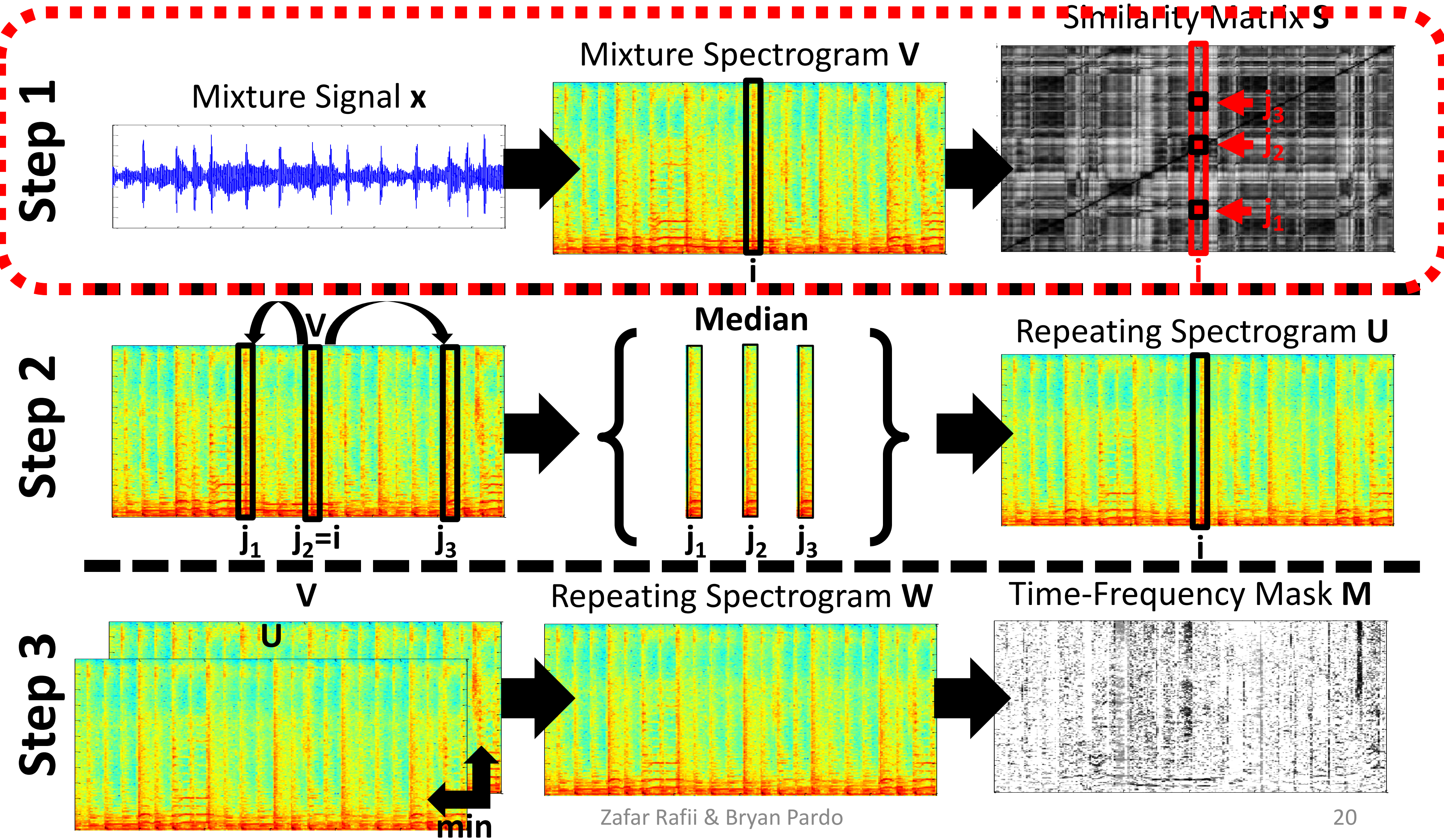


1. Repeating Elements

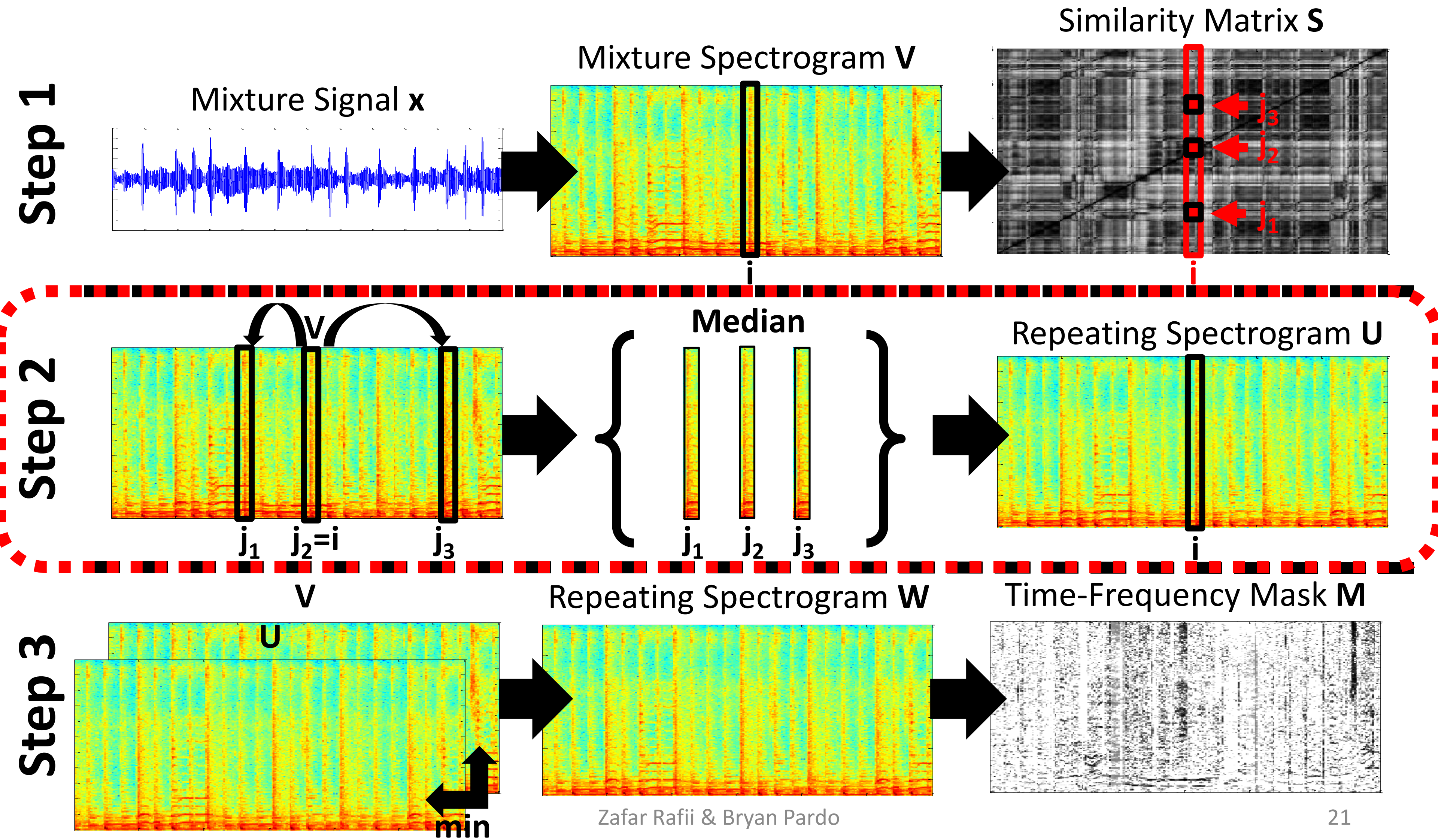
- The SM reveals for every frame i, the frames j_k that are **the most similar** to frame i



