REpeating Pattern Extraction Technique (REPET)

EECS 352: Machine Perception of Music & Audio

Observation

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

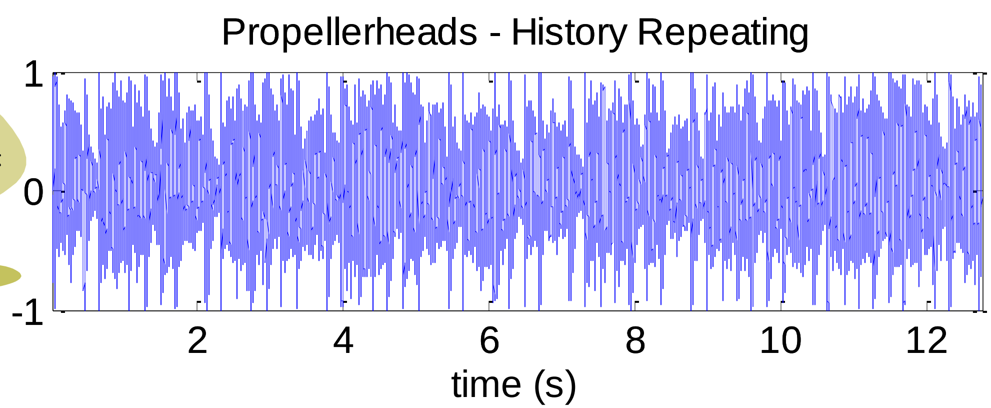
… in nature

… in art [http://http://en.wikipedia.org/wiki/Campbell's_Soup_Cans]

… in audio

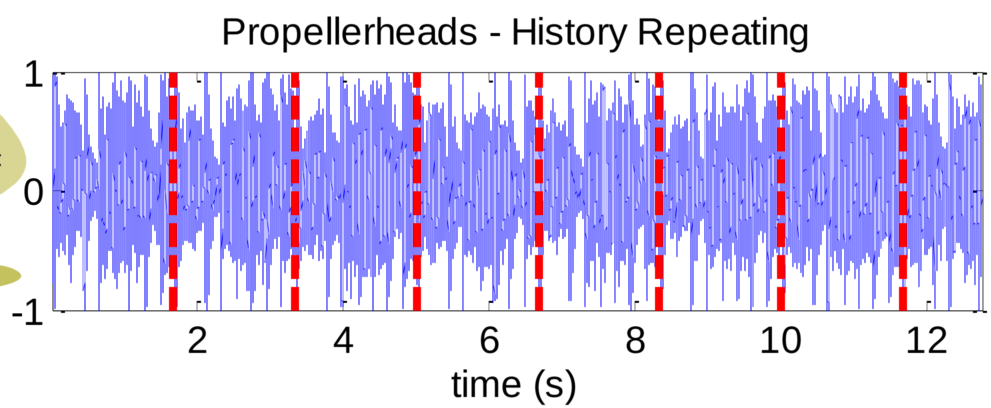
Observation

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



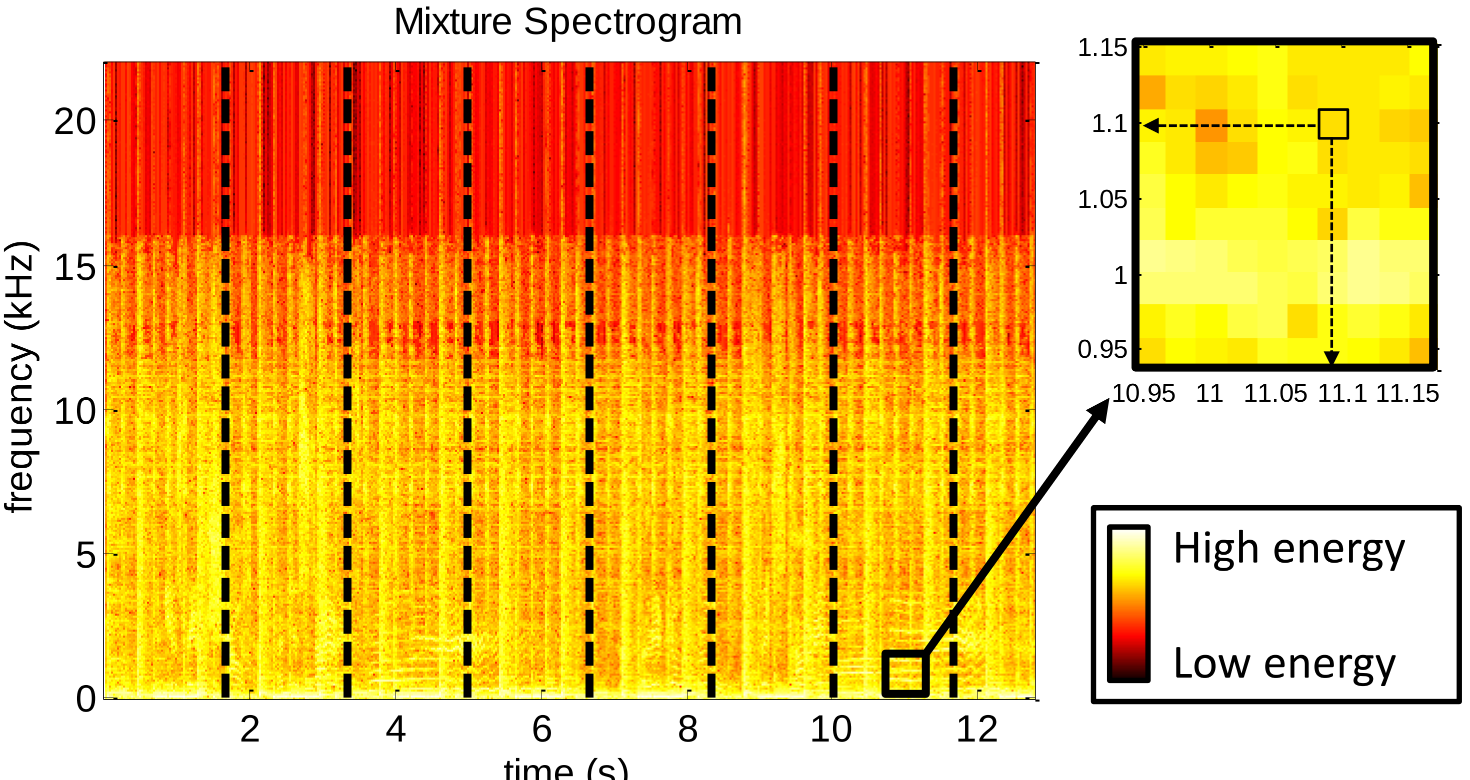
Observation

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



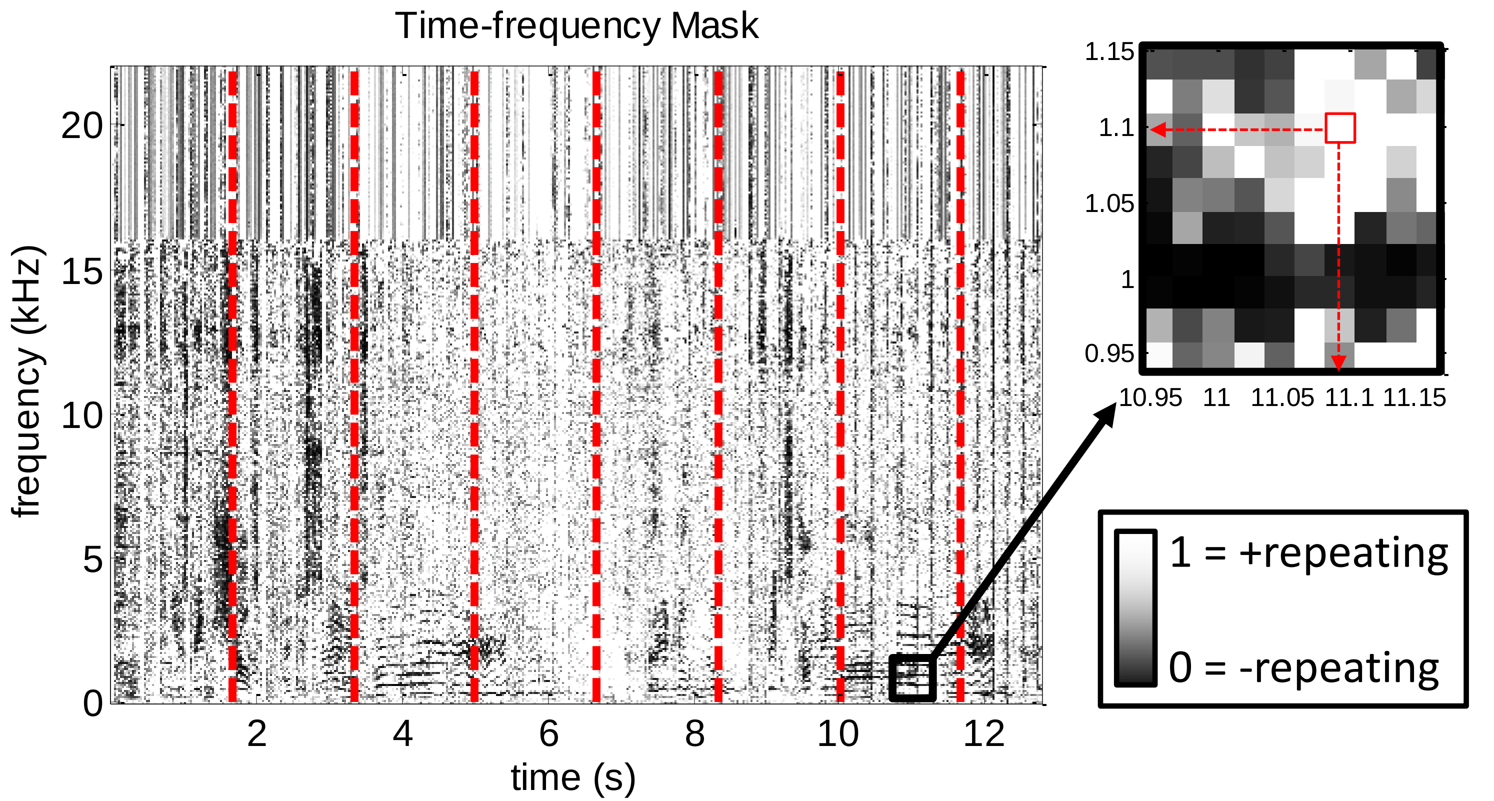
Assumption

- There should be patterns that are more or less **repeating in time and frequency**



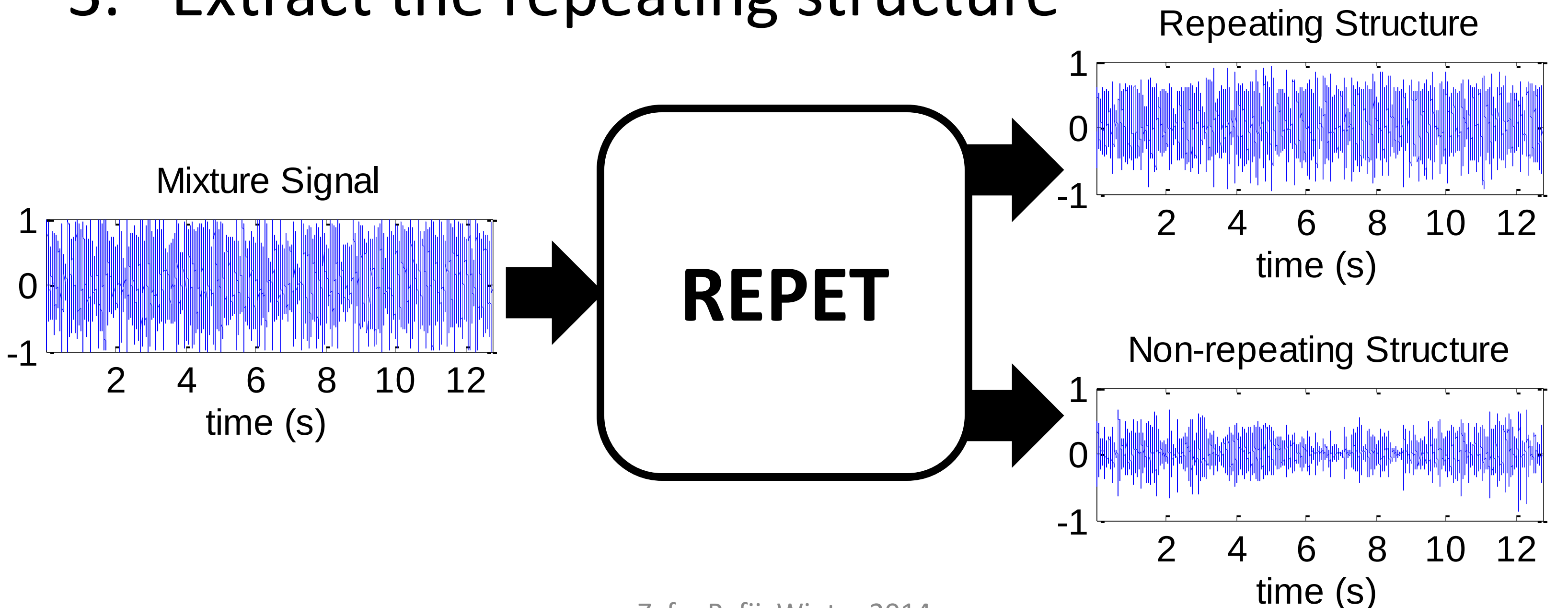
Assumption

- The repeating patterns could be identified and extracted using a **time-frequency mask**



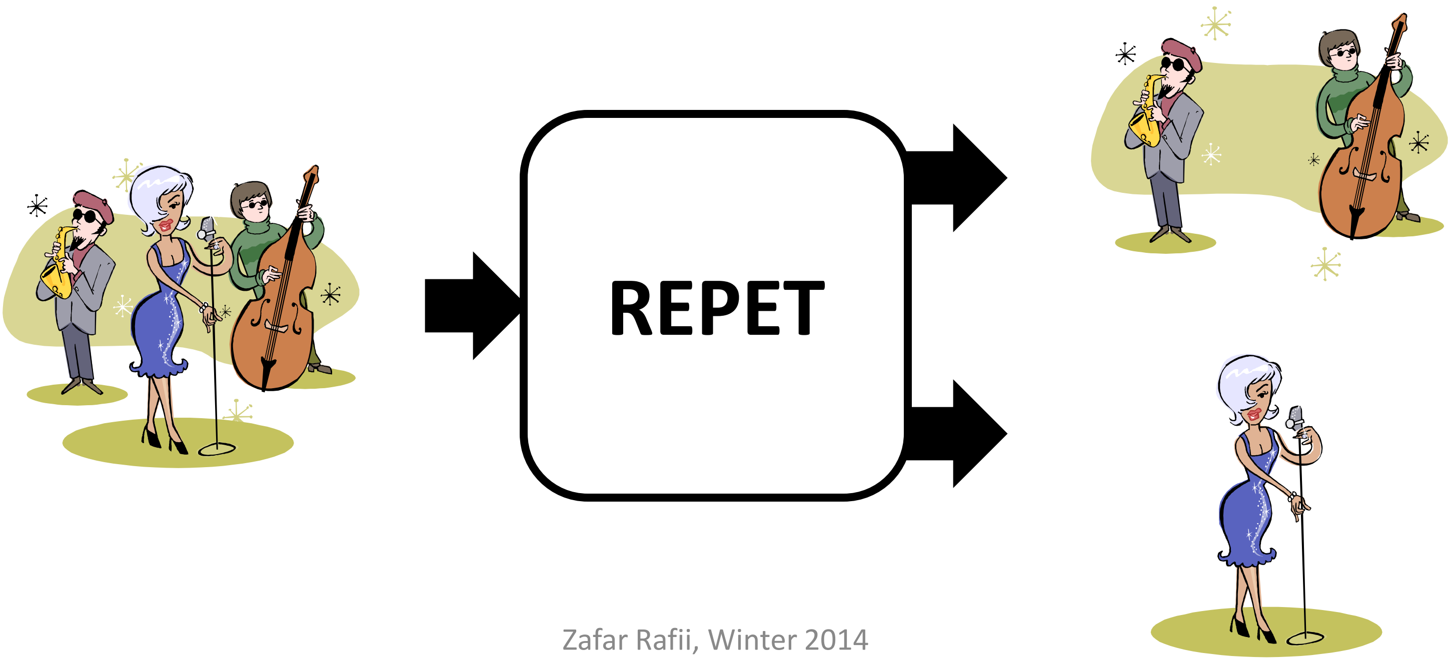
Idea

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

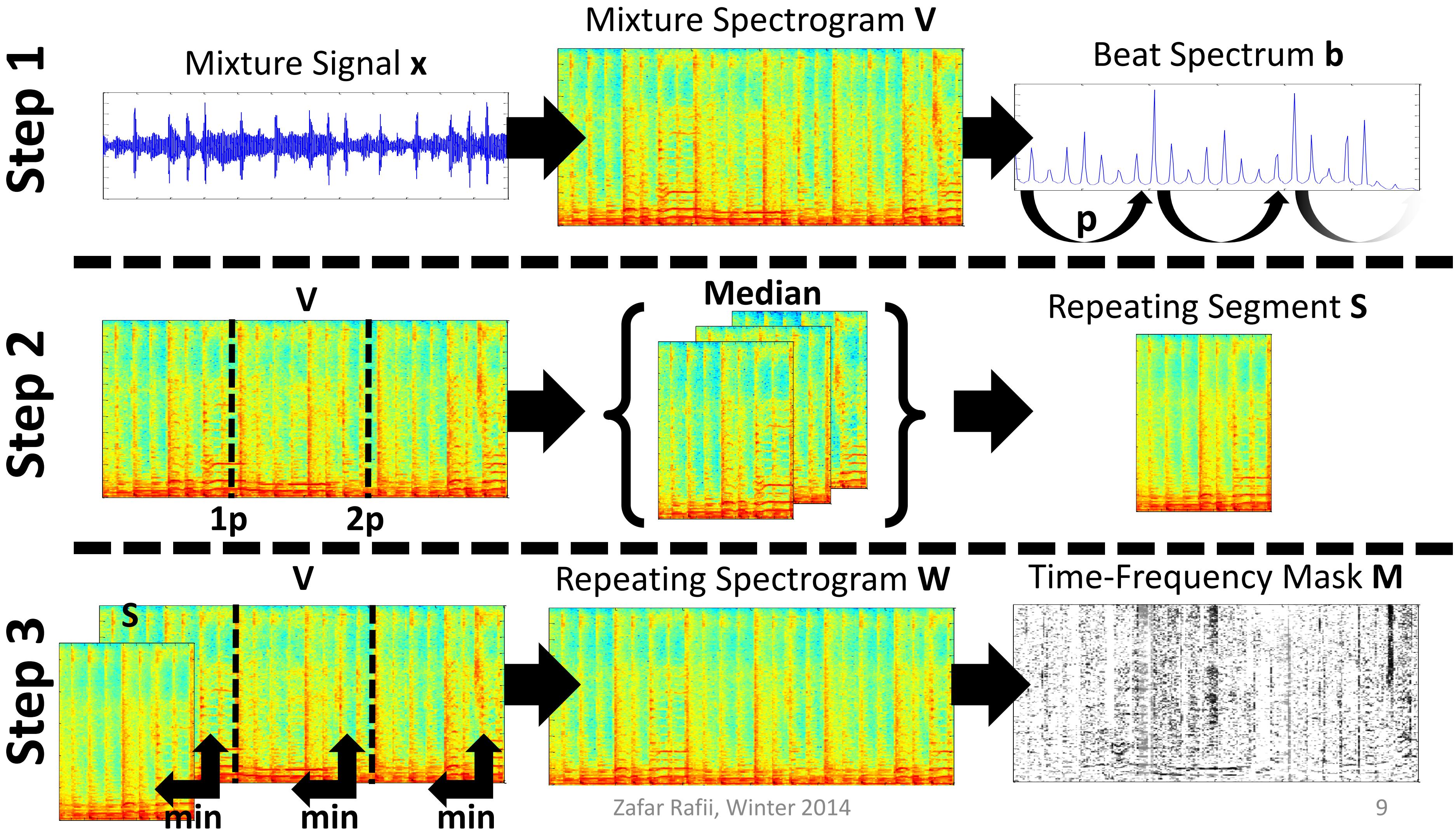


Idea

- Simple **music/voice separation** method!
  - Repeating structure = background music
  - Non-repeating structure = foreground voice



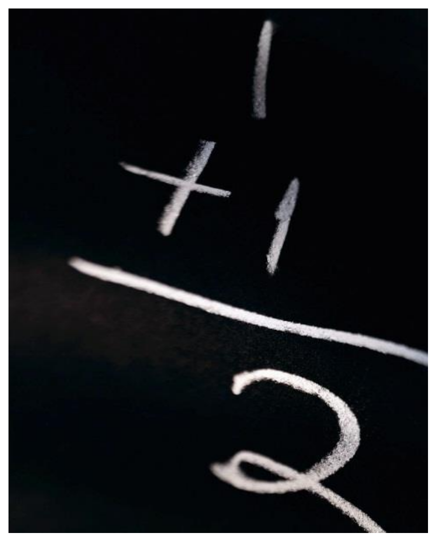
REPET



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**

Practical Advantages

- Does not depend on special parametrizations
- Does not rely on complex frameworks
- Does not require external information

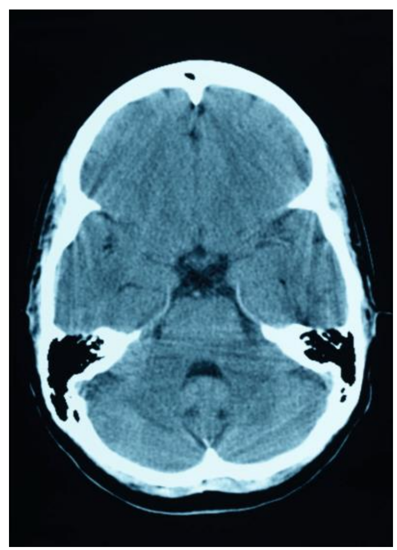


Practical Interests

- Karaoke gaming (need the music)
- Query-by-humming (need the voice)
- Audio remixing (need both components)

Intellectual Interests

- Music understanding
- Music perception
- Simply based on repetition!

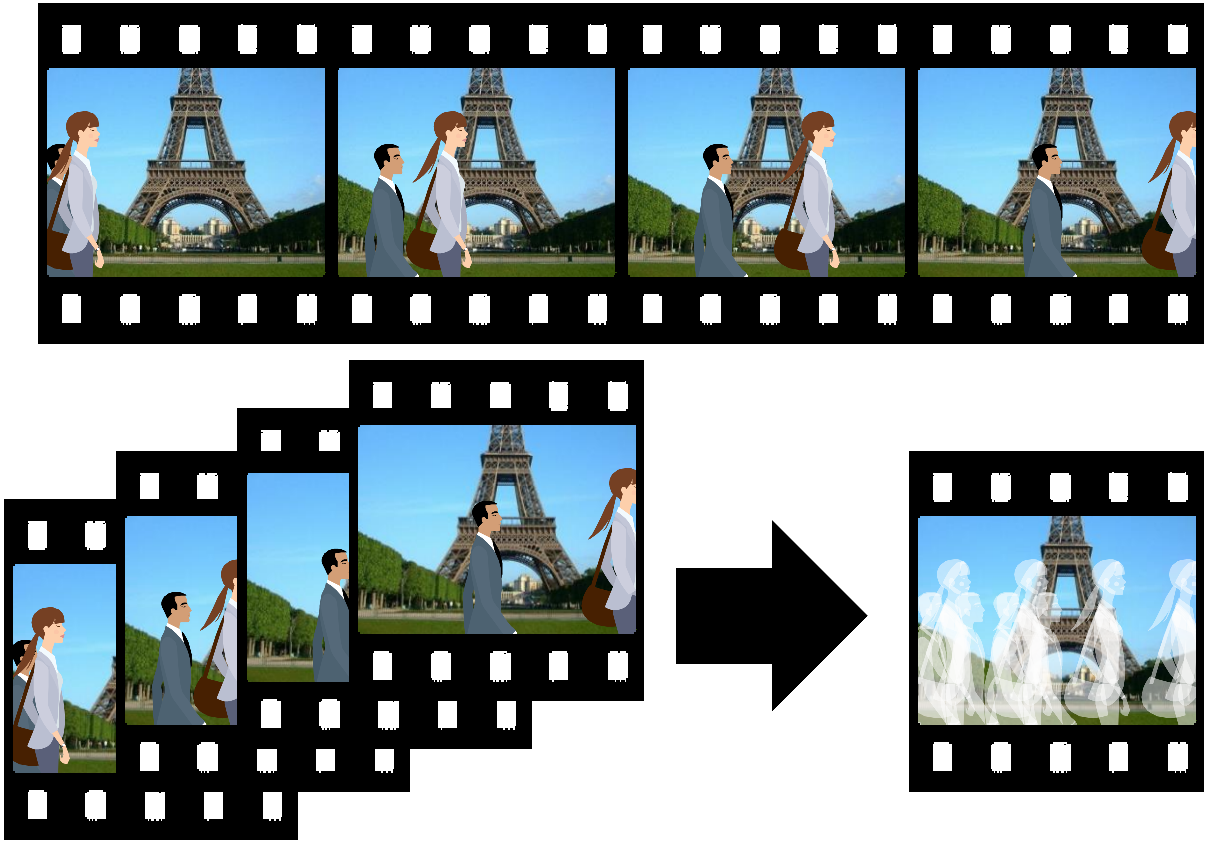


Parallels

- **Background subtraction** in computer vision

Sequence of video frames

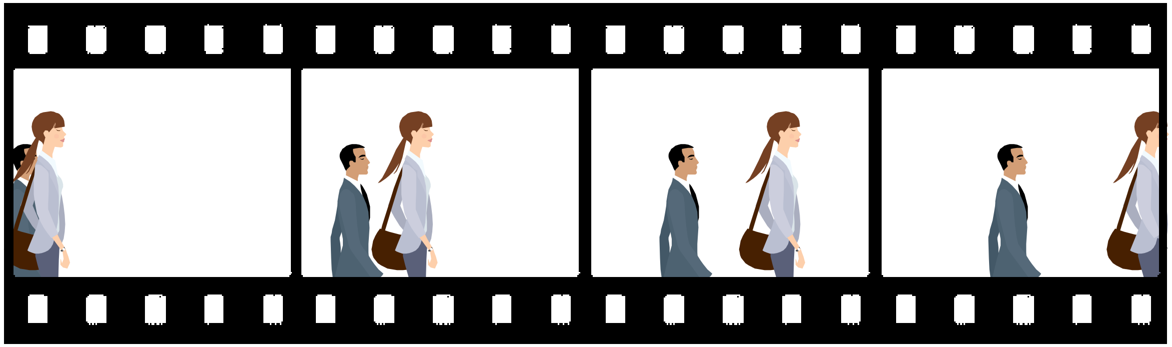
Compare frames to estimate a background model



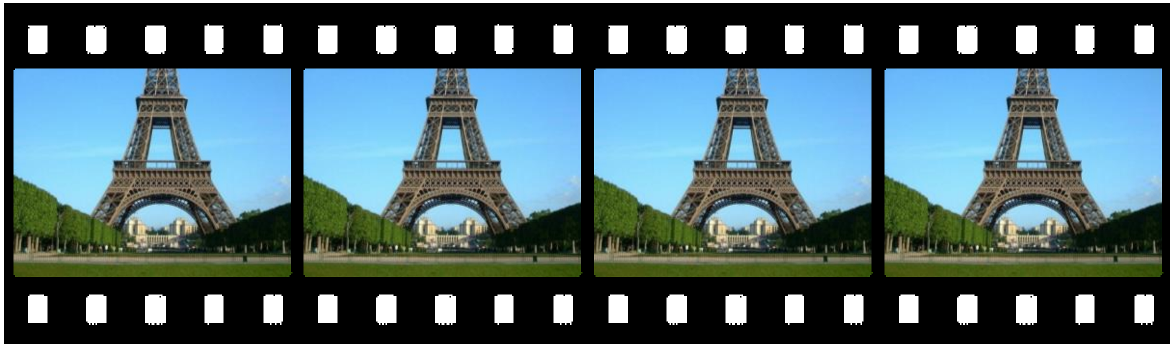
Parallels

- **Background subtraction** in computer vision

Extracted varying foreground scene

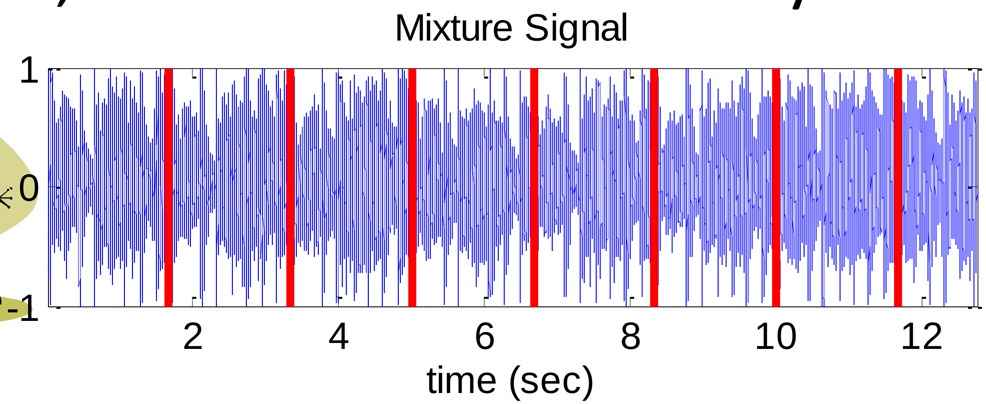


Extracted fixed background scene



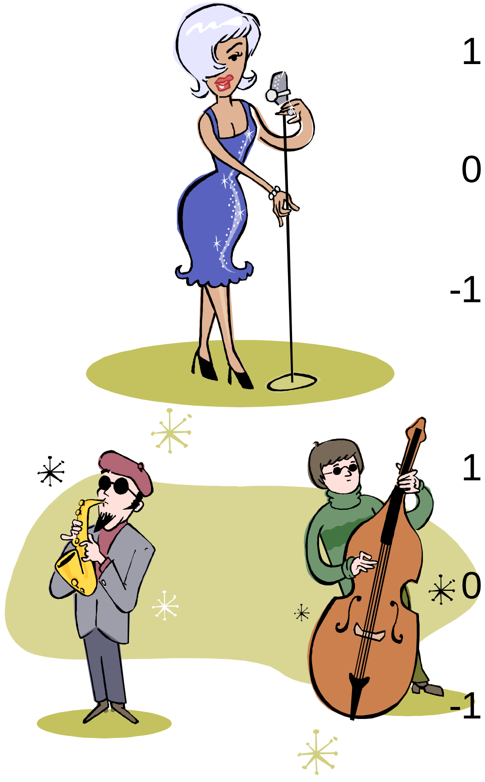
Parallels

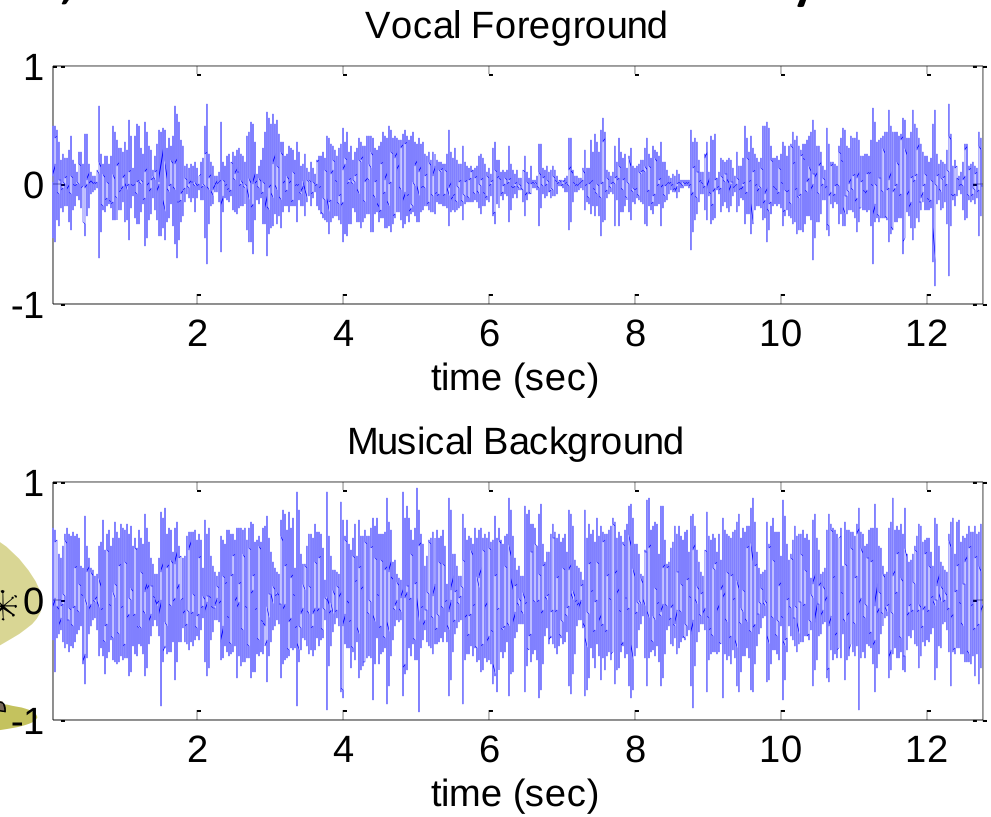
- **Background subtraction** in computer vision
  - In audio, we also need to identify the repetitions!



Parallels

- **Background subtraction** in computer vision
  - In audio, we also need to identify the repetitions!





Parallels

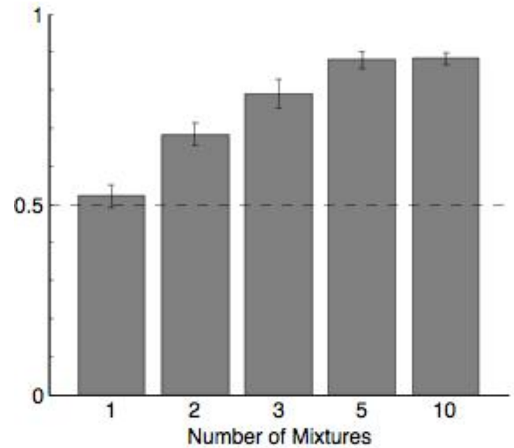
- **Auditory segregation** in human listeners

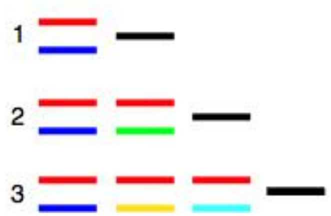
Handsome professor

Unknown audio mixtures
with the same target
and different distractors

Parallels

- **Auditory segregation** in human listeners





red/black = target/probe,
other colors = distractors

1 mixture:

2 mixtures:

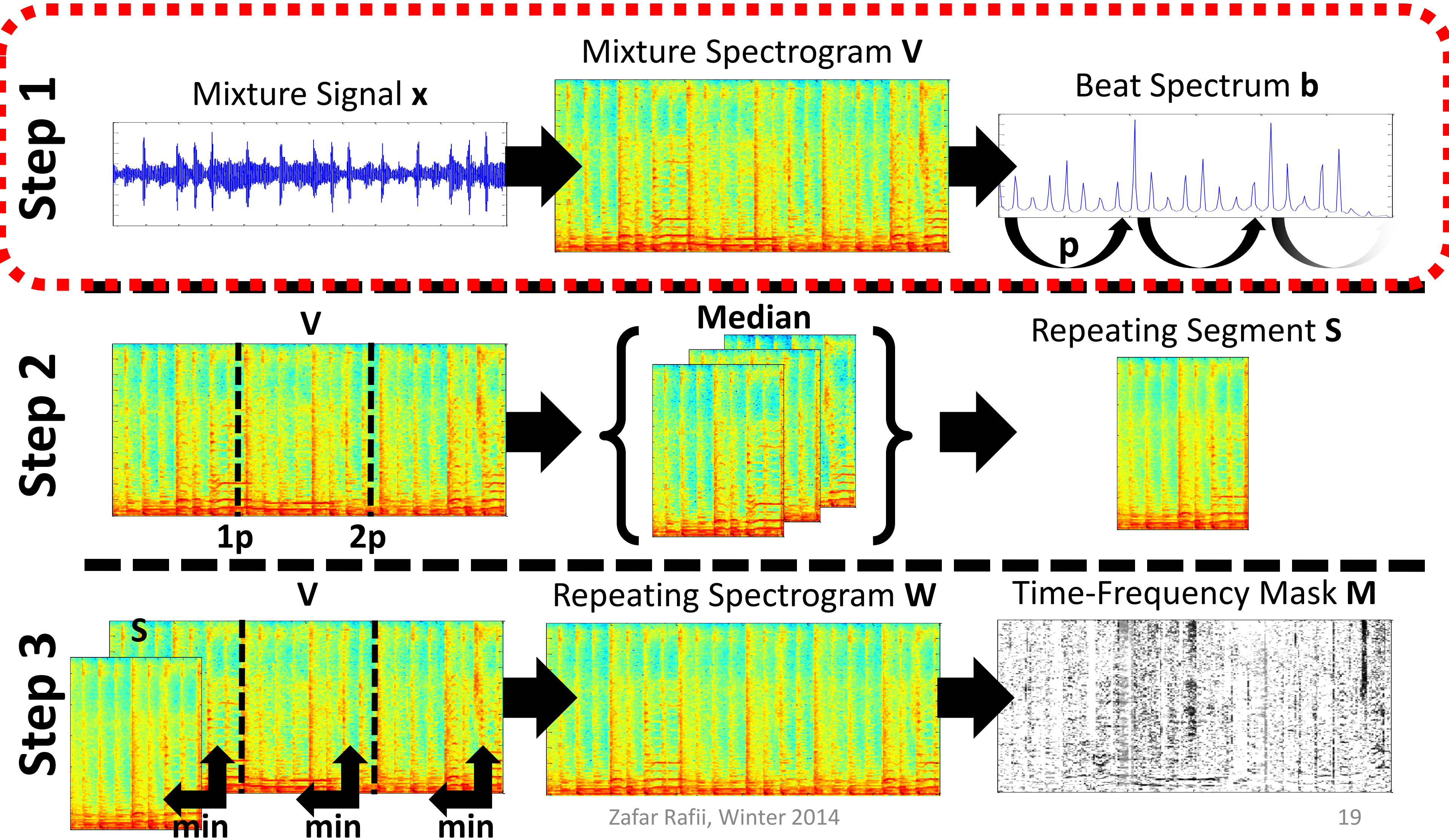
3 mixtures:

5 mixtures:

10 mixtures:

As the number of mixtures increases, the target becomes more apparent…
[courtesy of Josh McDermott]