1. Repeating Elements

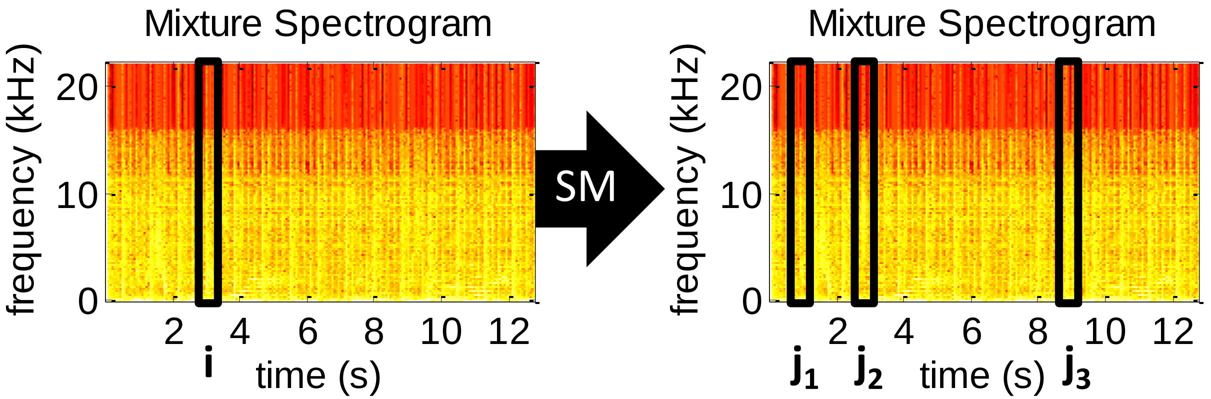


2. Repeating Model



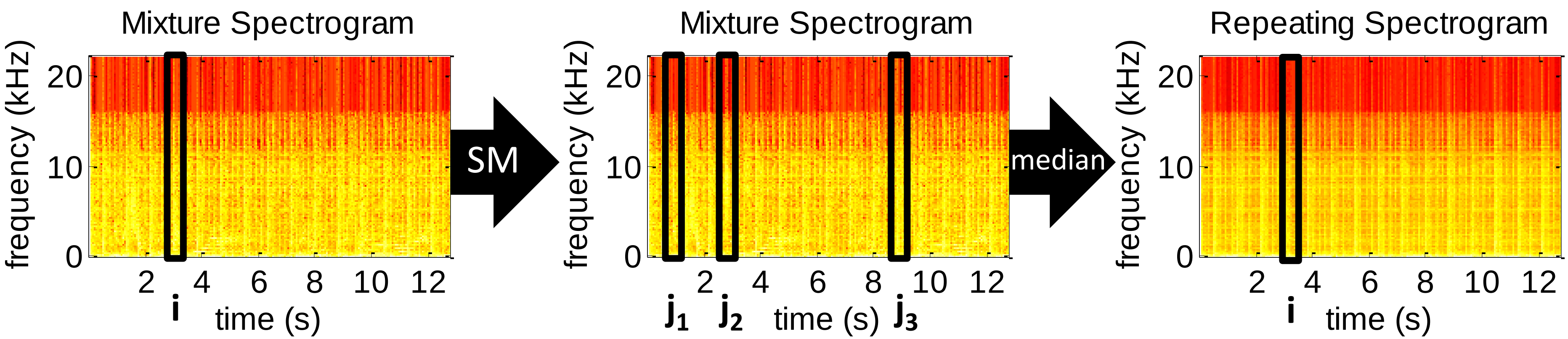
2. Repeating Model

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

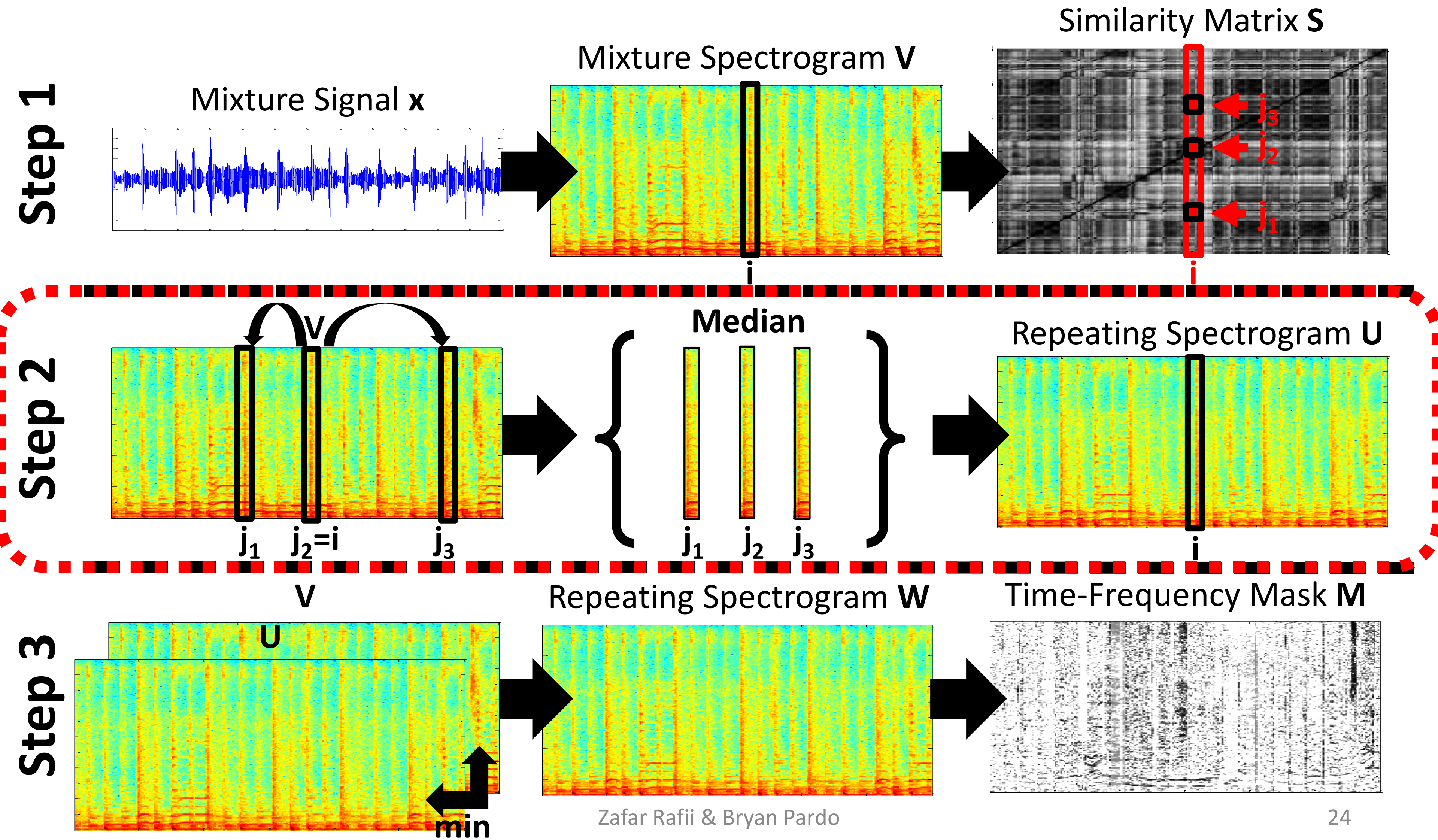


2. Repeating Model

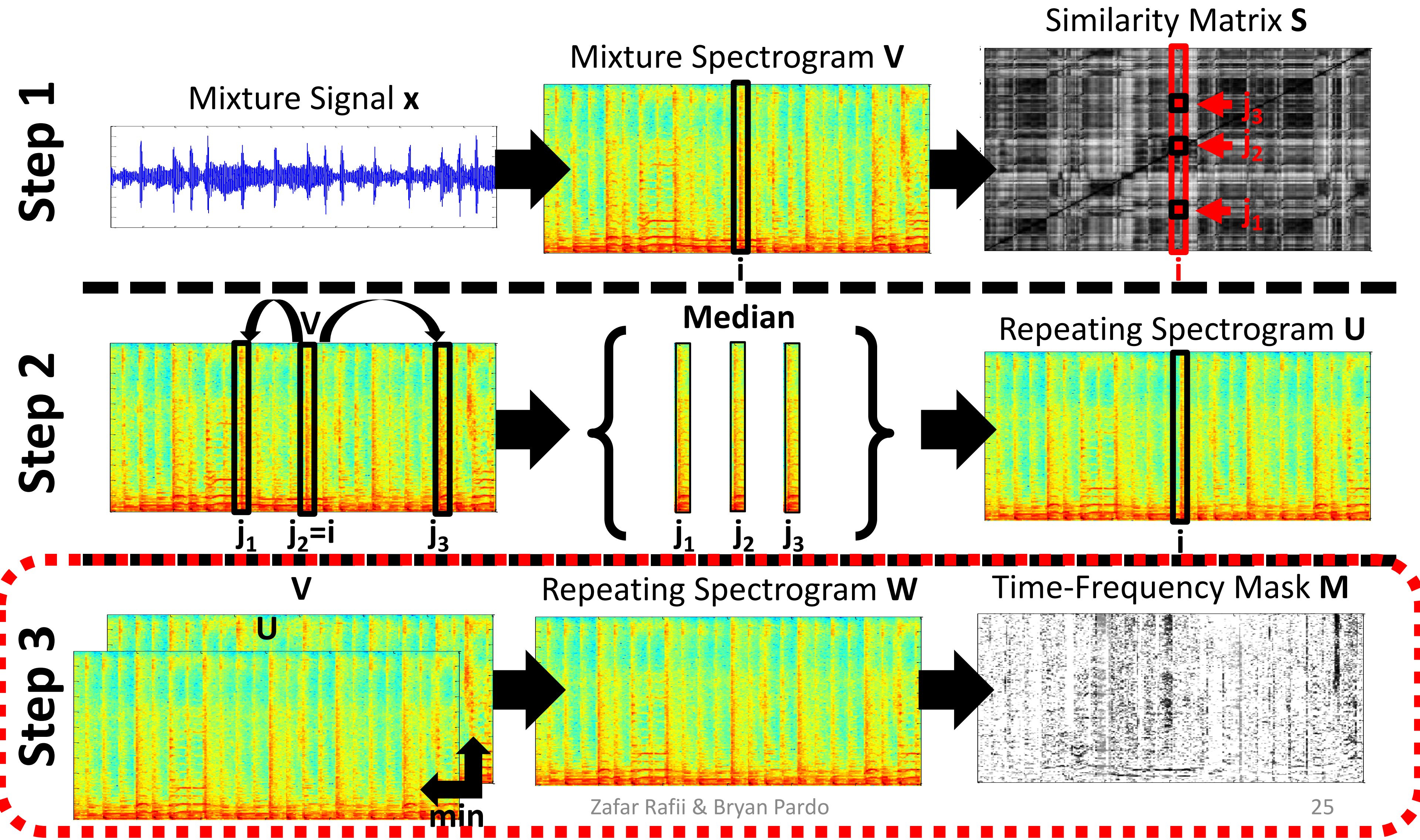
- We obtain an initial **repeating spectrogram model**