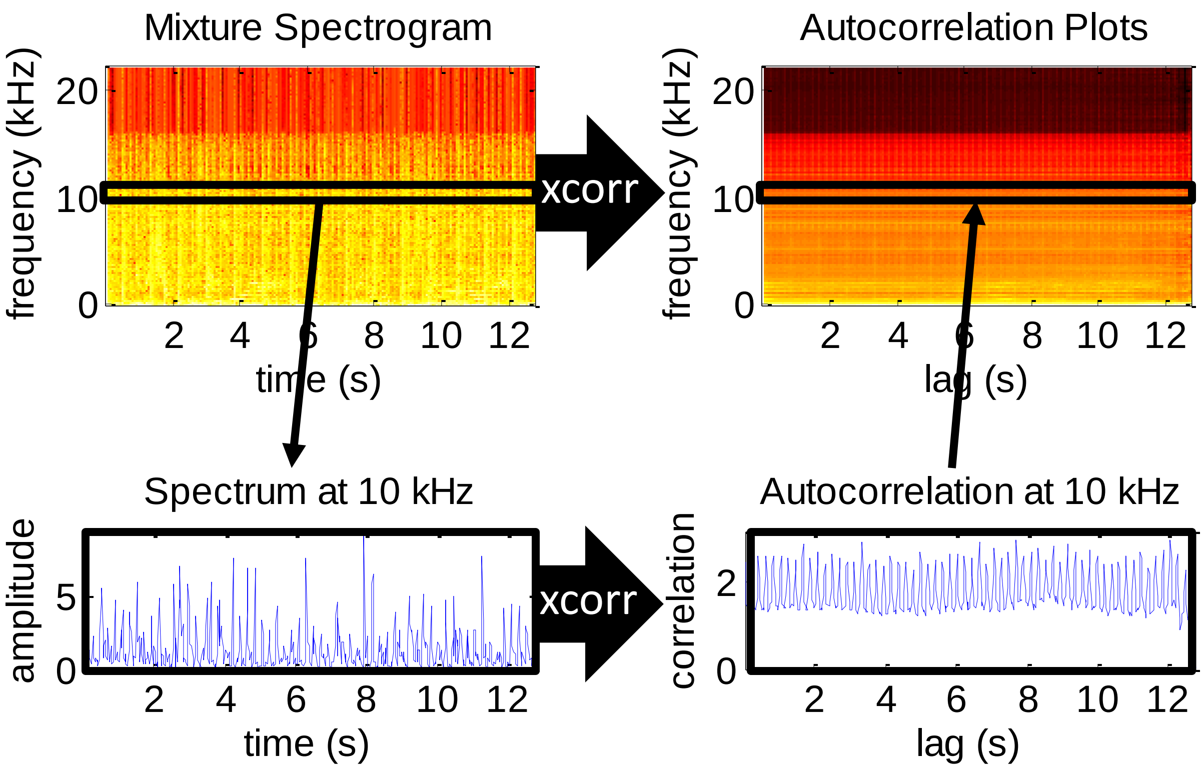
REPET



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

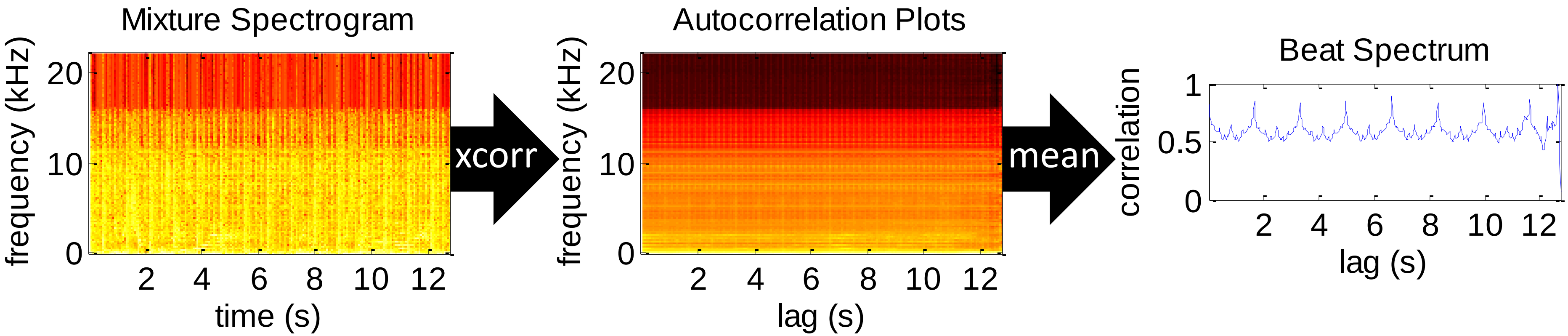
1. Repeating Period

- We compute the **autocorrelations** of the frequency rows of the mixture spectrogram



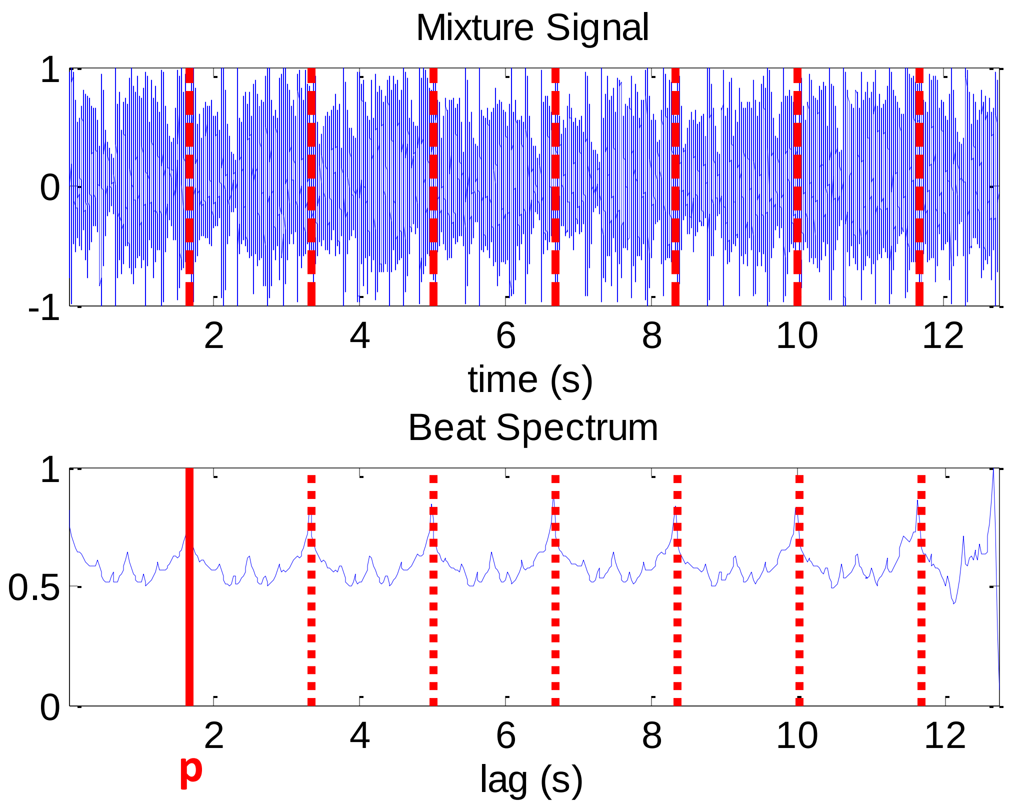
1. Repeating Period

- We take the mean of the autocorrelation rows and obtain the **beat spectrum**

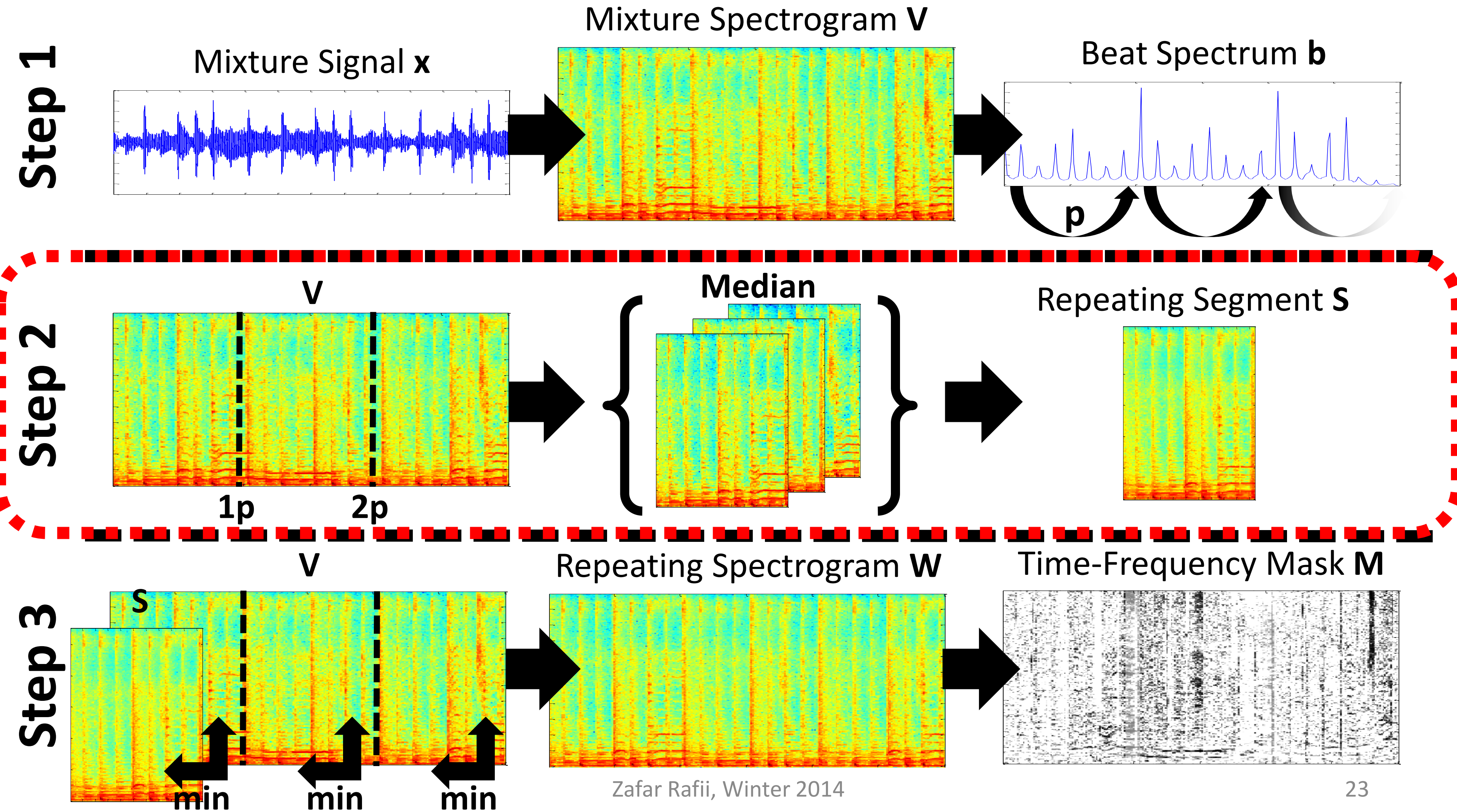


1. Repeating Period

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

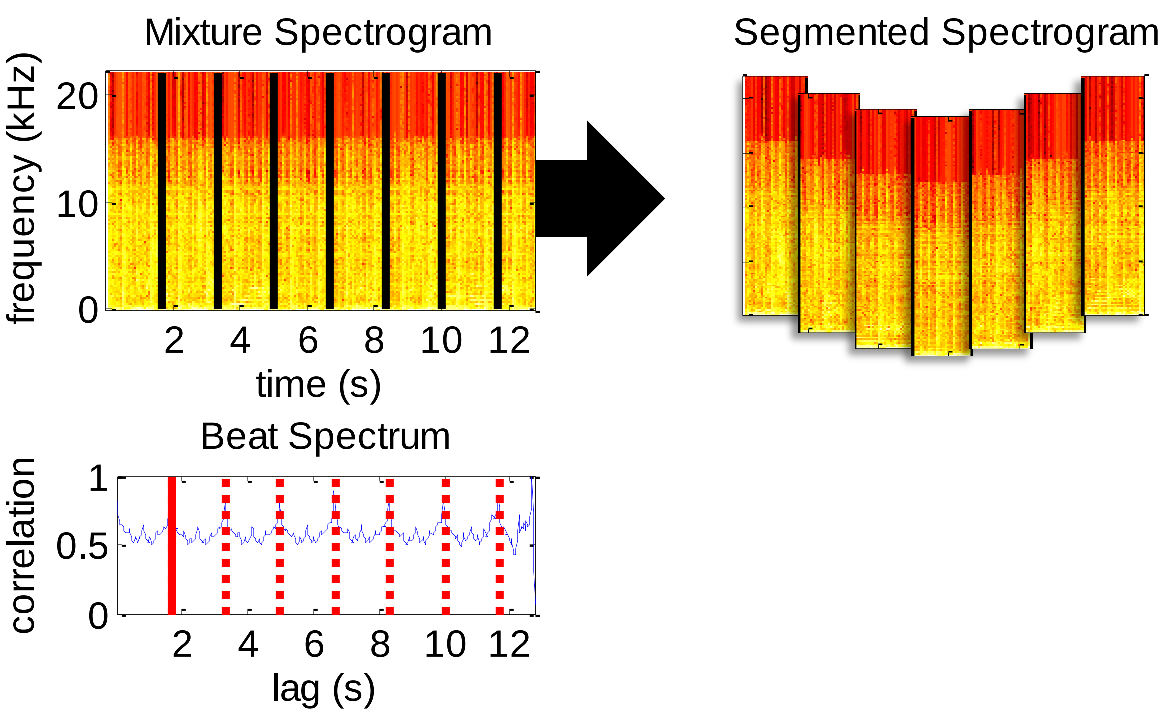


REPET


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**

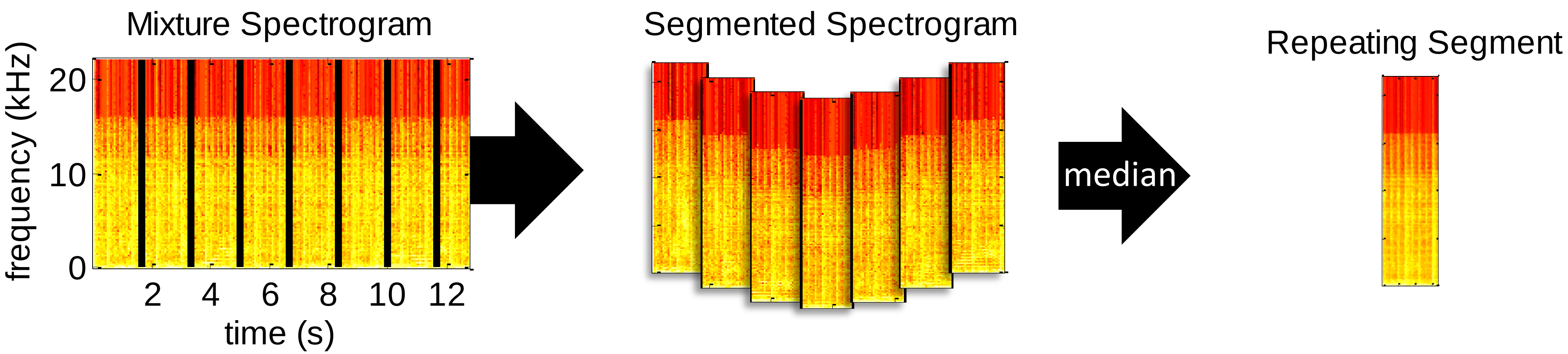
2. Repeating Segment

- We then use the repeating period to **segment** the mixture spectrogram at period rate



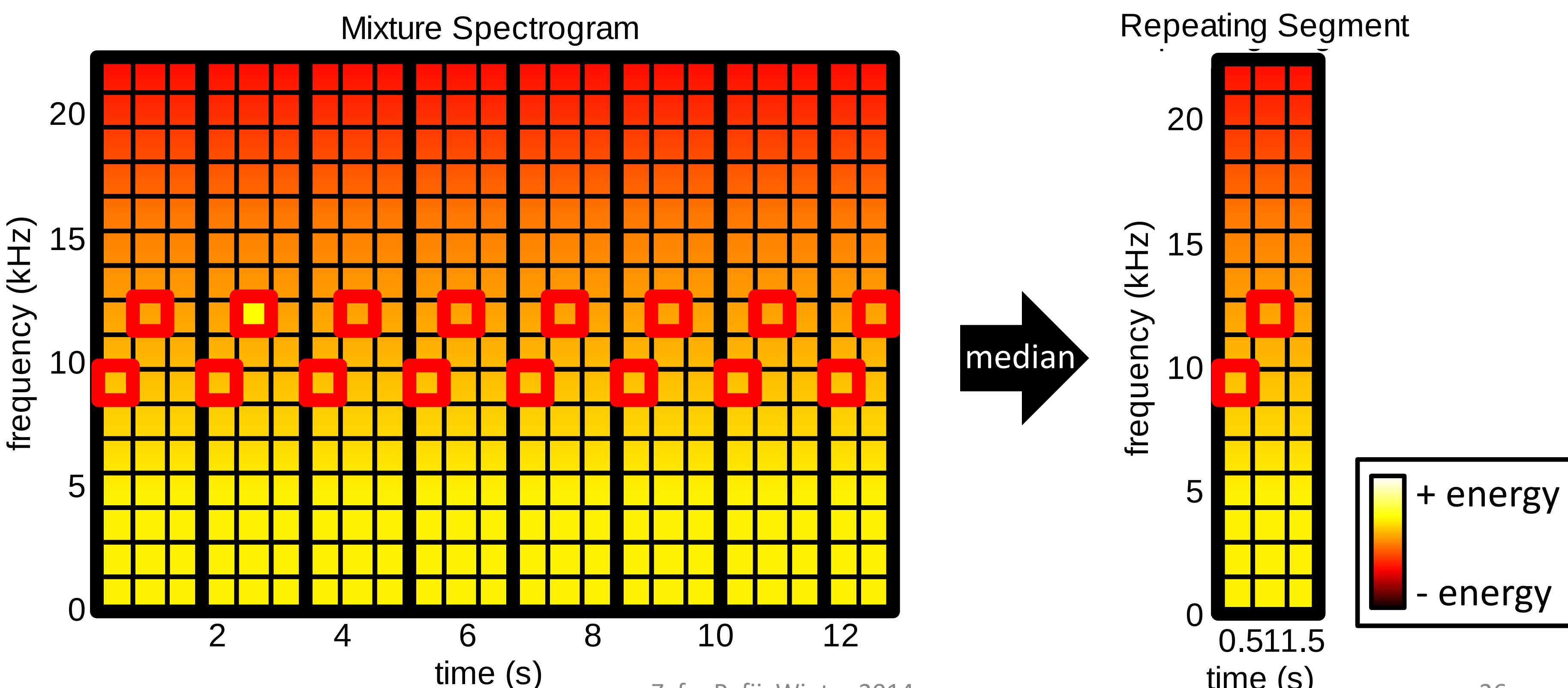
2. Repeating Segment

- We derive a **repeating segment model** by taking the element-wise median of segments

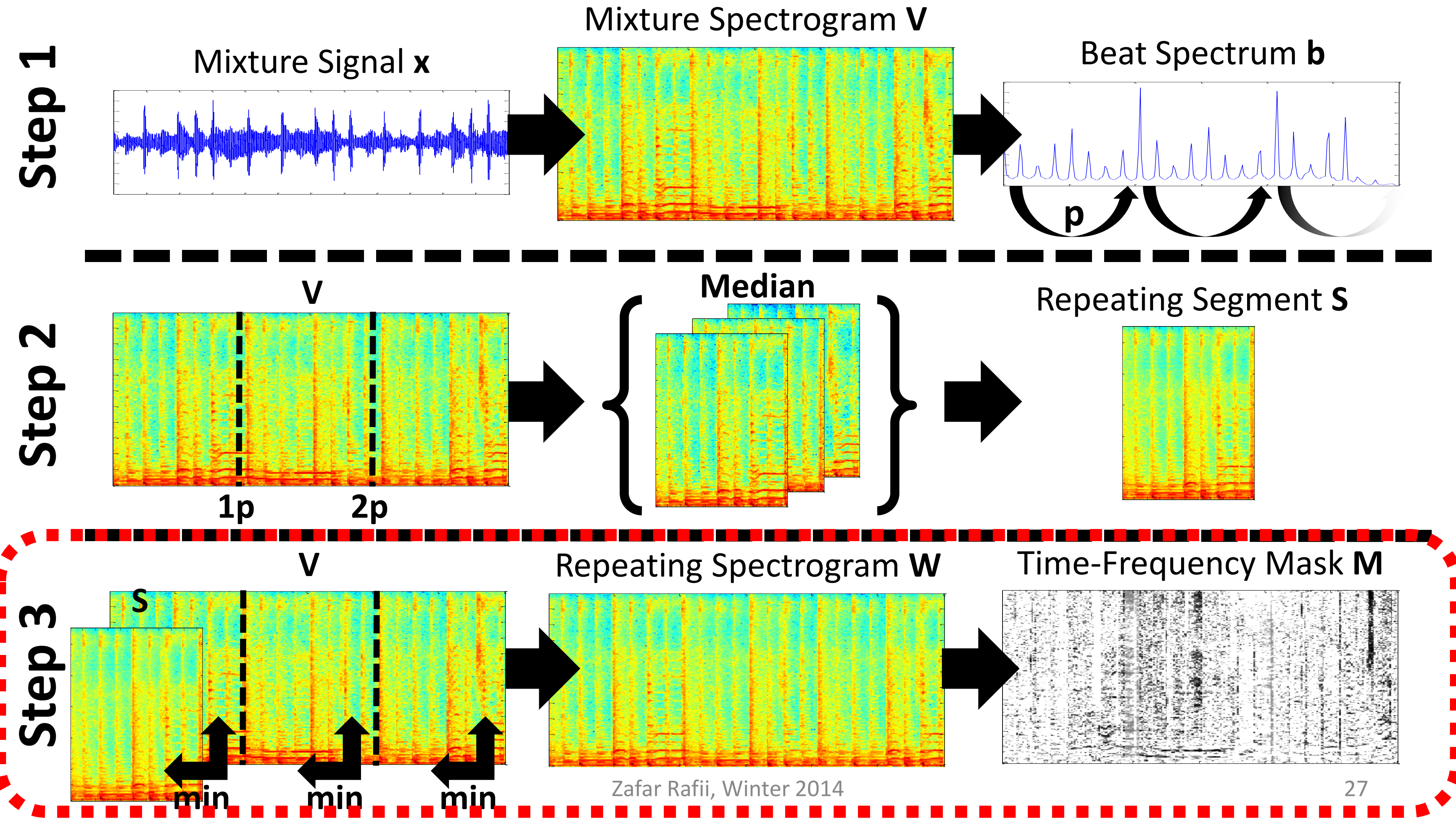


2. Repeating Segment

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

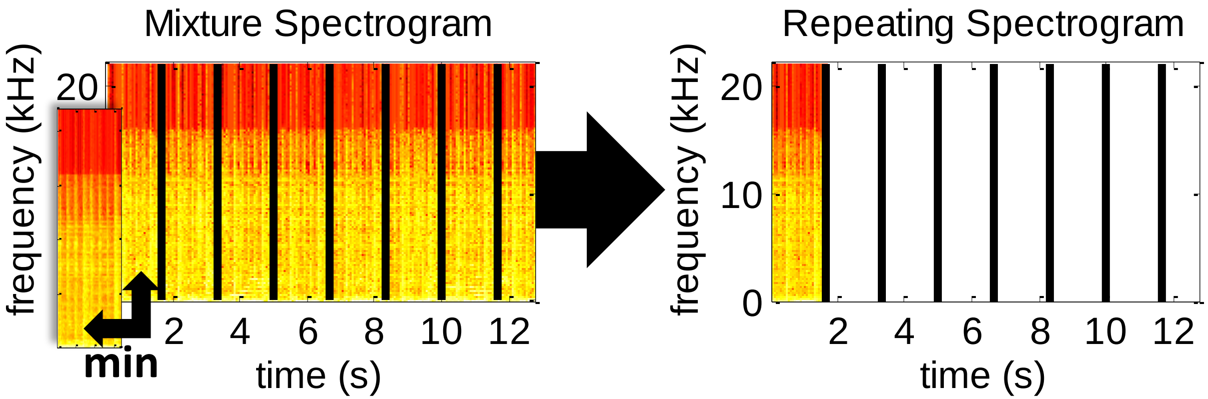


REPET


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

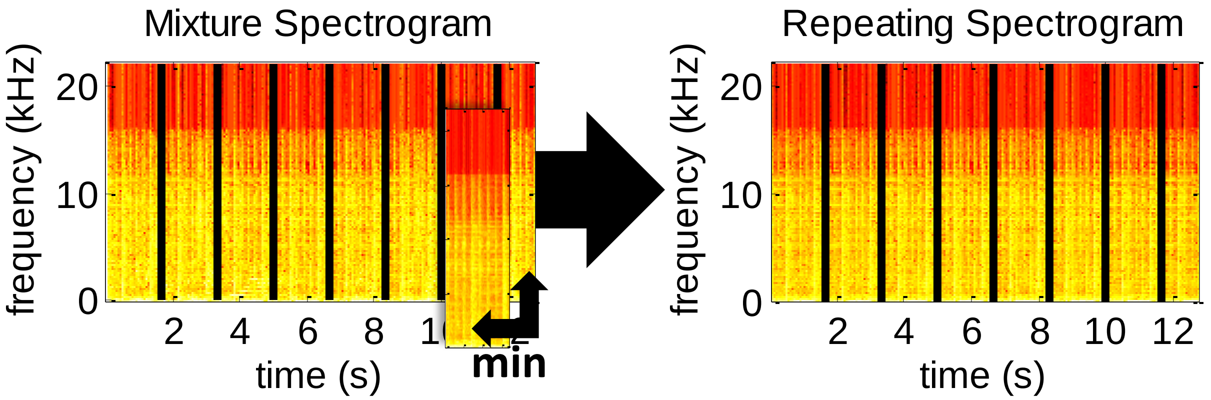
3. Repeating Structure

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



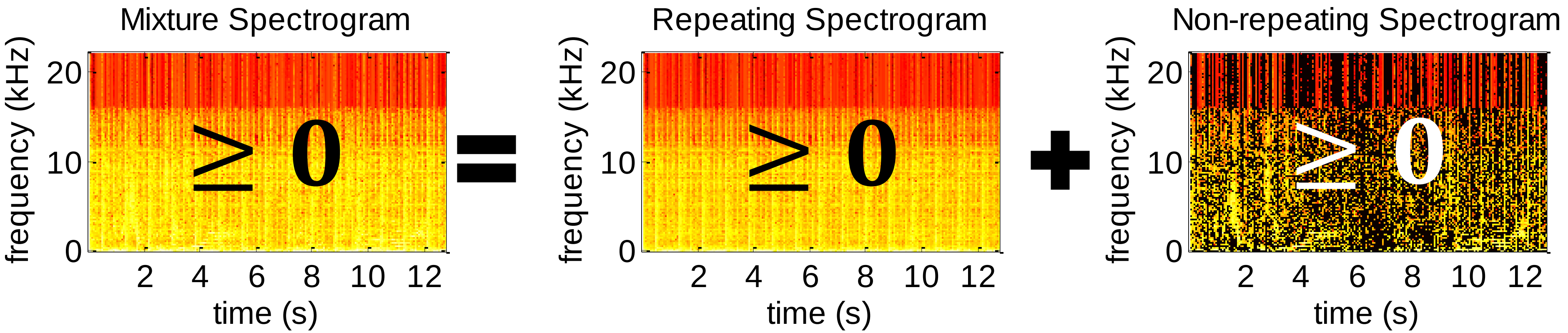
3. Repeating Structure

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



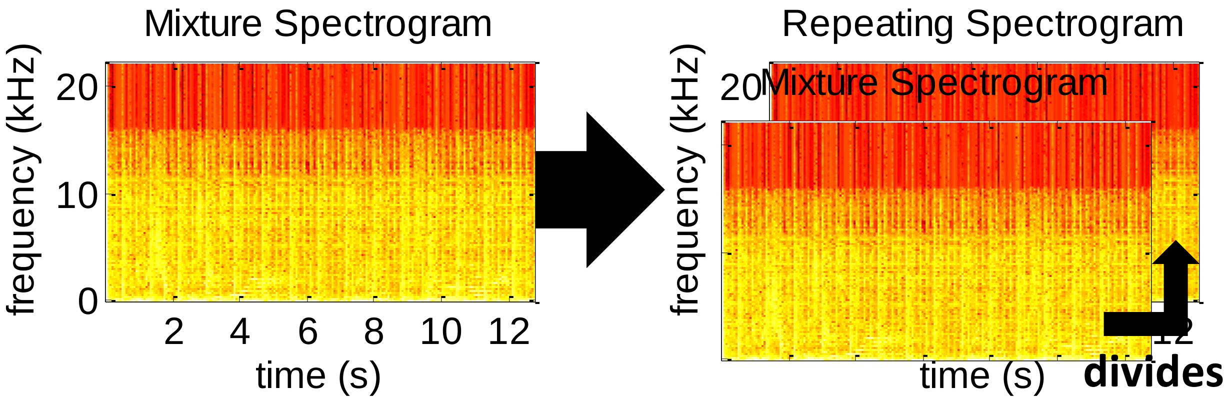
3. Repeating Structure

- The repeating spectrogram **should not have values higher than** the mixture spectrogram



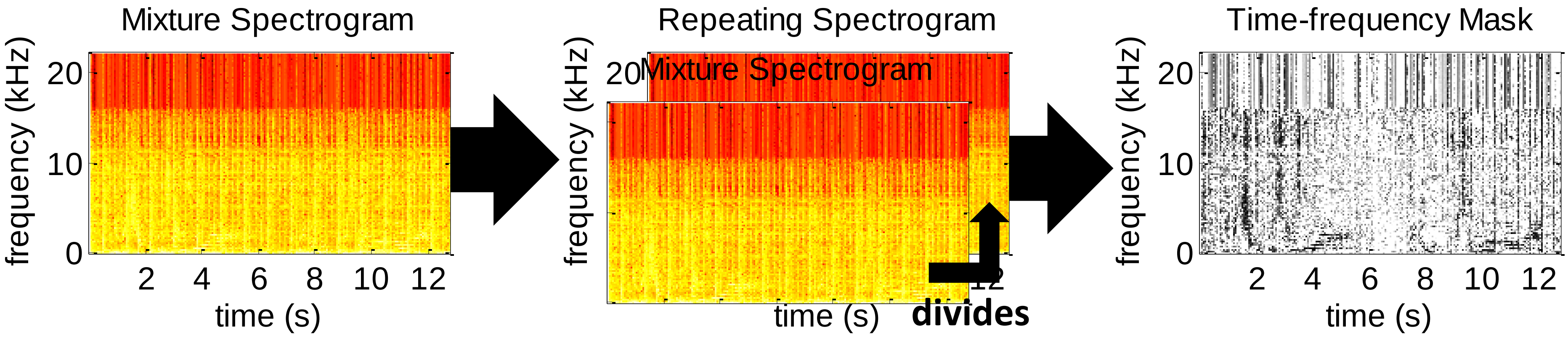
3. Repeating Structure

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



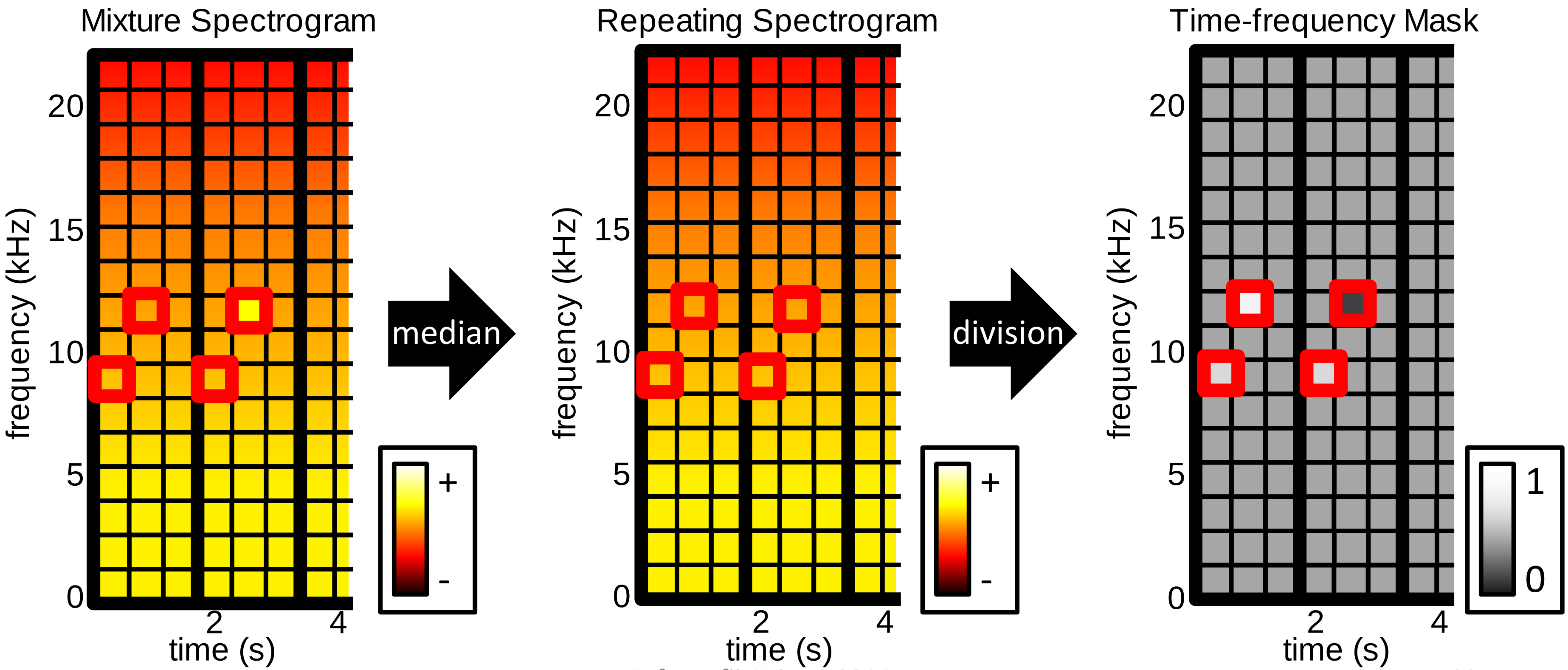
3. Repeating Structure

- We obtain a **soft time-frequency mask** (with values between 0 and 1)



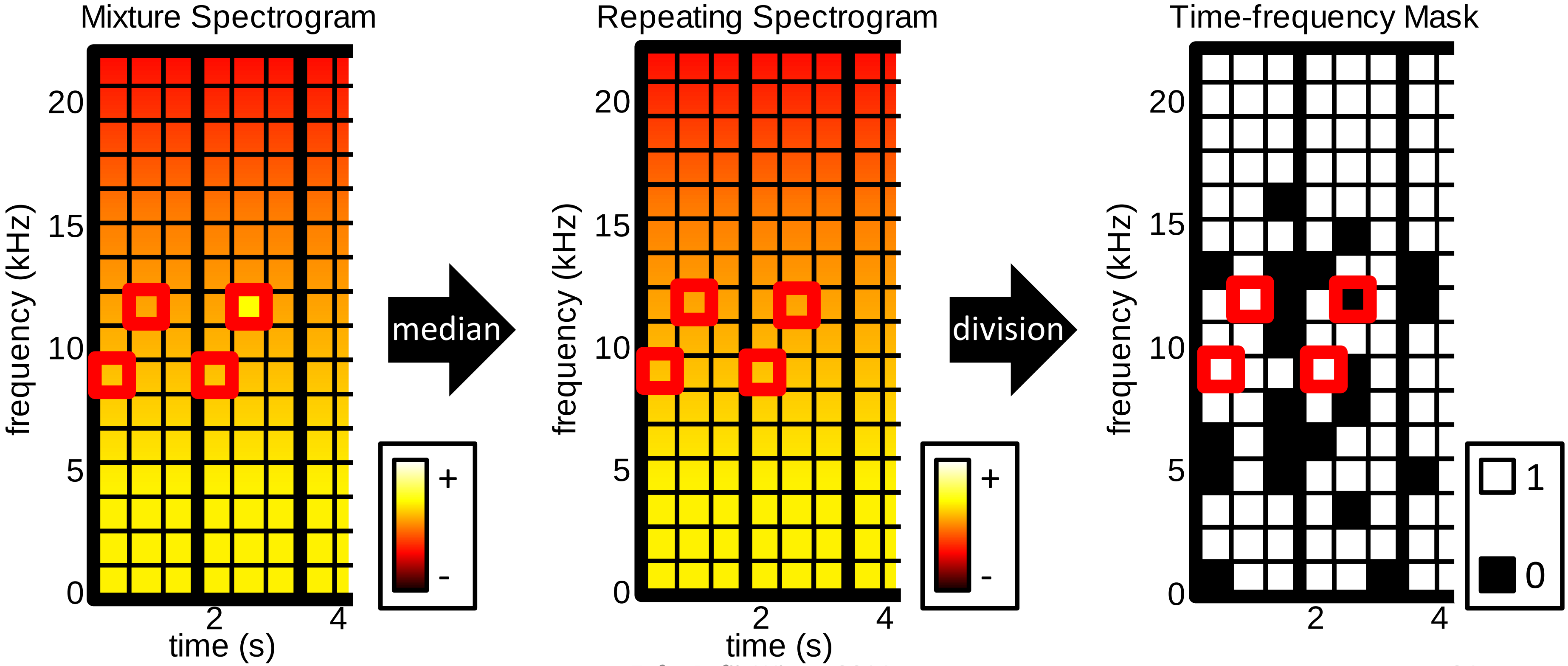
3. Repeating Structure

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



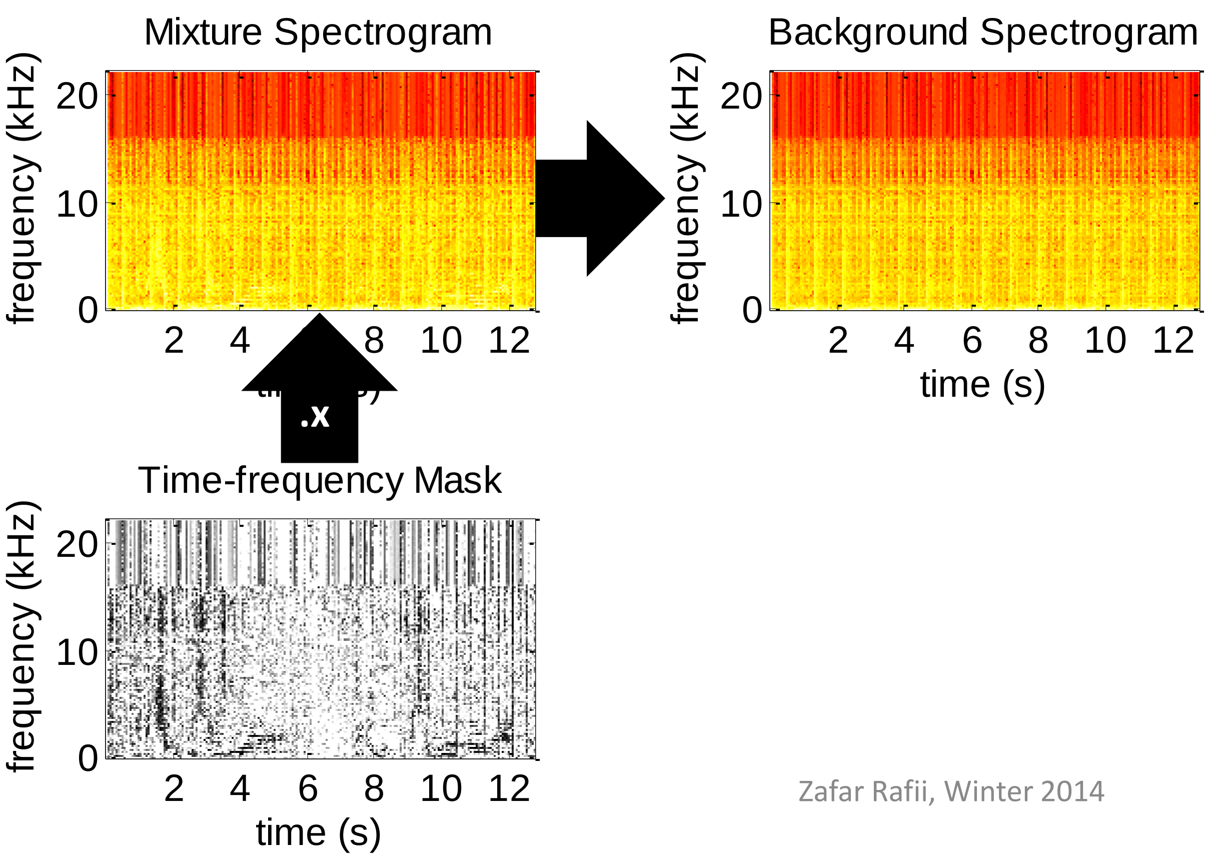
3. Repeating Structure

- We could further derive a **binary t-f mask** by fixing a threshold between 0 and 1