2. Repeating Model

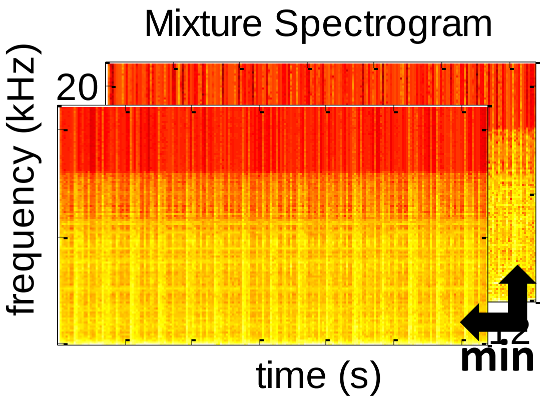


3. Repeating Structure



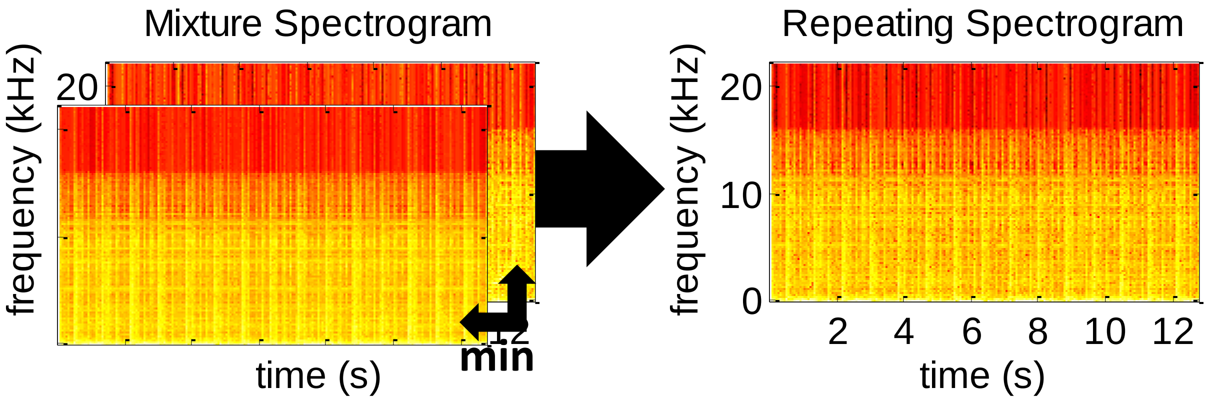
3. Repeating Structure

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



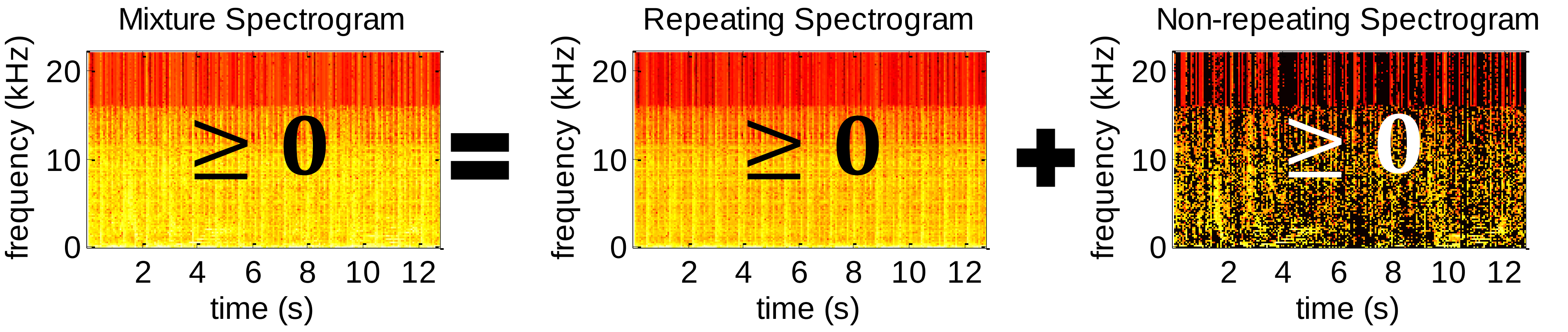
3. Repeating Structure

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



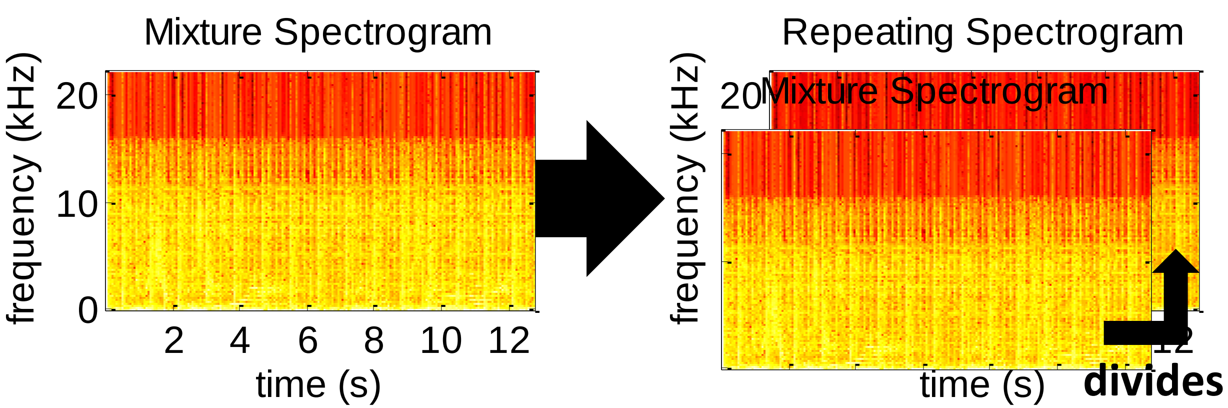
3. Repeating Structure

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