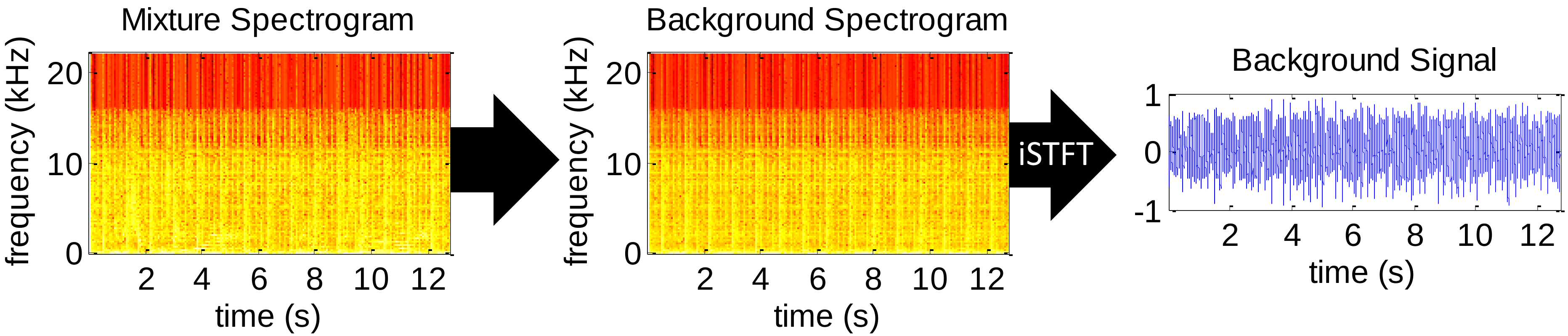
3. Repeating Structure

- We **multiply**, element-wise, the t-f mask with the mixture STFT to get the background STFT



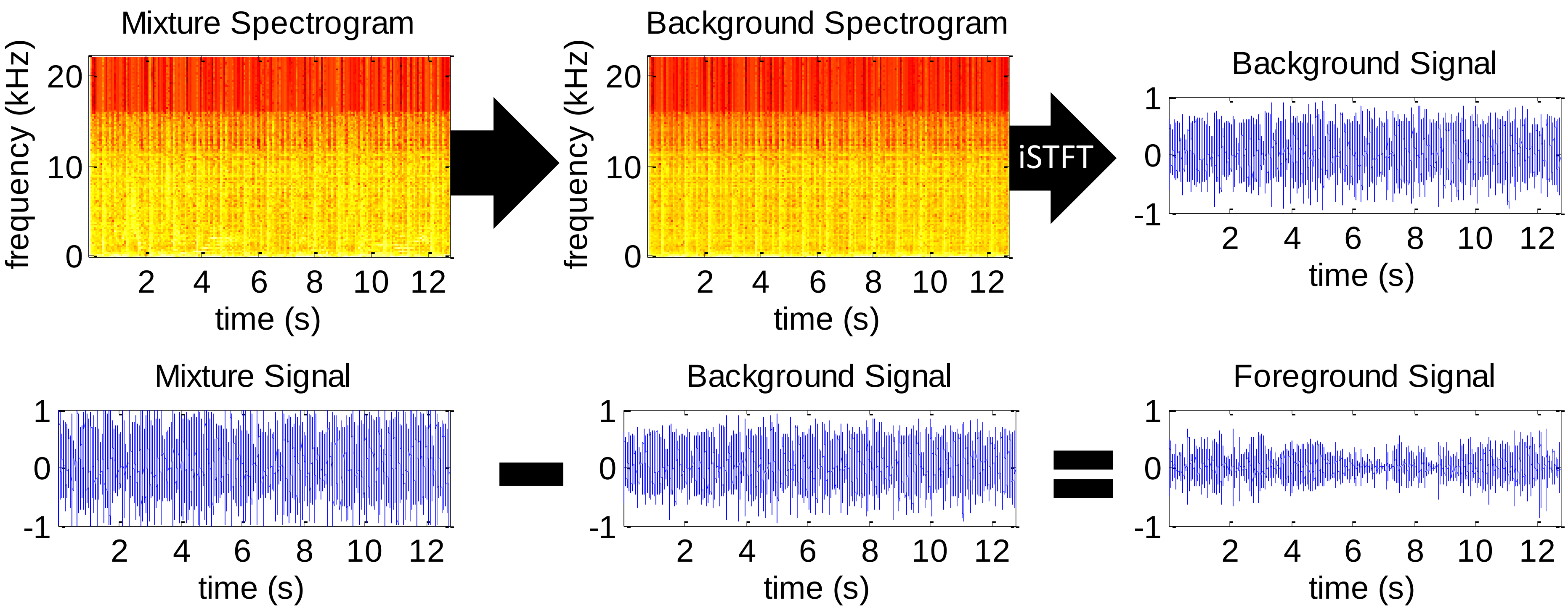
3. Repeating Structure

- We obtain the **repeating background** signal by inverting its STFT into the time domain



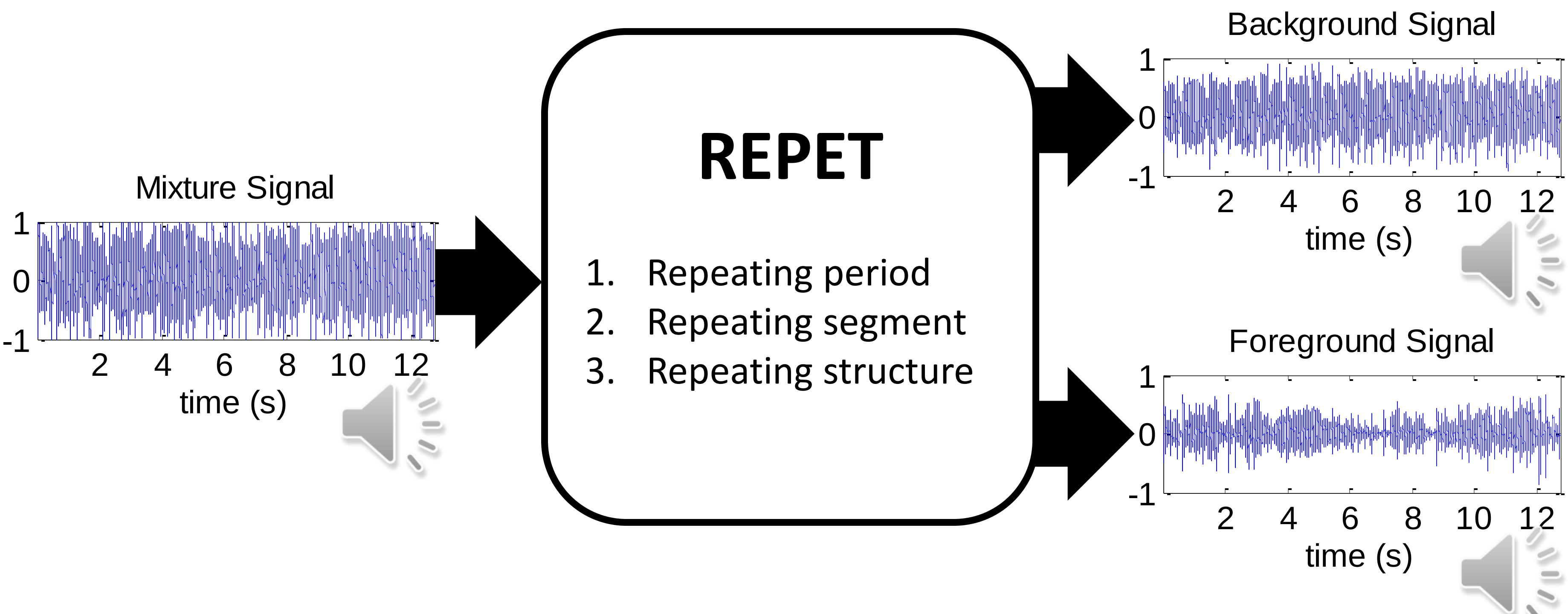
3. Repeating Structure

- We obtain the **non-repeating foreground** signal by subtracting background from mixture



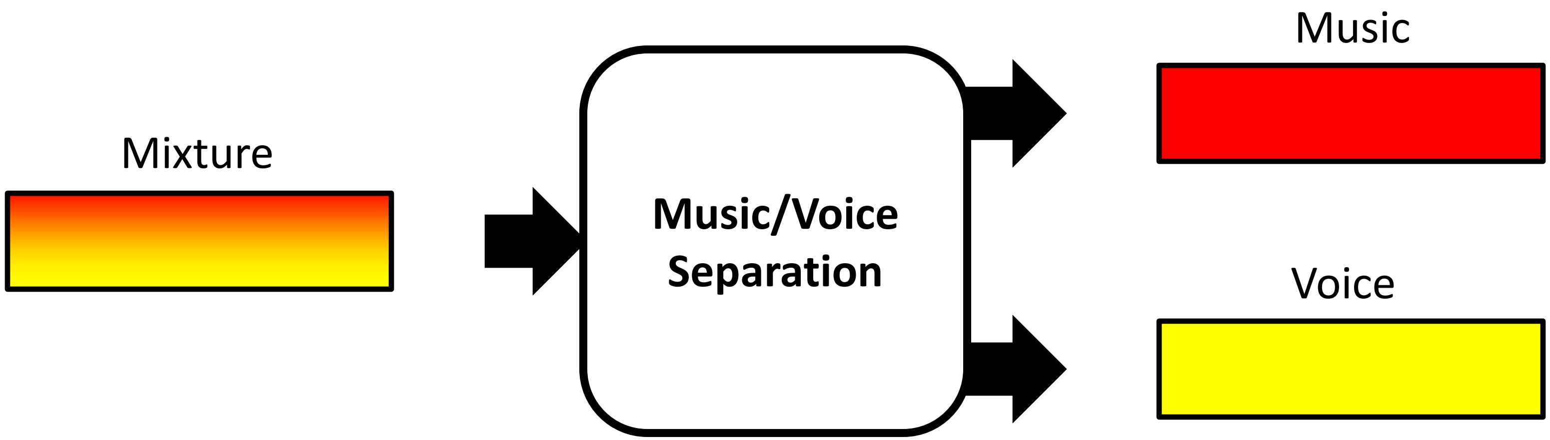
Summary

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



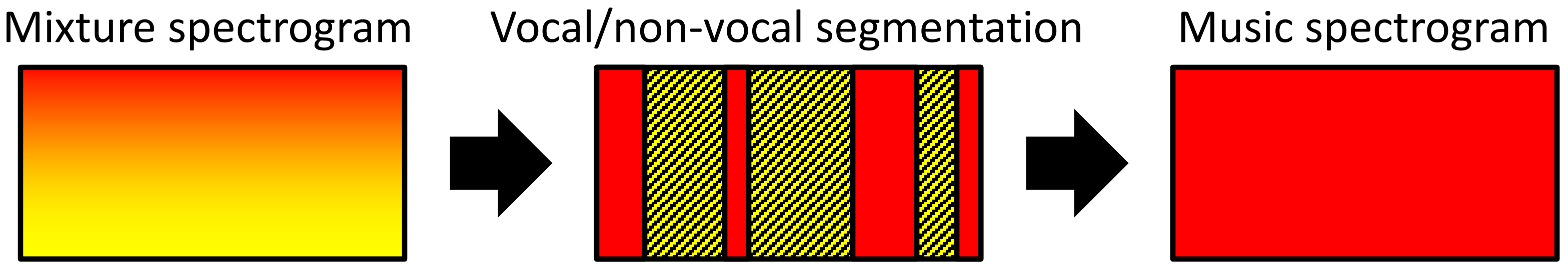
Music/Voice Separation

- A variety of techniques has been proposed to separate **music** and **voice** from a mixture
  - Accompaniment modeling, Pitch-based inference, Non-negative Matrix Factorization (NMF), etc.



Music/Voice Separation

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

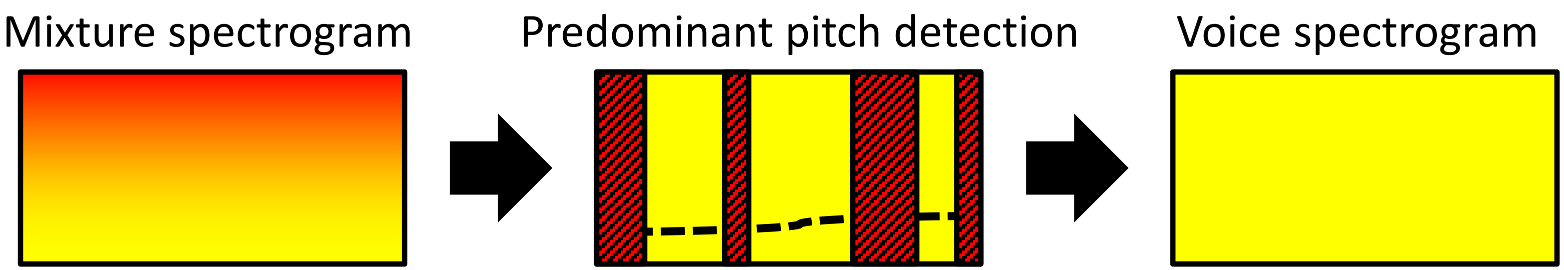


→ Need an accurate vocal/non-vocal segmentation!

→ 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

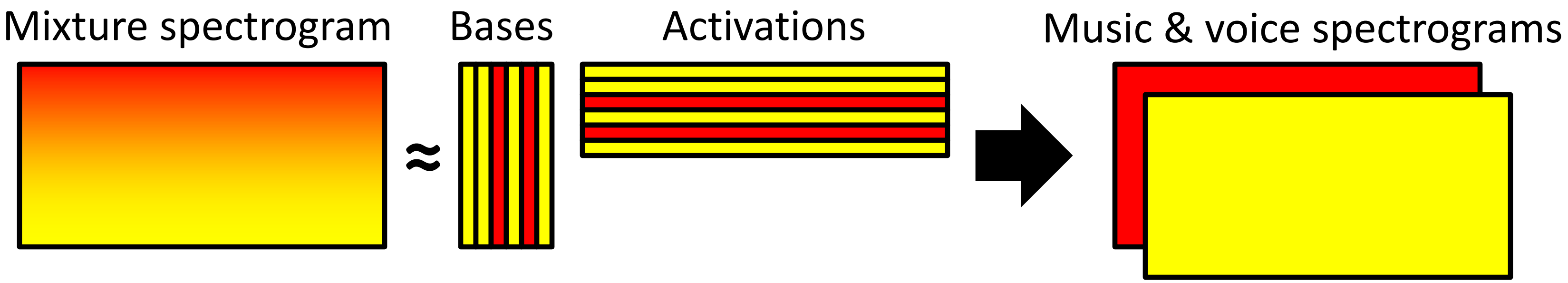


→ Need an accurate predominant pitch detection!

→ Cannot extract unvoiced vocals!

Music/Voice Separation

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



→ Need to know the number of components!

→ Need a proper initialization!

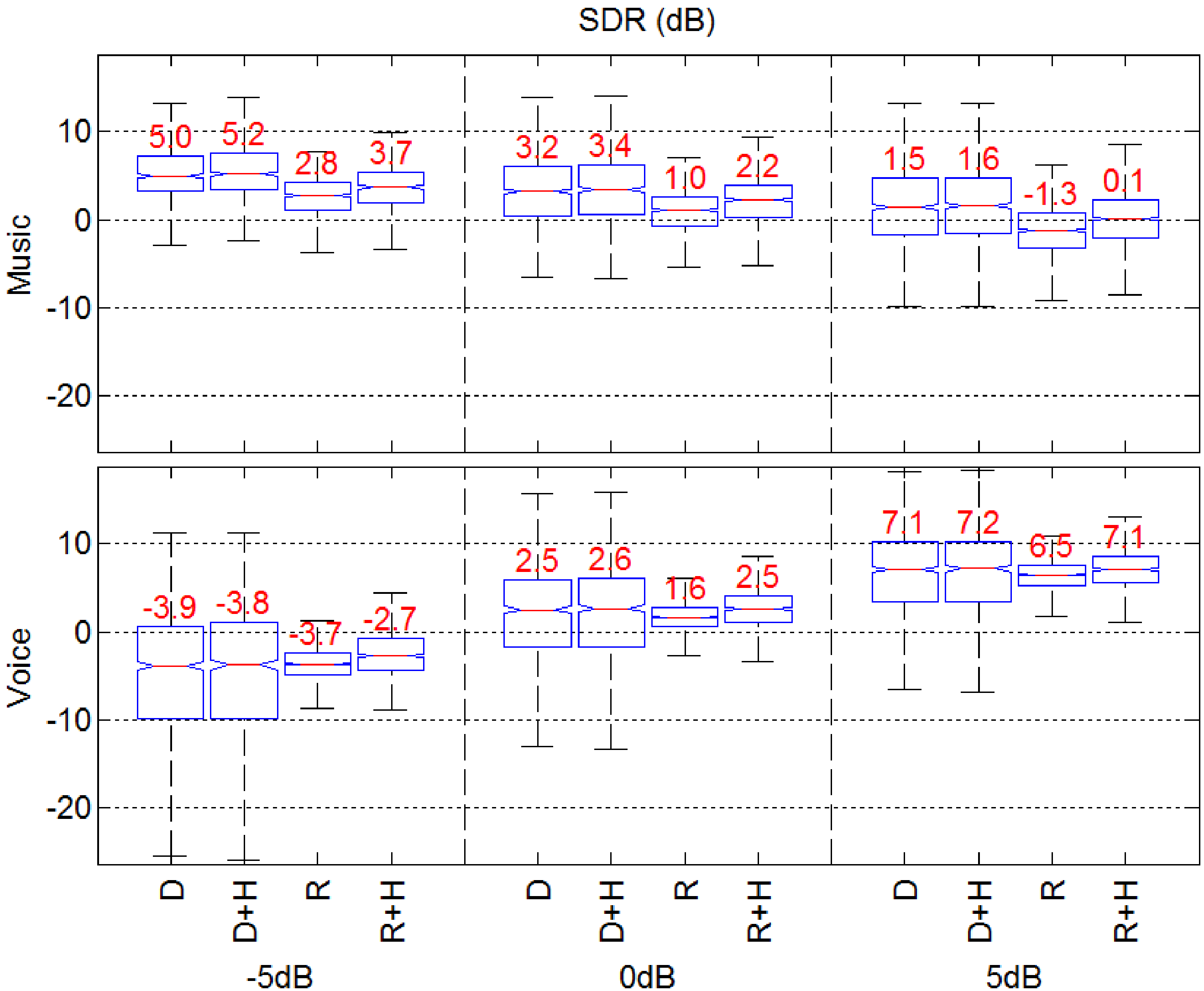
Evaluation

- **REPET** [Rafii et al., 2013]
  - 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