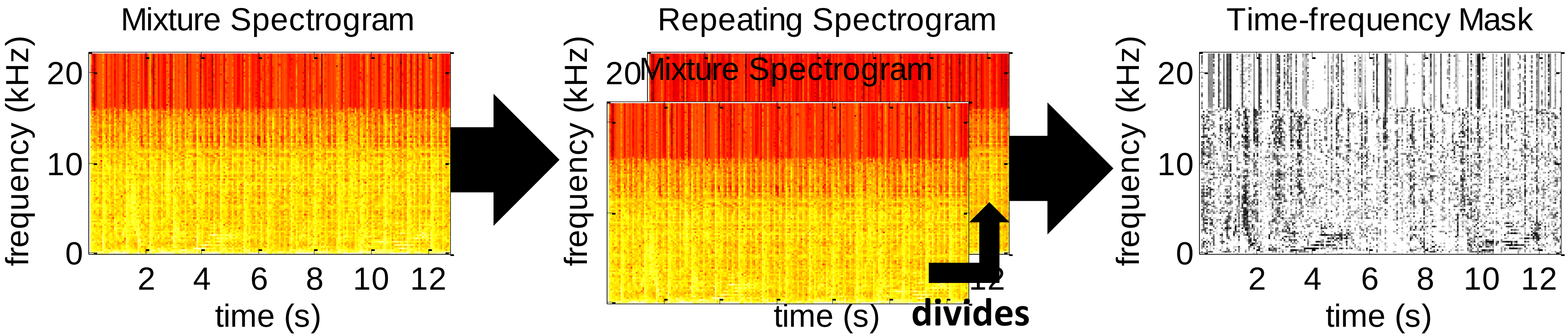
3. Repeating Structure

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



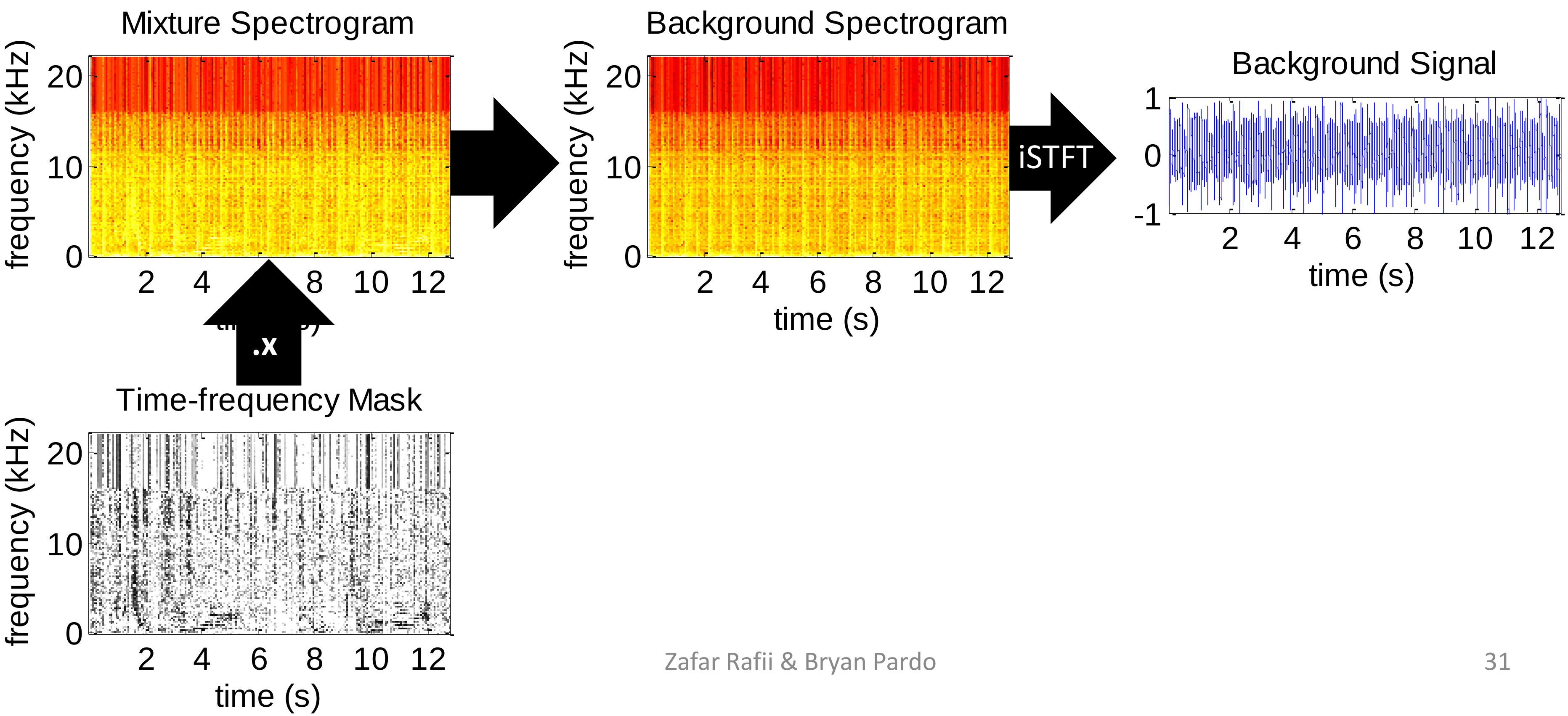
3. Repeating Structure

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



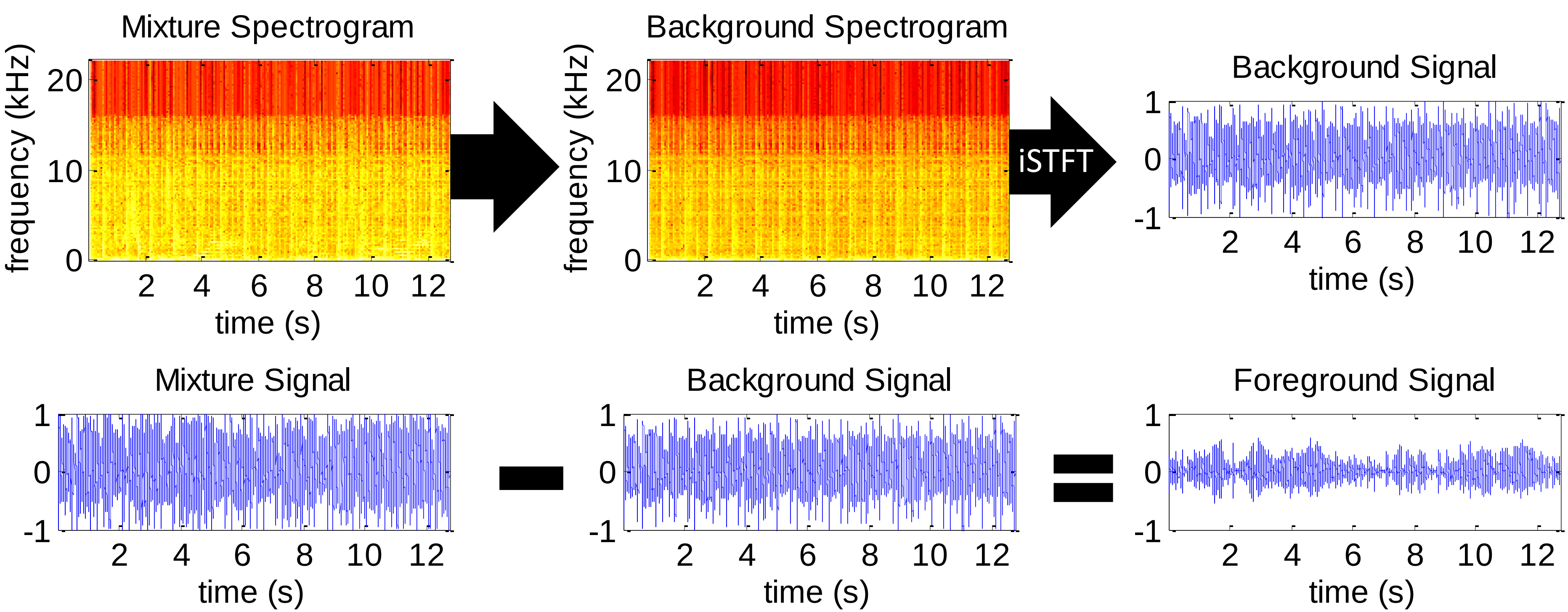
3. Repeating Structure

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



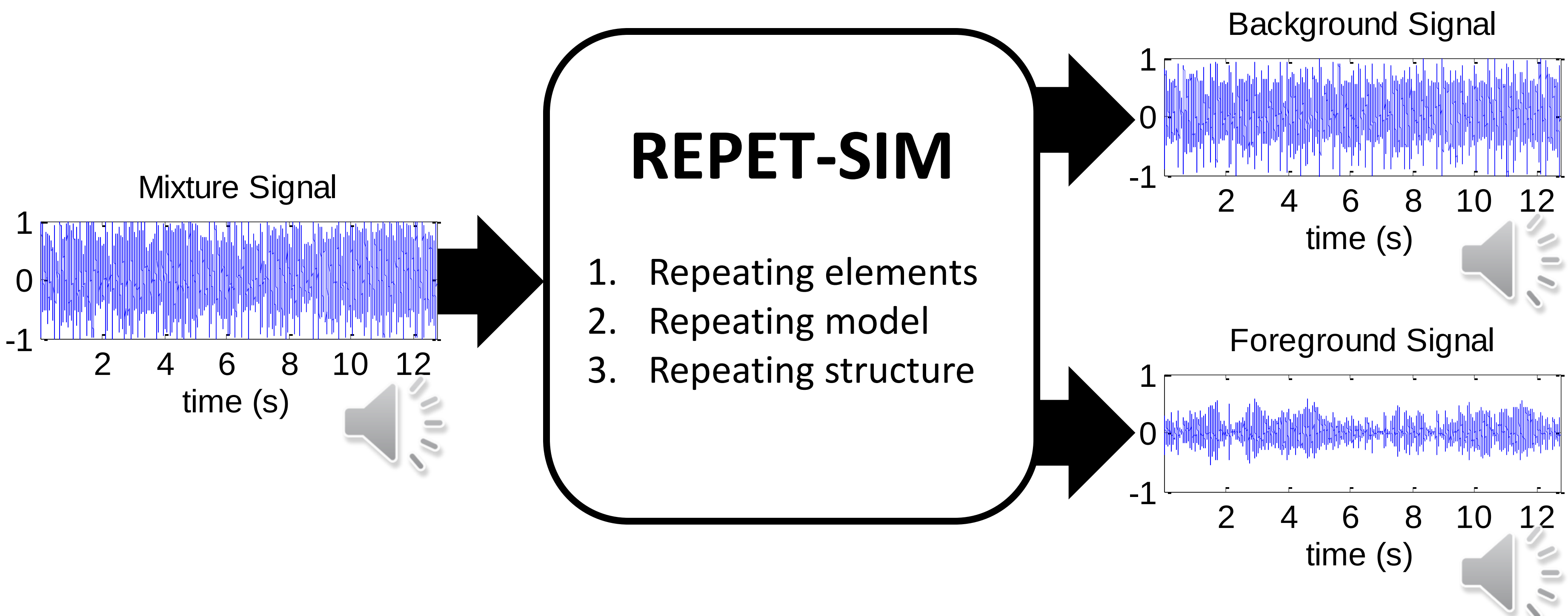
3. Repeating Structure