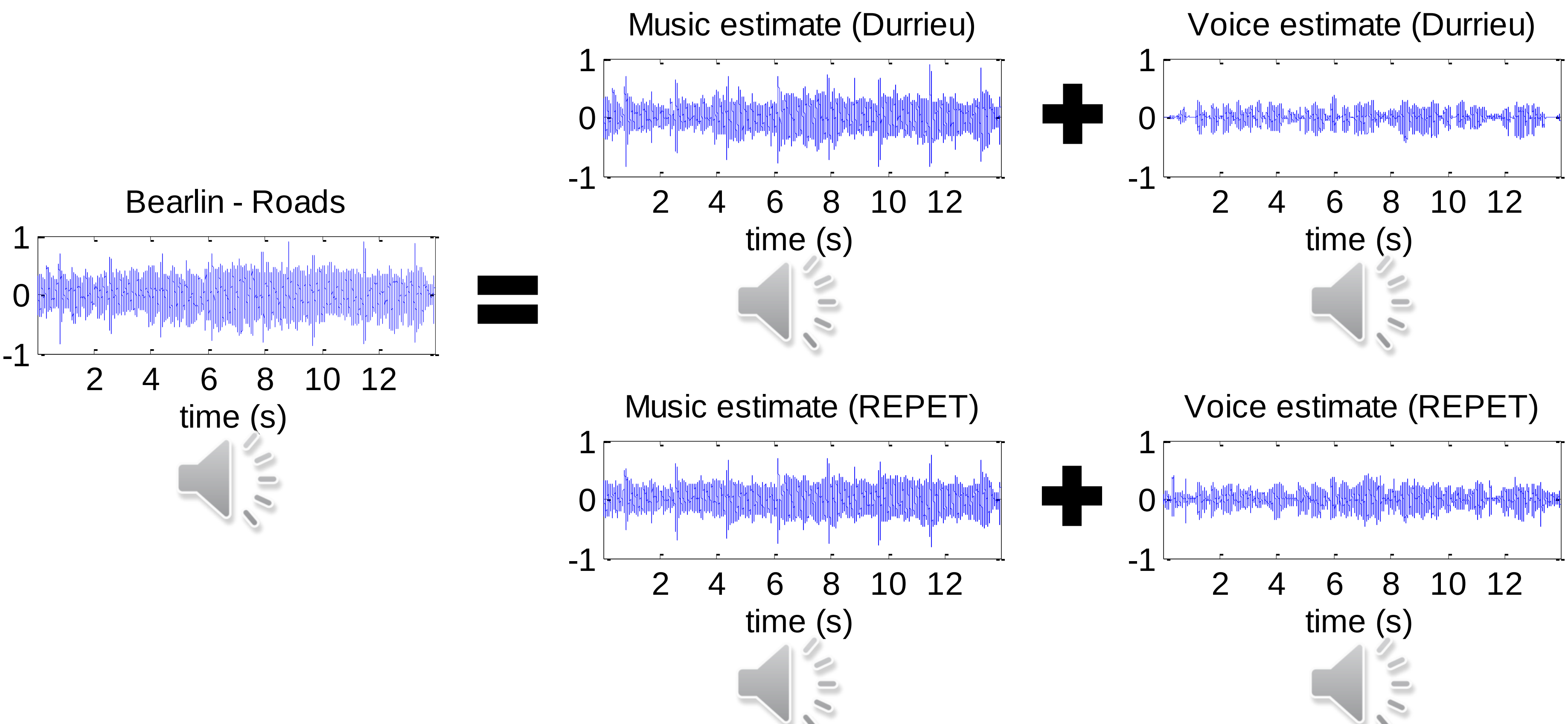
D = Durrieu
D+H = Durrieu + High-pass
R = REPET
R+H = REPET + High-pass

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)

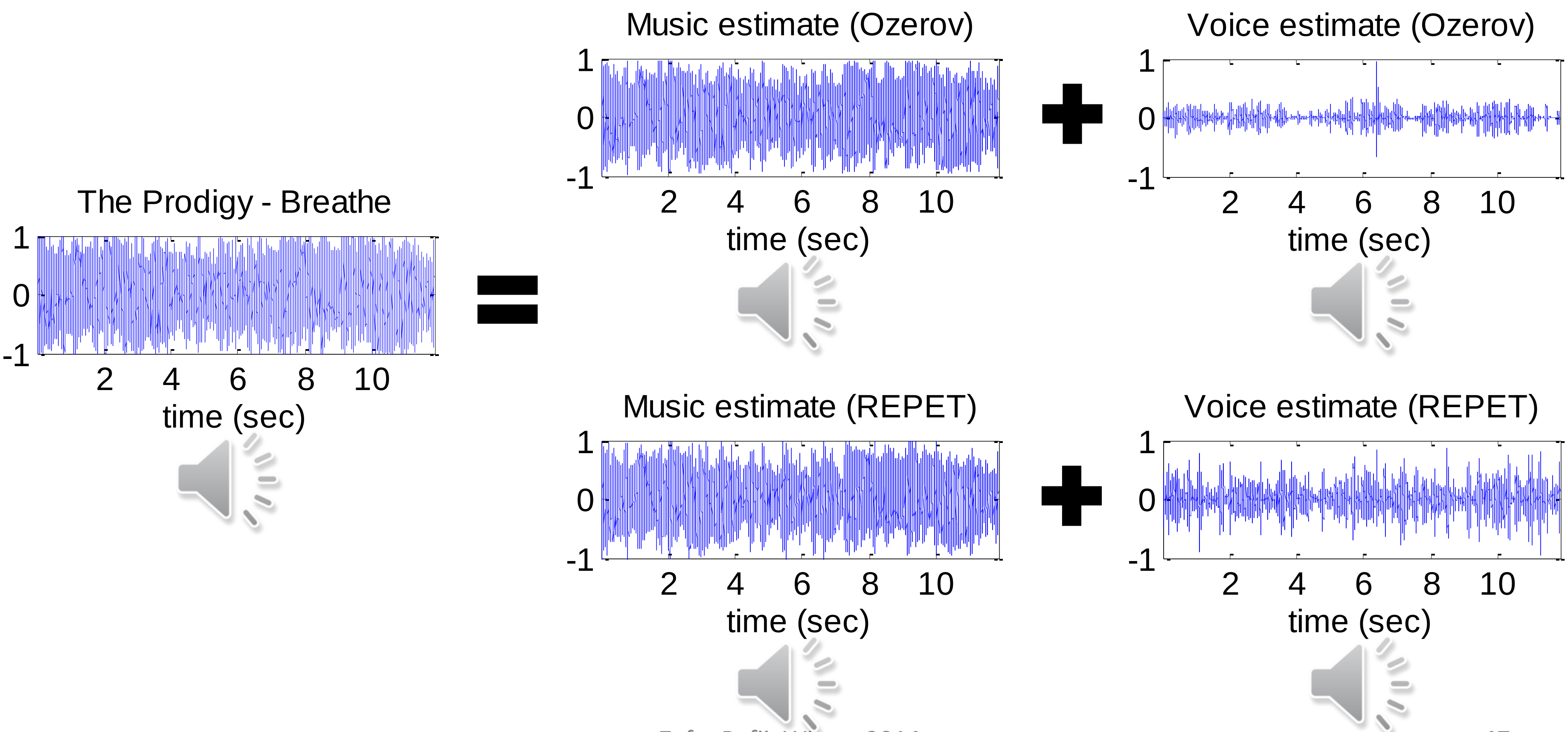
Examples

- REPET vs. Durrieu (source-filter + NMF)



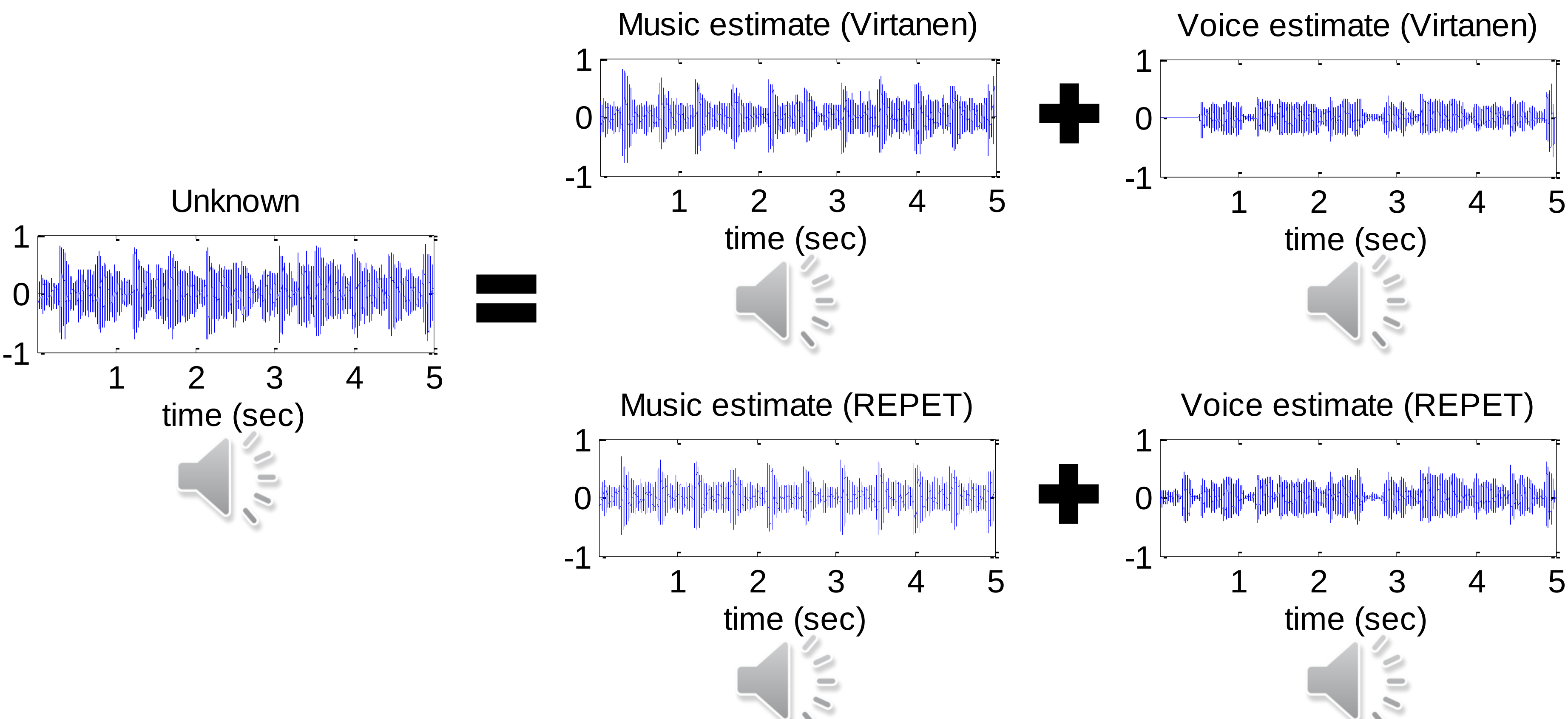
Examples

- REPET vs. Ozerov (accompaniment modeling)



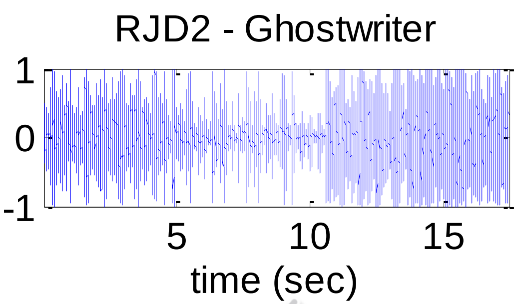
Examples

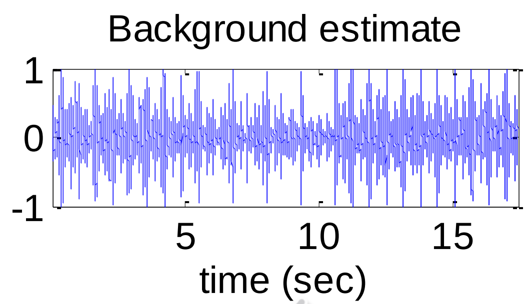
- REPET vs. Virtanen (NMF + pitch-based)

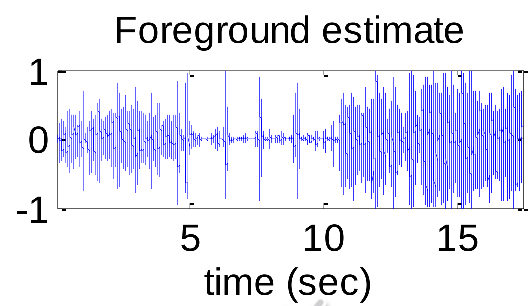


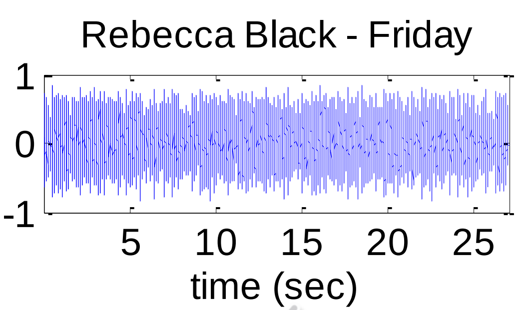
Examples

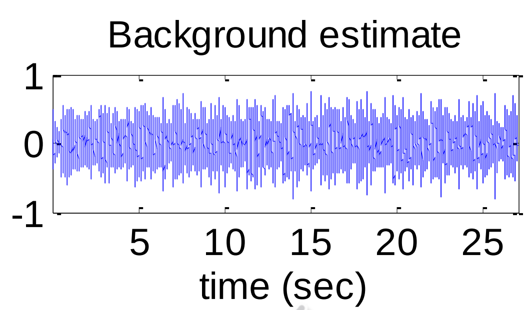
- REPET (more examples...)

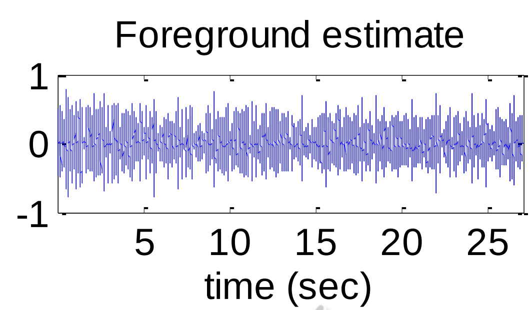




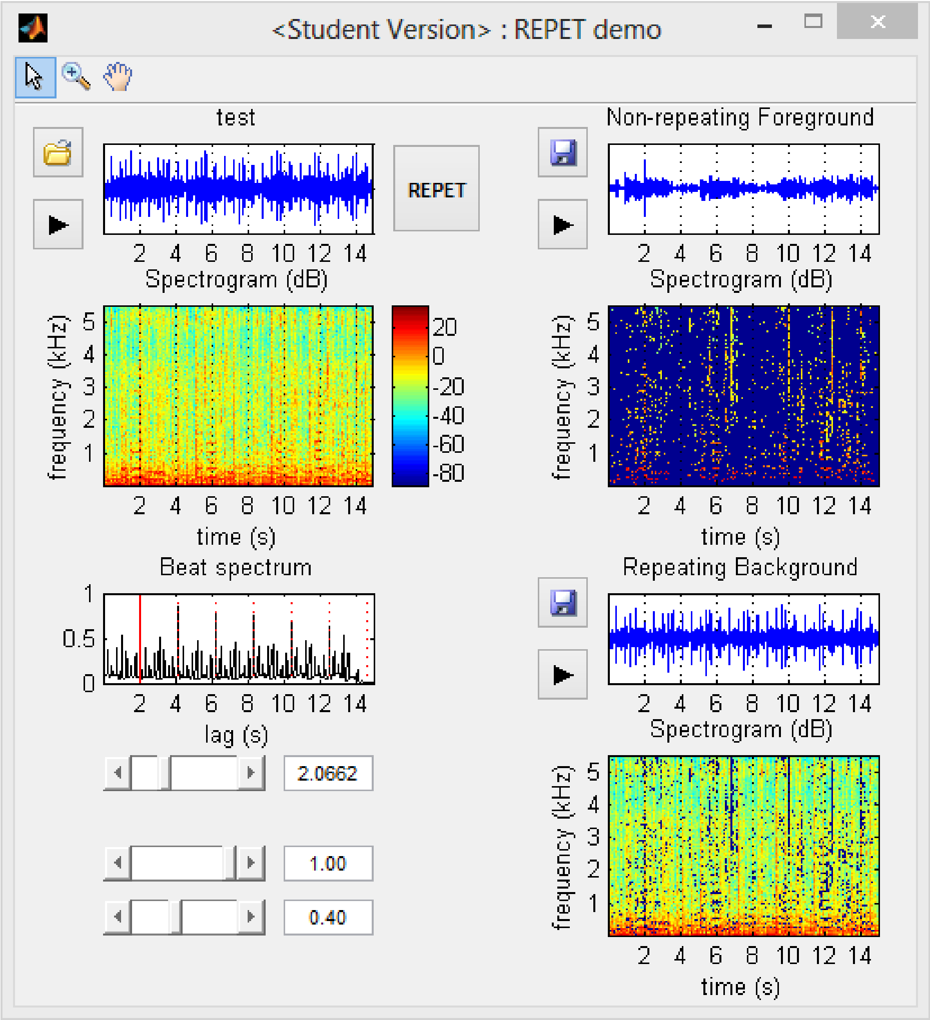








Demo



<Student Version> : REPET demo
test
REPET
Non-repeating Foreground
2 4 6 8 10 12 14
Spectrogram (dB)
frequency (kHz)
5
4
3
2
1
20
0
-20
-40
-60
-80
time (s)
Beat spectrum
1
0.5
0
lag (s)
Repeating Background
2.0662
1.00
0.40

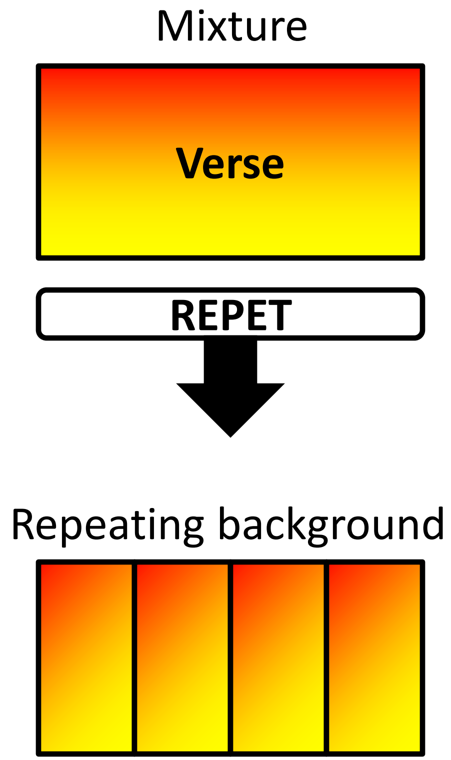
Thank you!

References

- 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.
- 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.
- A. Liutkus, Z. Rafii, R. Badeau, B. Pardo, and G. Richard, “Adaptive Filtering for Music/Voice Separation exploiting the Repeating Musical Structure,” in *37^{th} International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, March 25-30, 2012.
- J. H. McDermott, D. Wrobleski, and A. J. Oxenham, “Recovering Sound Sources from Embedded Repetition,” in *National Academy of Sciences*, vol. 108, pp. 1188-1193, 2011.
- A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, “Adaptation of Bayesian Models for Single-Channel Source Separation and its Application to Voice/Music Separation in Popular Songs,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1564-1578, July 2007.
- M. Piccardi, “Background Subtraction Techniques: a Review,” *IEEE International Conference on Systems, Man and Cybernetics*, The Hague, Netherlands, October 10-13, 2004.
- Z. Rafii and B. Pardo, “A Simple Music/Voice Separation Method based on the Extraction of the Repeating Musical Structure,” *36^{th} International Conference on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, May 22-27, 2011.
- Z. Rafii and B. Pardo, “Music/Voice Separation using the Similarity Matrix,” in *13^{th} International Society for Music Information Retrieval*, Porto, Portugal, October 8-12, 2012.
- Z. Rafii and B. Pardo, “REpeating Pattern Extraction Technique (REPET): A Simple Method for Music/Voice Separation,” in *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 21, no. 1, pp. 22-27, January, 2013.
- T. Virtanen, A. Mesaros, and M. Ryynänen, “Combining Pitch-based Inference and Non-Negative Spectrogram Factorization in Separating Vocals from Polyphonic Music,” *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition*, Brisbane, Australia, pp. 17-20, September 21, 2008.