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



Music/Voice Separation

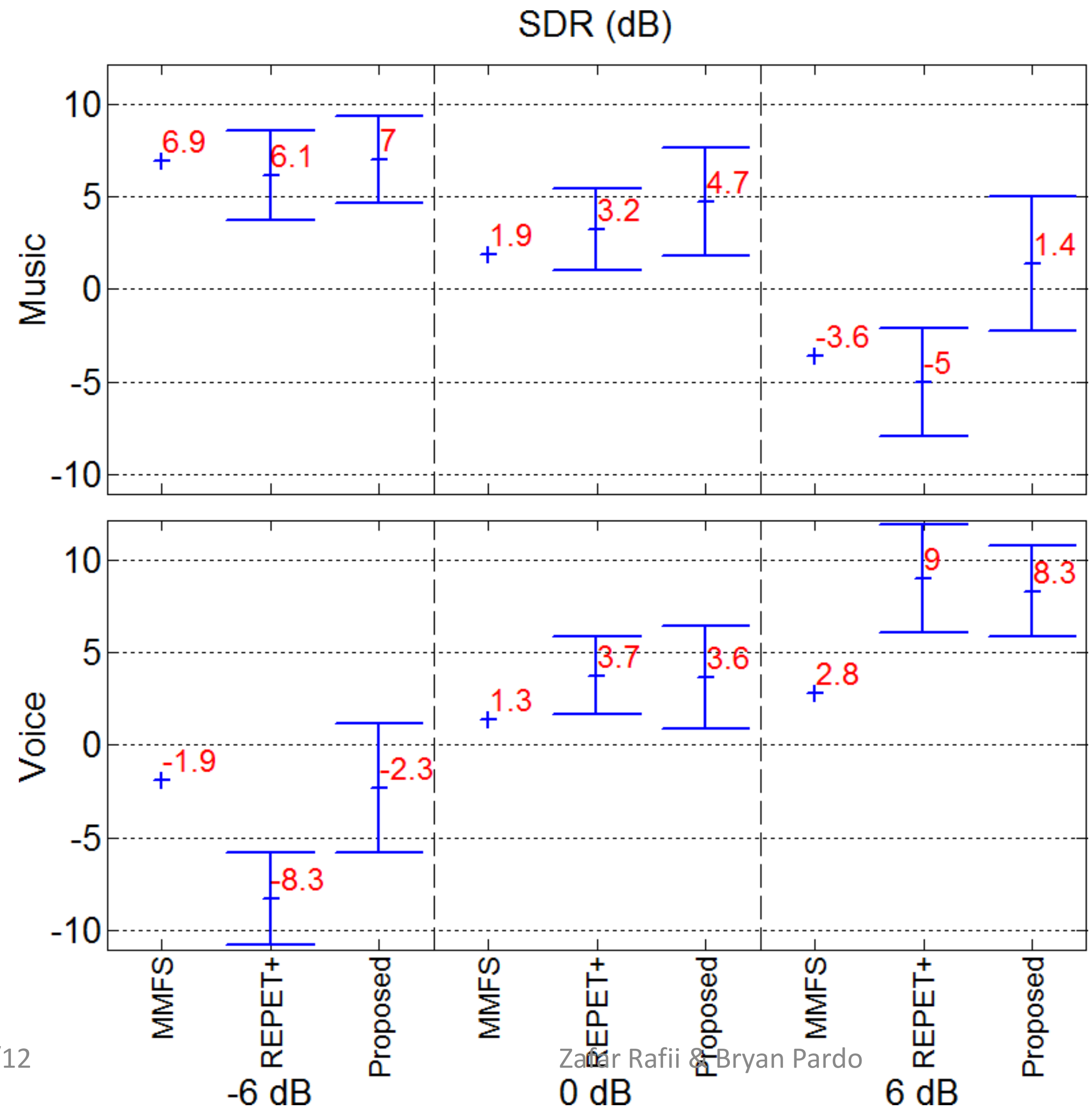
- Repeating background ≈ **music component**
- Non-repeating foreground ≈ **voice component**



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)

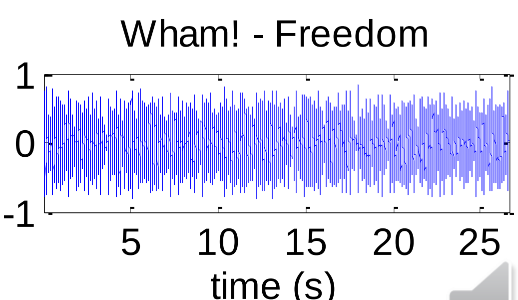
Evaluation

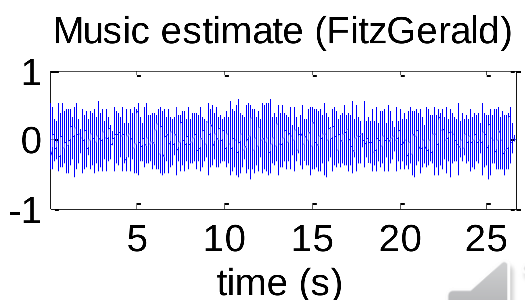


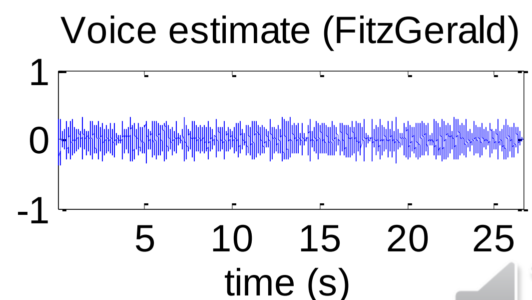
MMFS = FitzGerald et al.
REPET+ = Liutkus et al.
Proposed = REPET-SIM