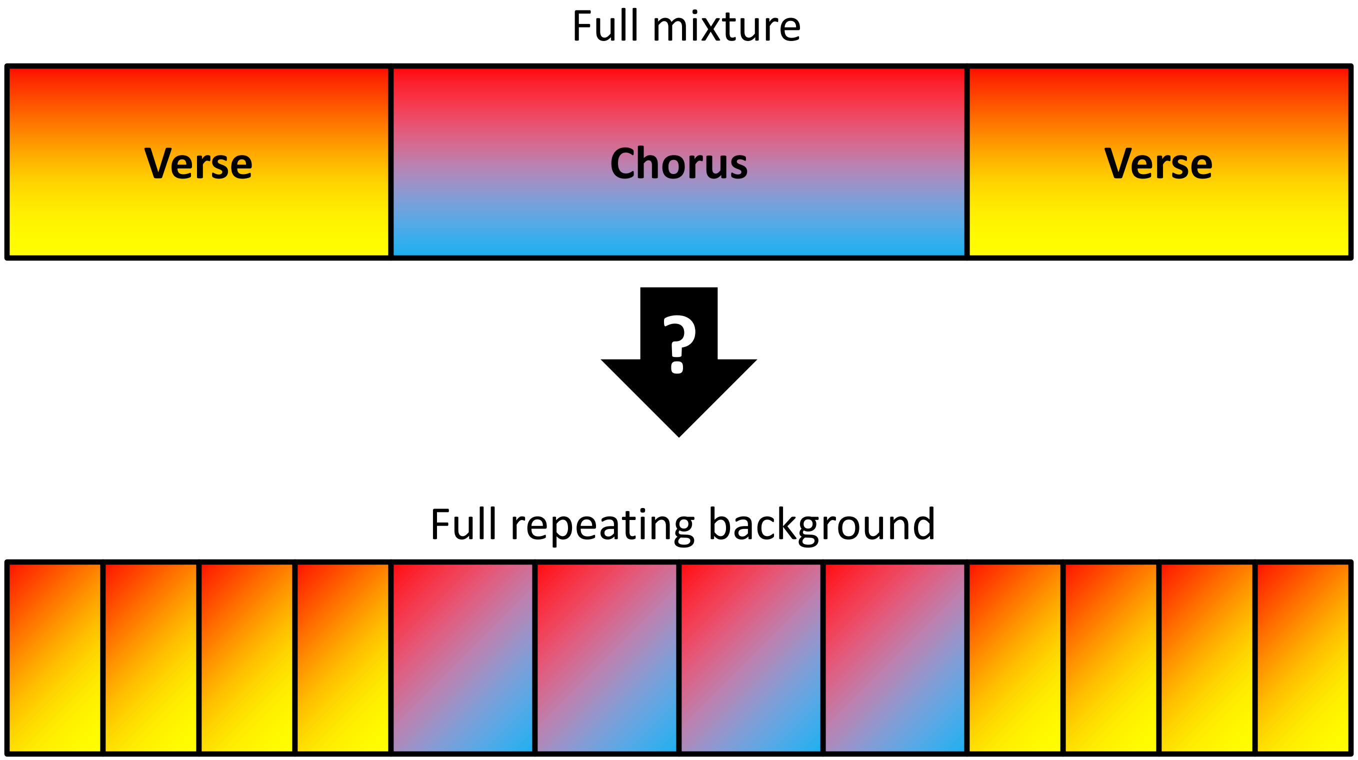
Extensions

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



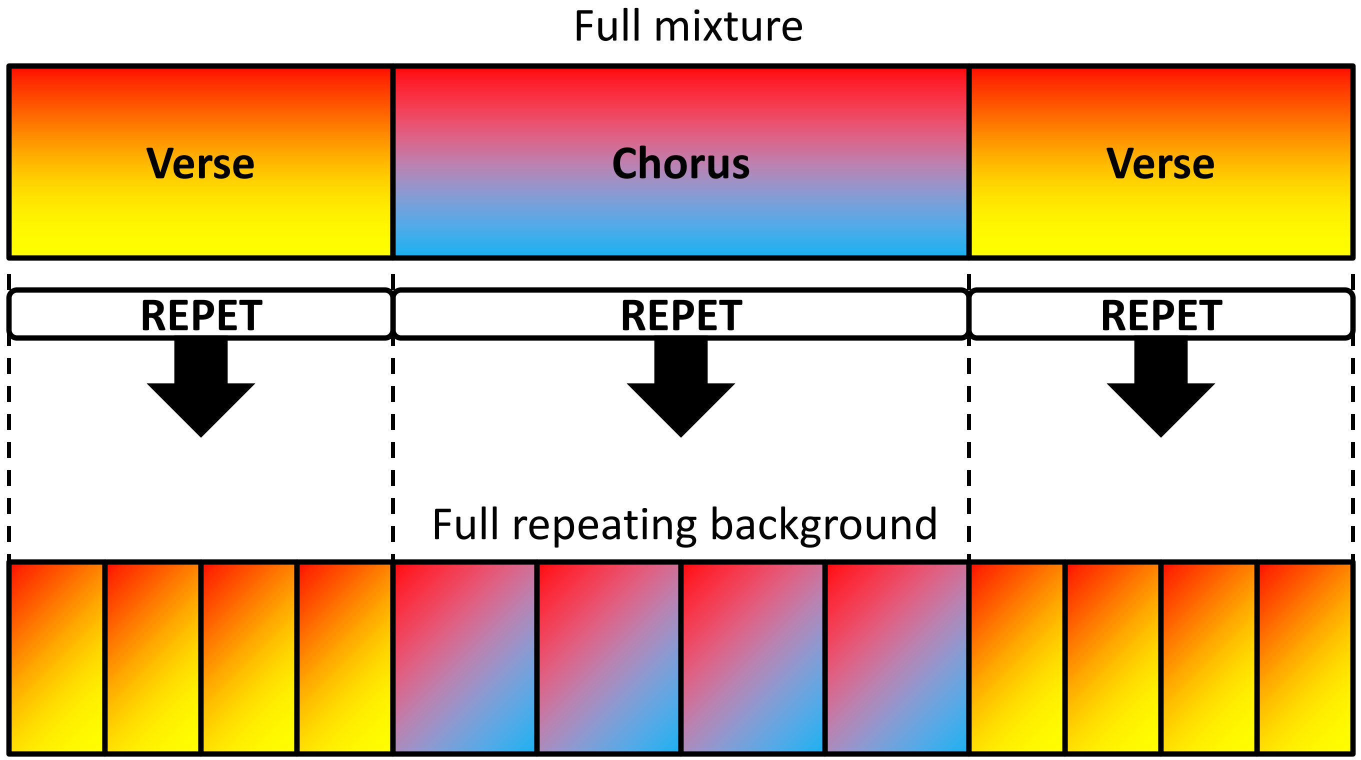
Extensions

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



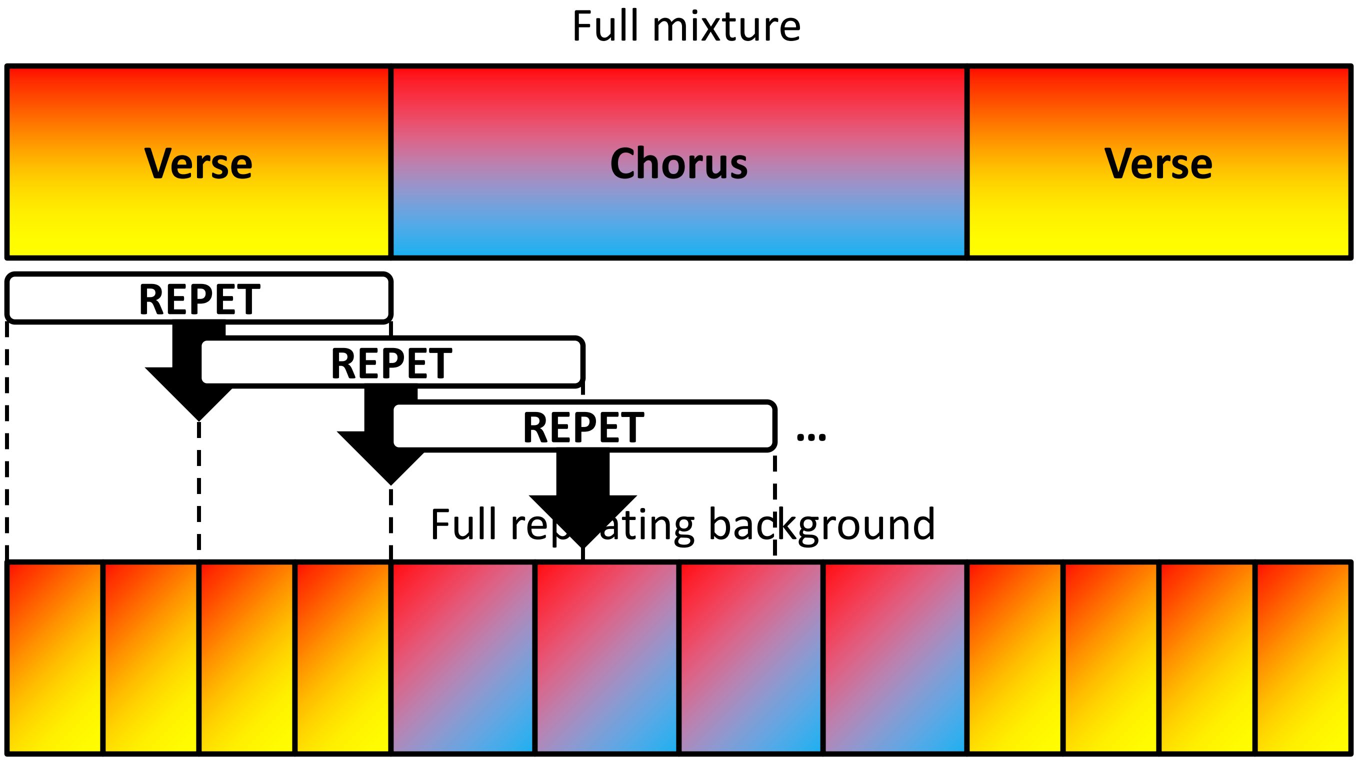
Prior Segmentation

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



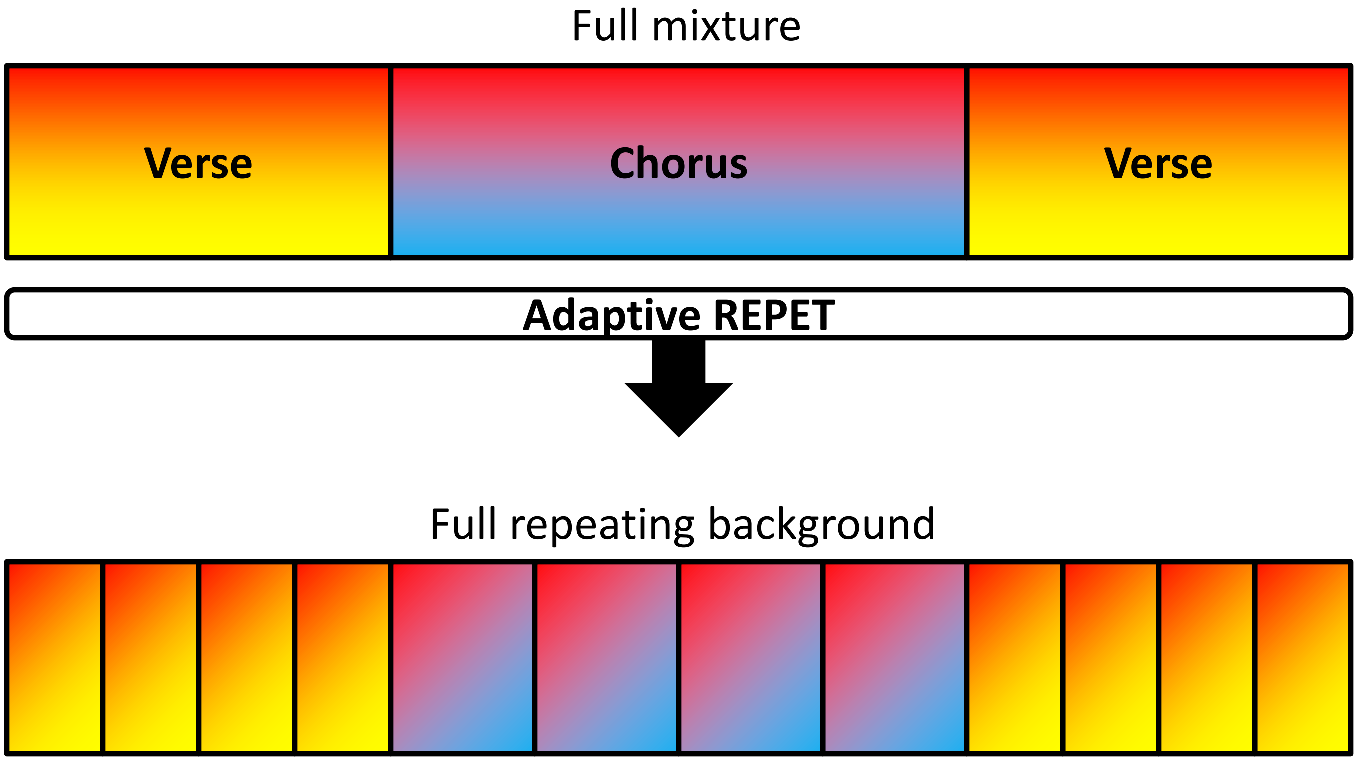
Sliding Window

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

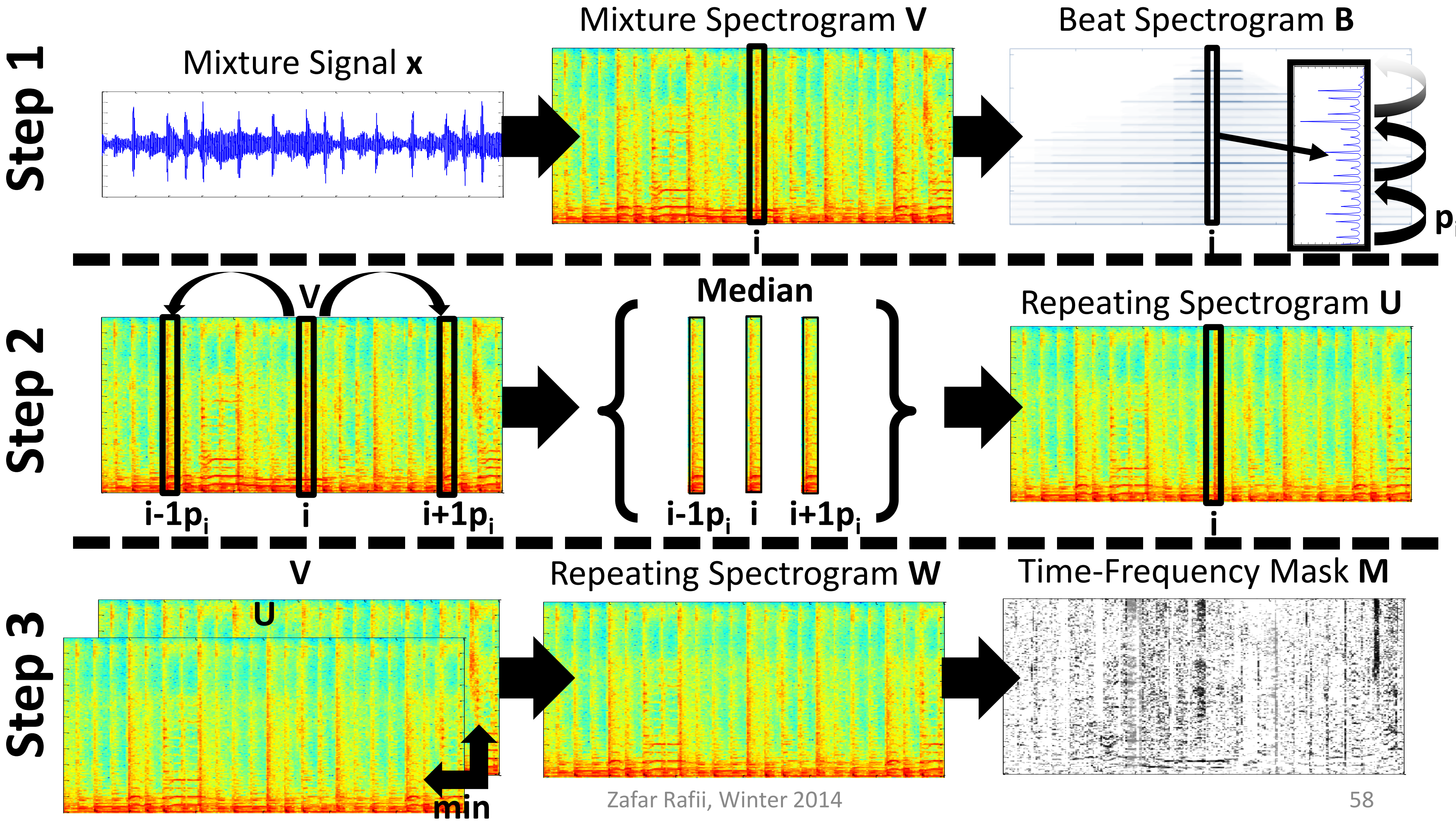


Adaptive REPET

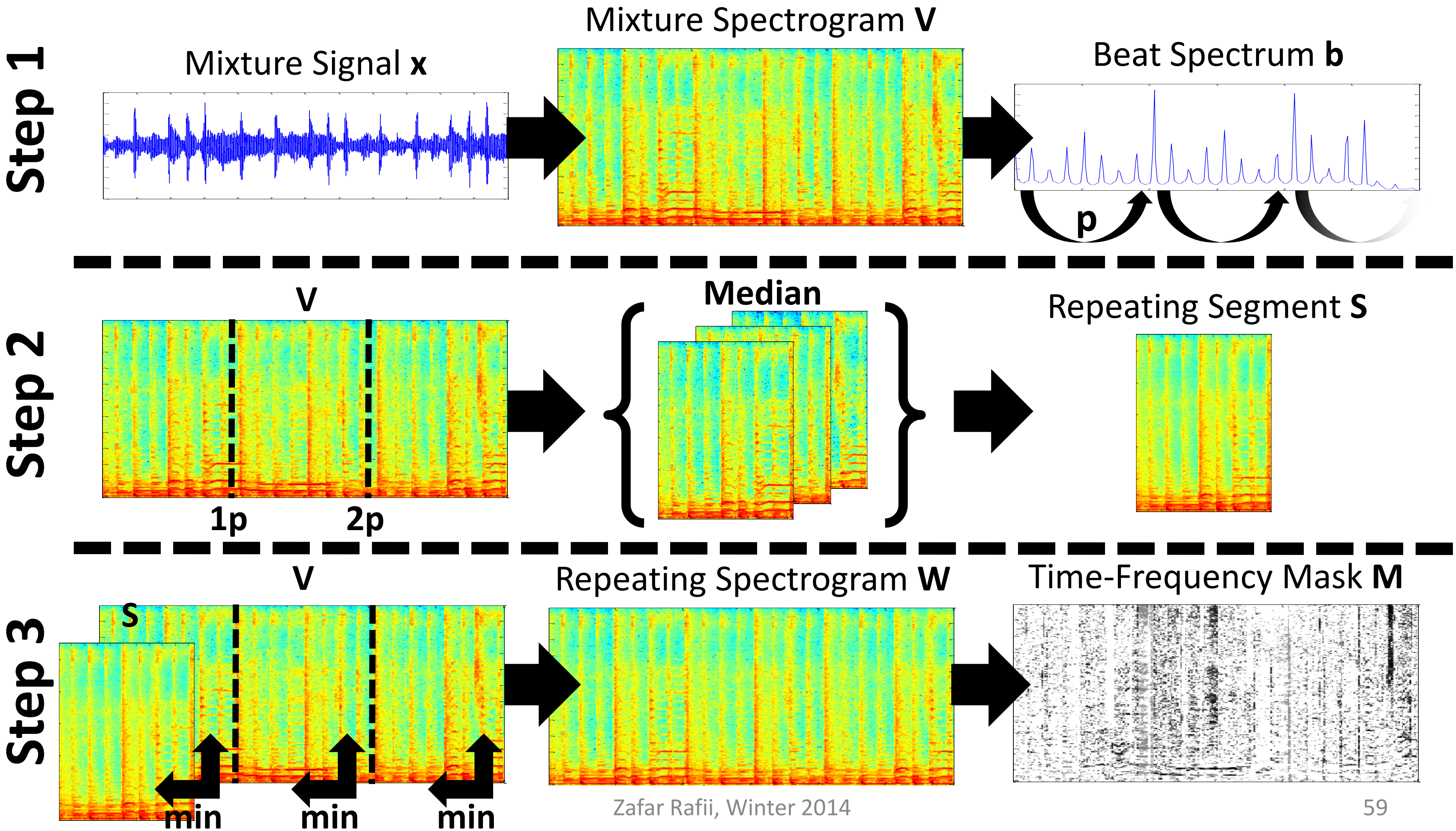
- We could directly **adapt REPET** along time by locally modeling the repeating background