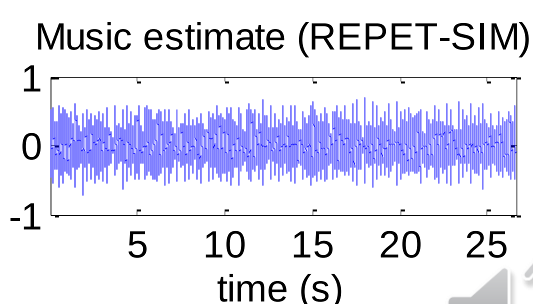
Examples

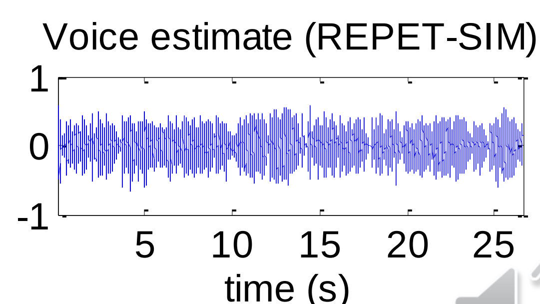
- REPET-SIM vs. FitzGerald et al.





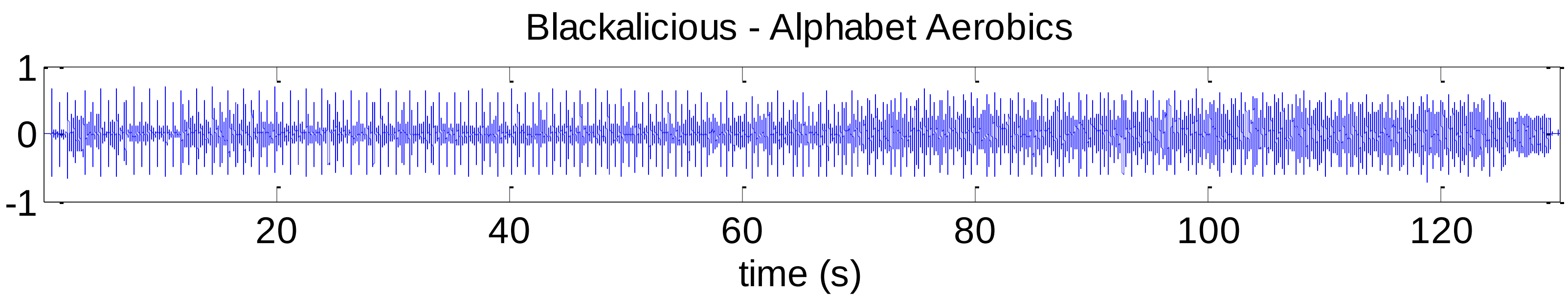


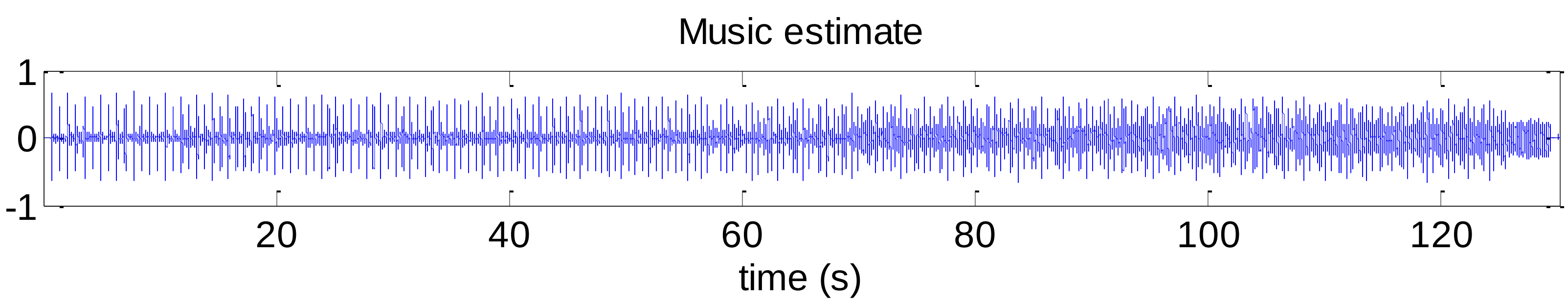


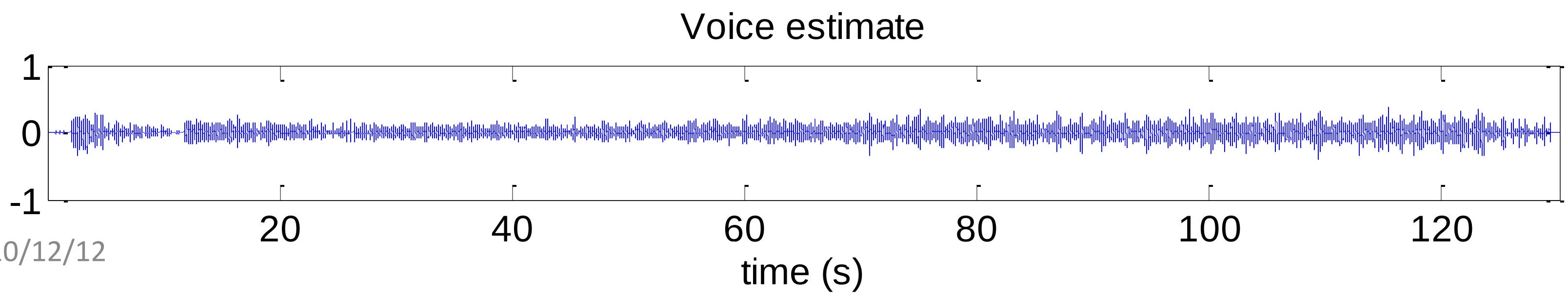


Examples

- REPET-SIM

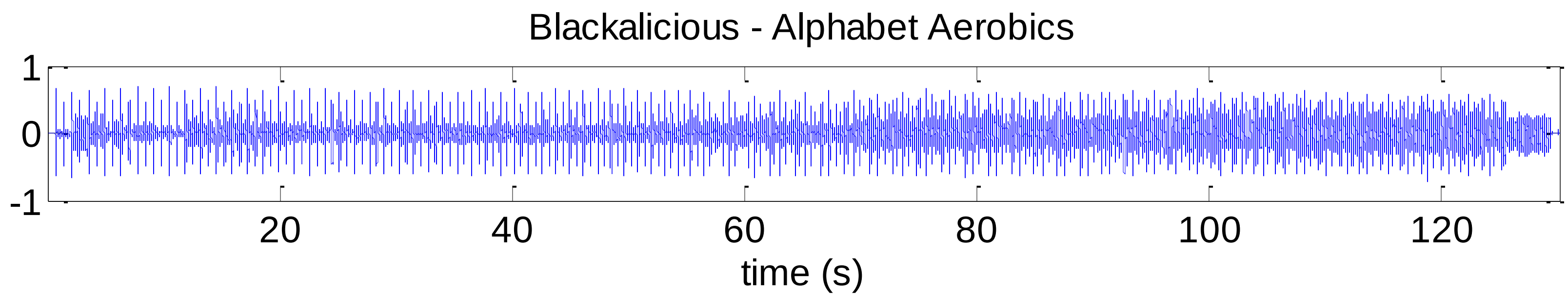




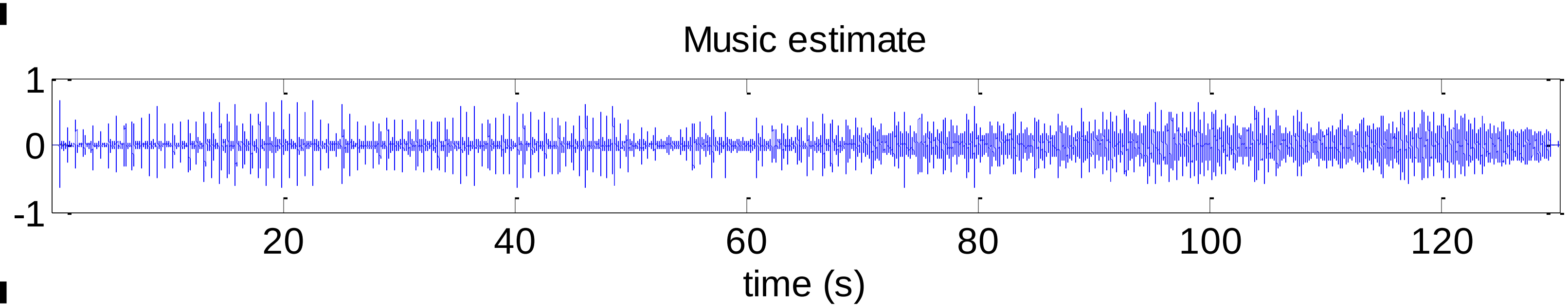


Examples

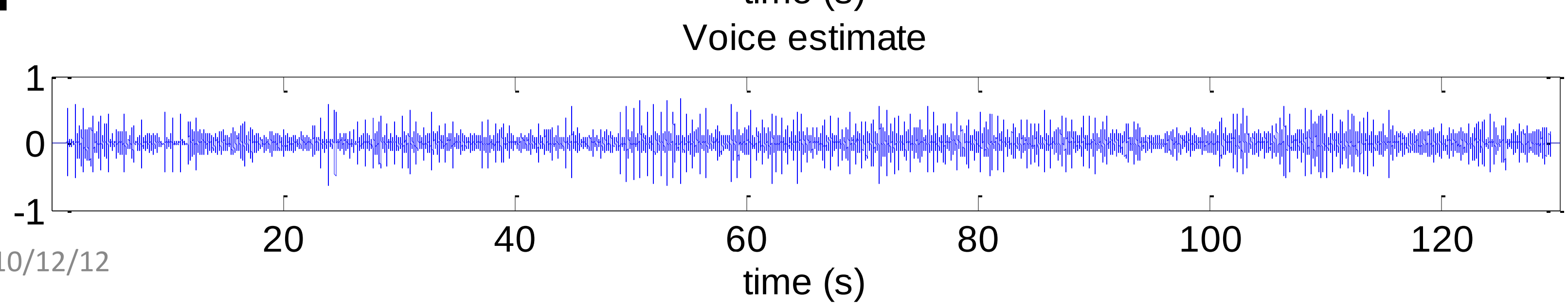
- Adaptive REPET



=

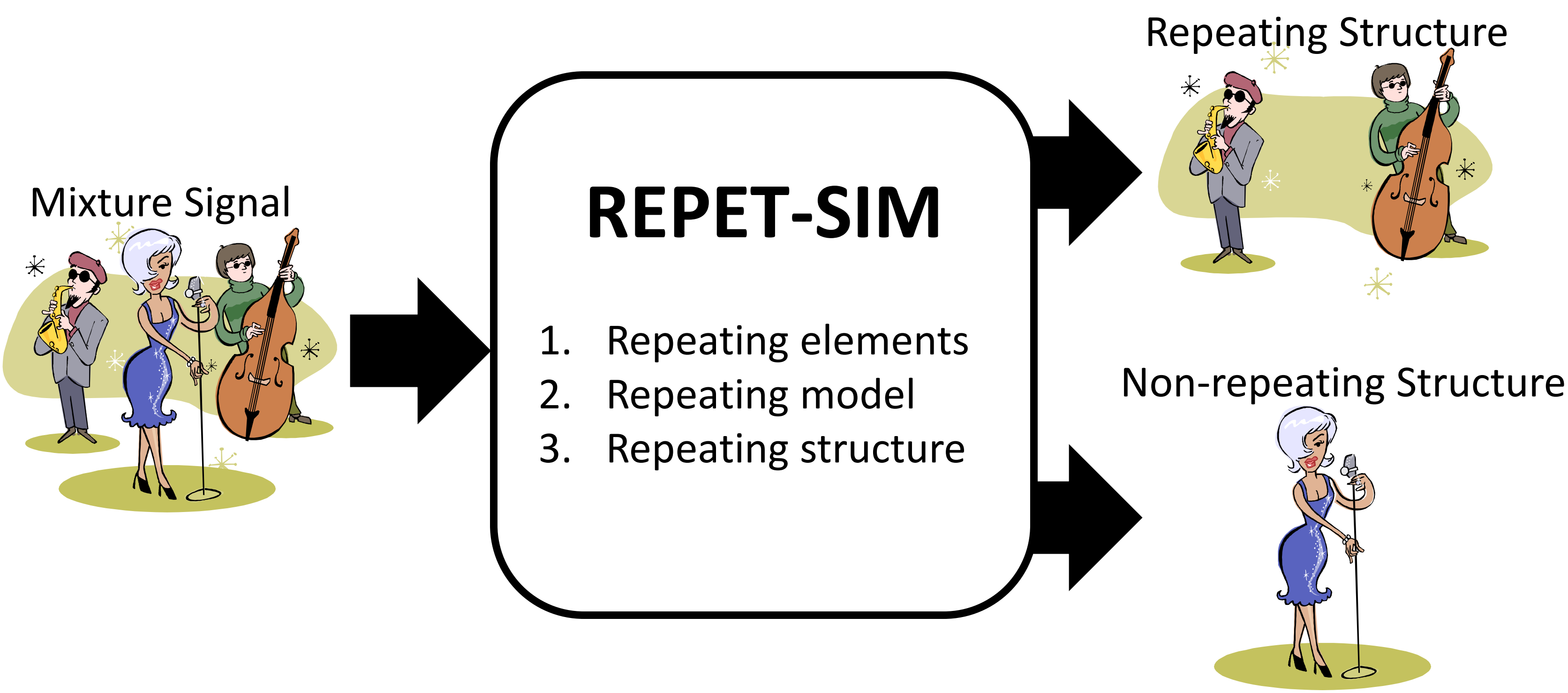


+



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.**