Adaptive REPET

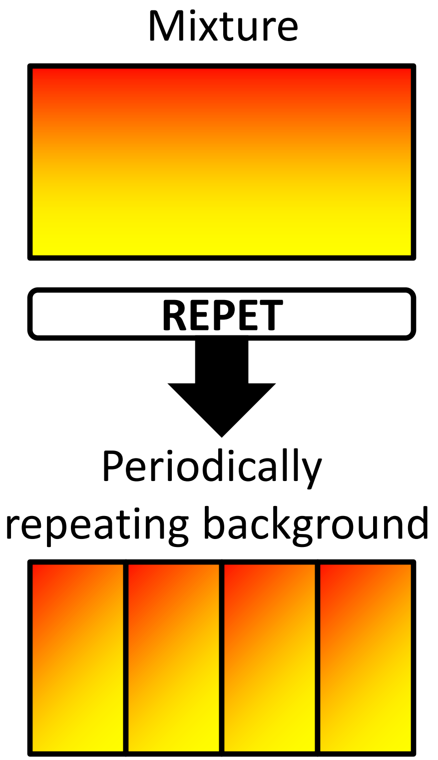


Original REPET



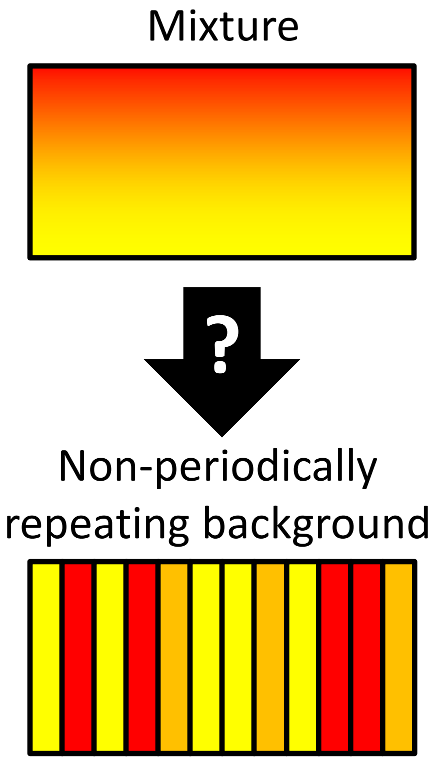
Generalization

- REPET (and its extension) assumes **periodically repeating patterns**



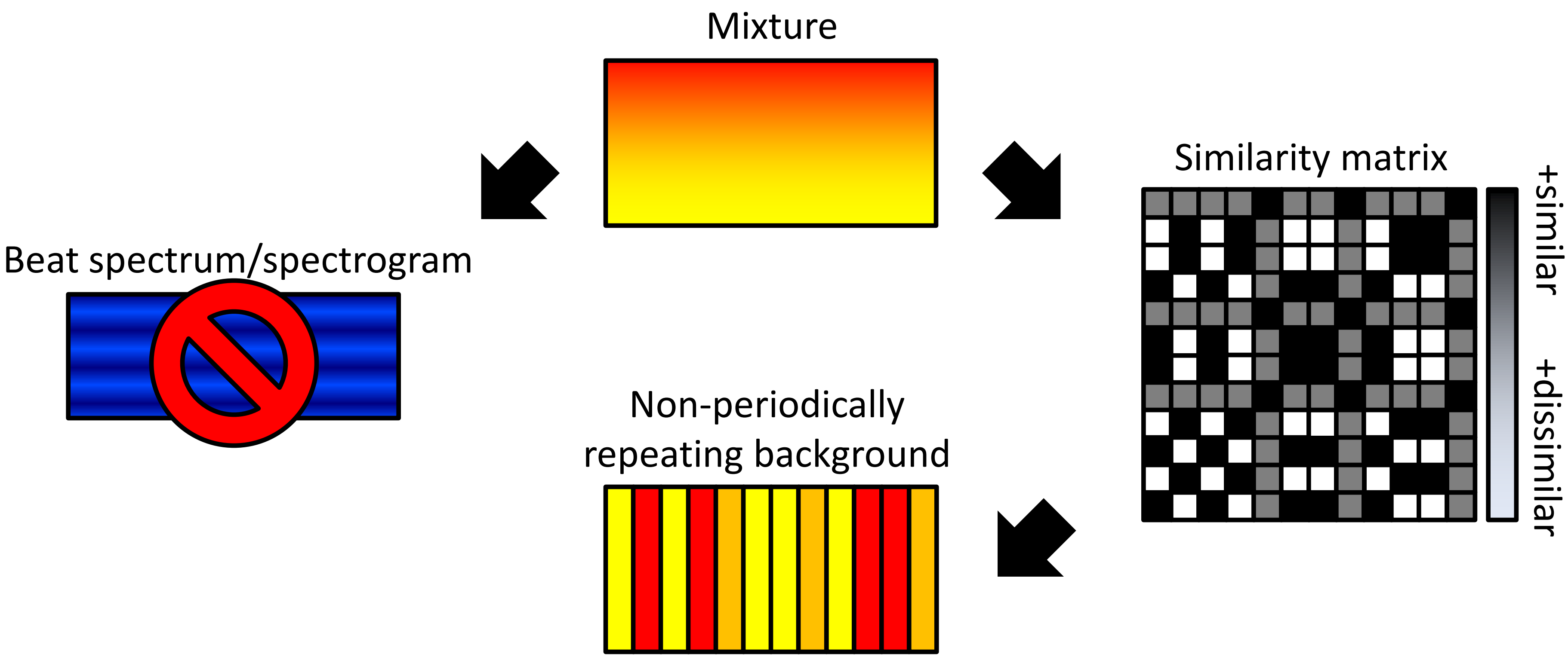
Generalization

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



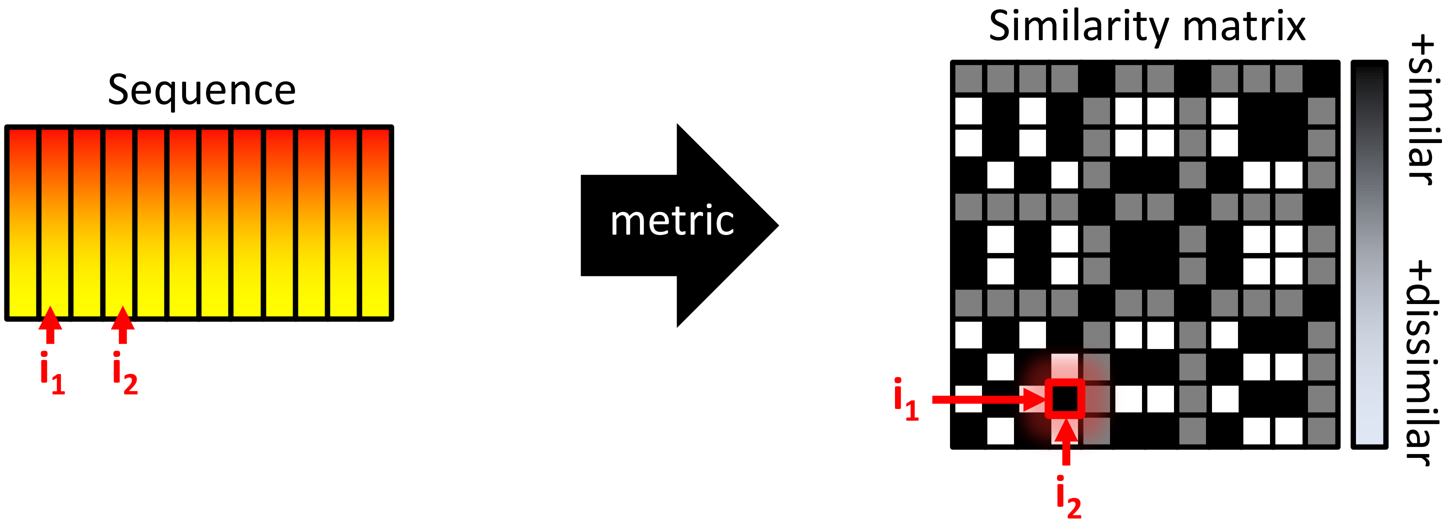
Generalization

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

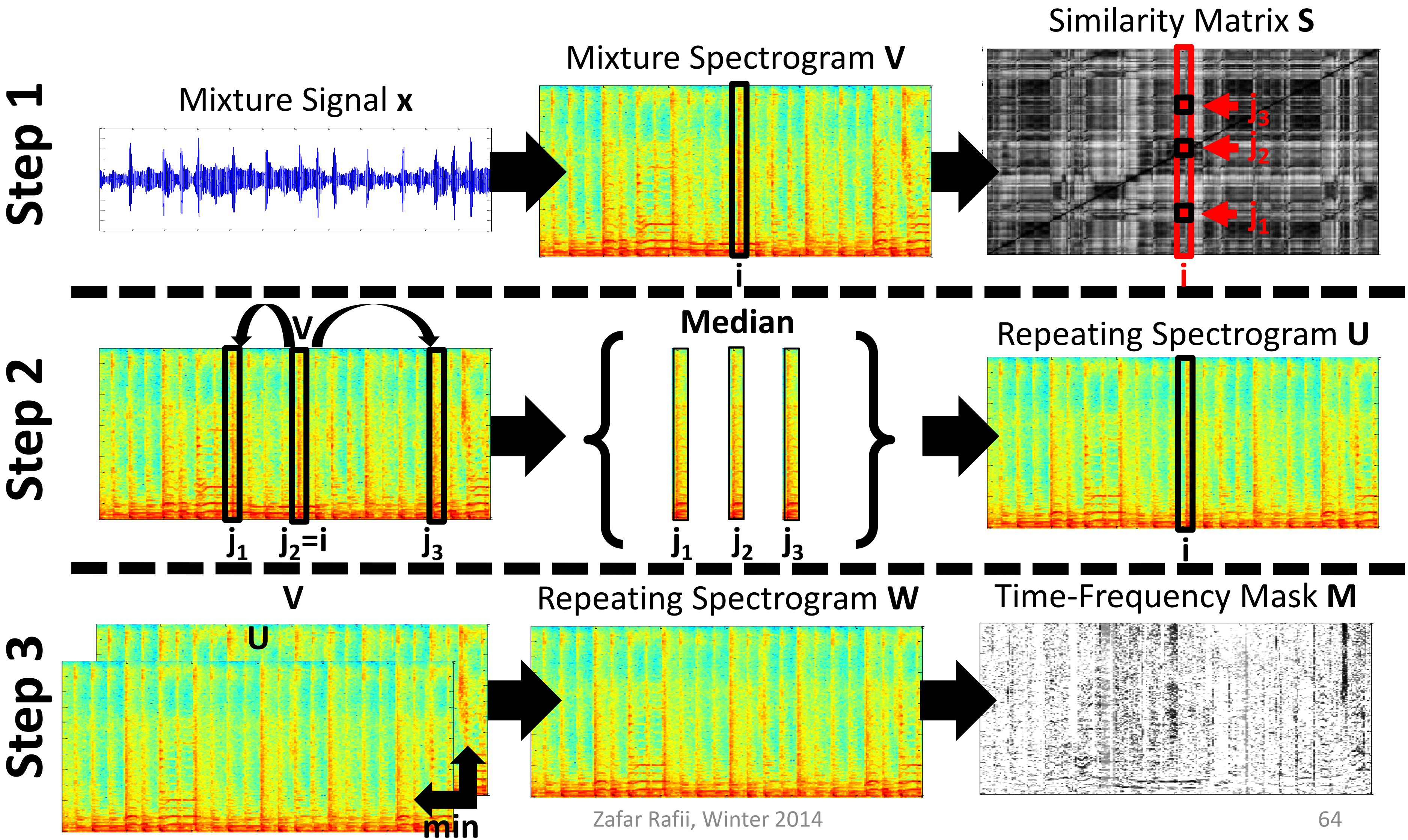


Generalization

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

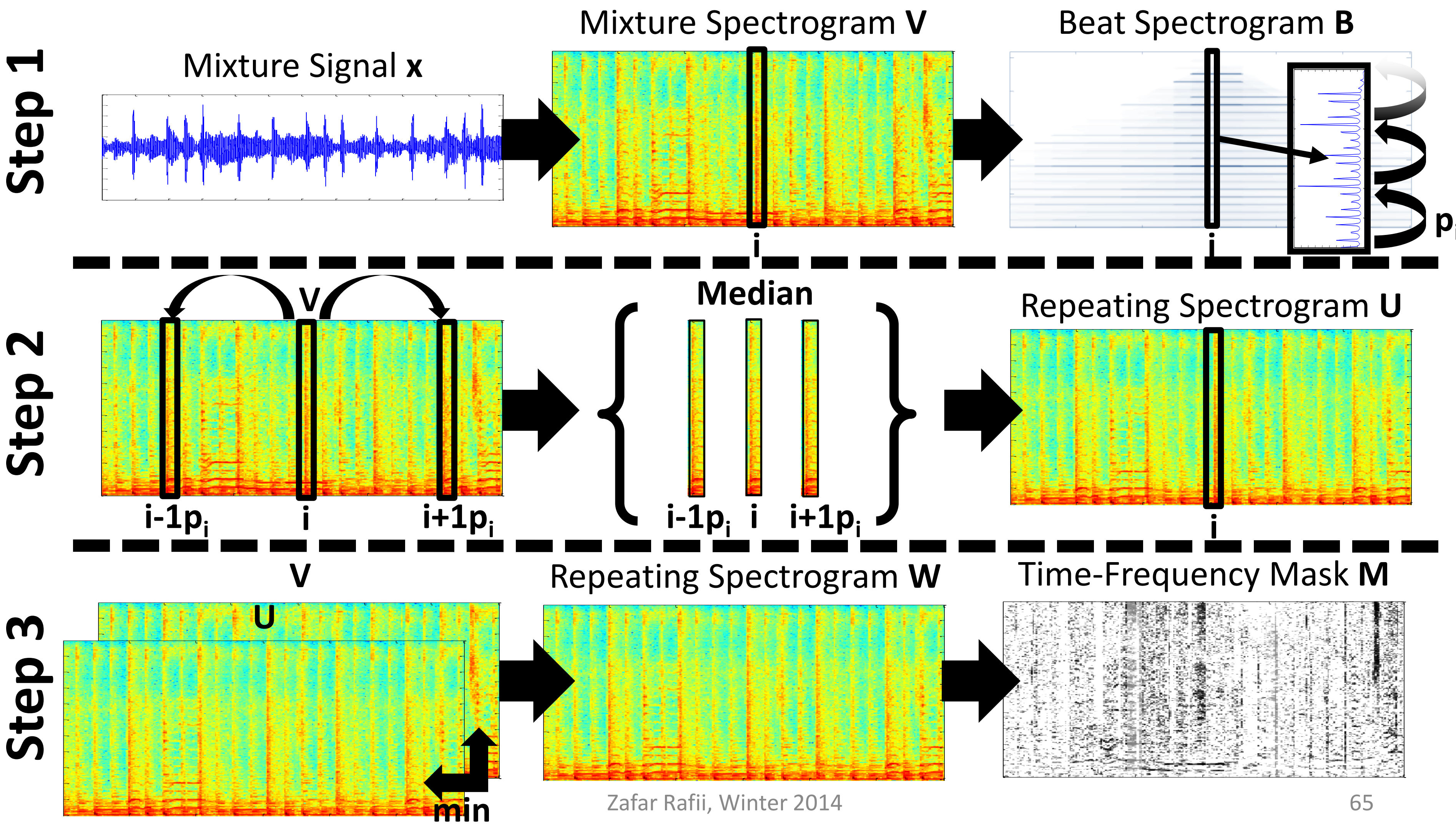


REPET-SIM



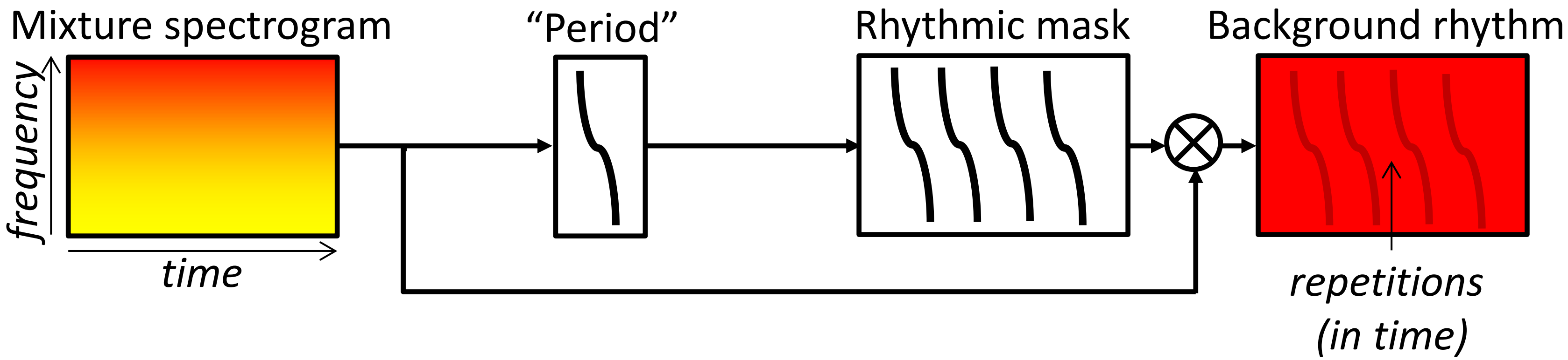
Step 1
Mixture Signal **x**
Mixture Spectrogram **V**
Similarity Matrix **S**
i
j1
j2
j3
i
Step 2
V
Median
Repeating Spectrogram **U**
j1
j2=i
j3
j1
j2
j3
i
Step 3
V
U
min
Repeating Spectrogram **W**
Time-Frequency Mask **M**

Adaptive REPET

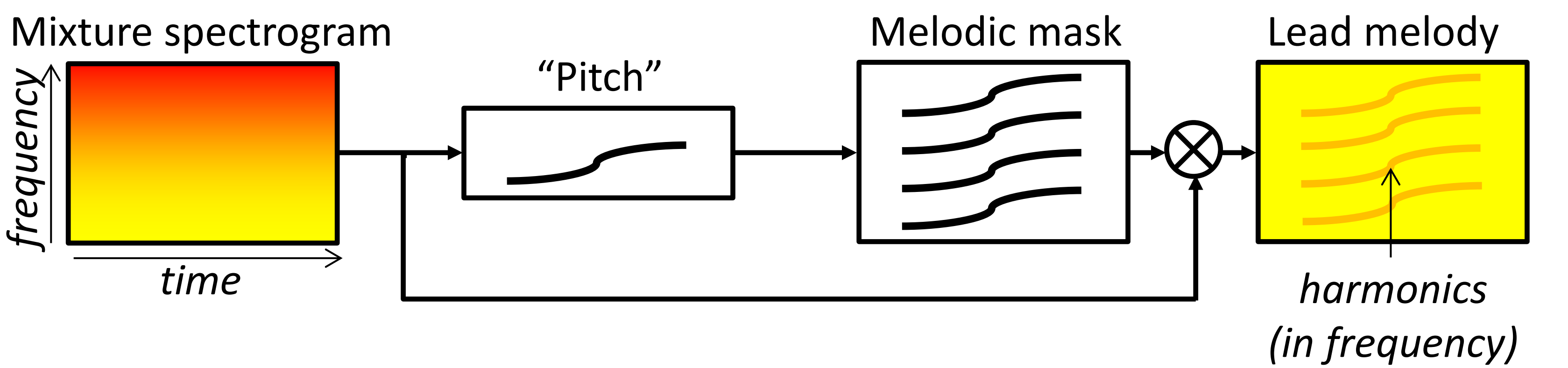


REPET + Pitch

- REPET models the **background rhythm**



- Pitch-based methods model the **lead melody**



REPET + Pitch

- **Auditory processing** in human listeners