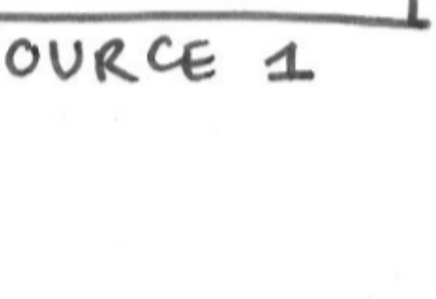
ON THE IMPORTANCE OF TEMPORAL CONTEXT IN PROXIMITY KERNELS: A VOCAL SEPARATION CASE STUDY

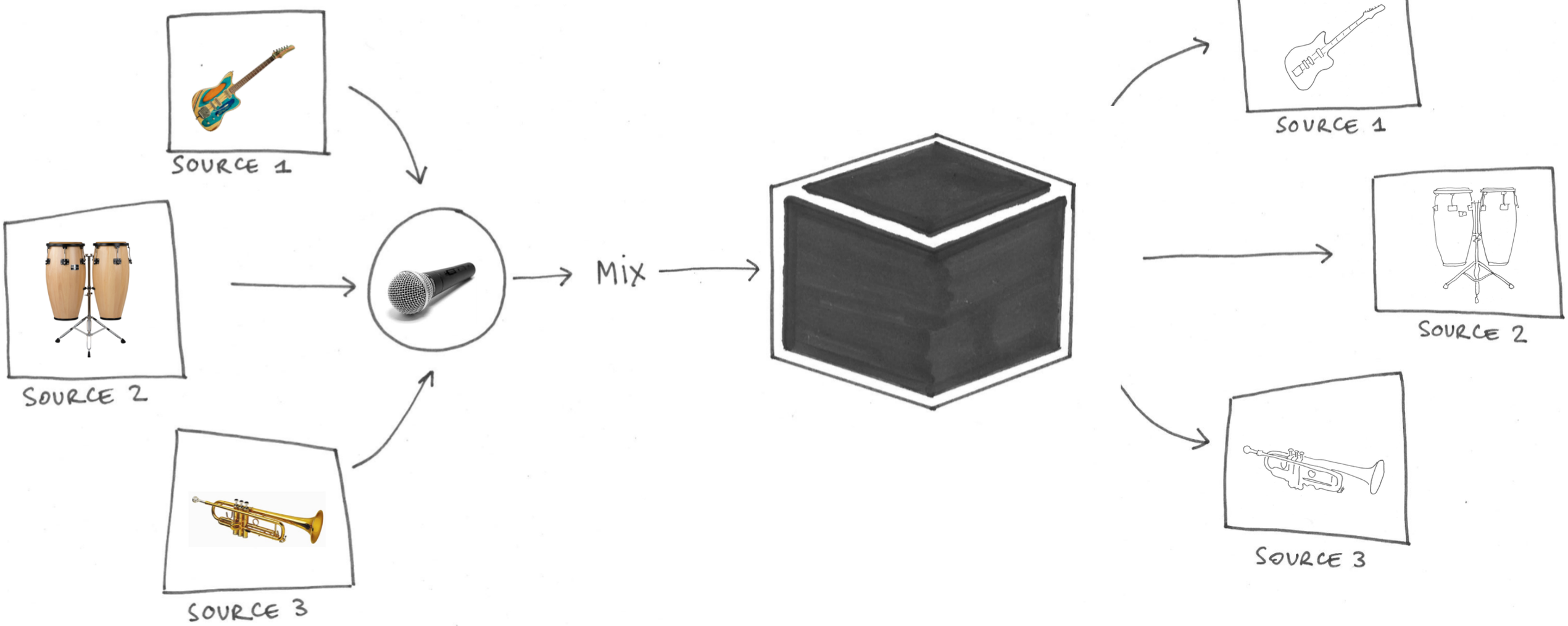
SOURCE 1

SOURCE 2



SOURCE 3

Mix


SOURCE 1
SOURCE 2
SOURCE 3
Mix
SOURCE 1
SOURCE 2
SOURCE 3

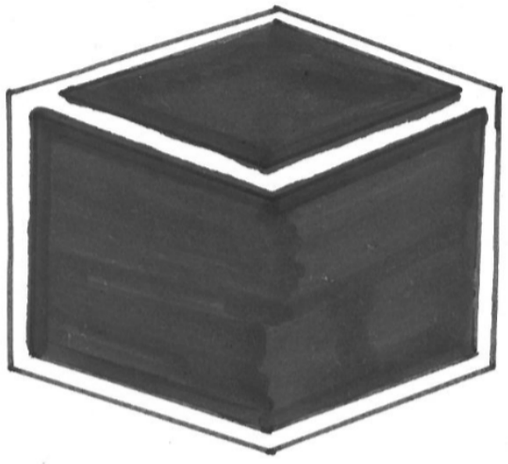
SOURCE SEPARATION

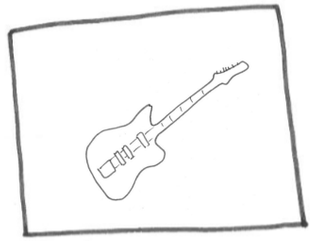
SOURCE 1

SOURCE 2

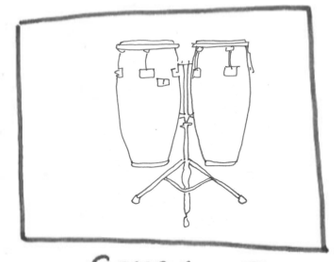
SOURCE 3

Mix



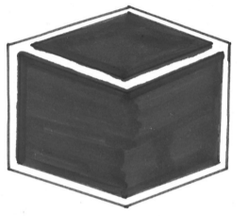


SOURCE 1

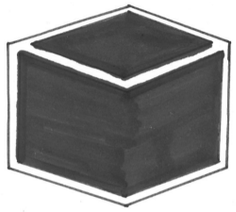


SOURCE 2

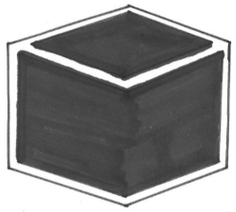
SOURCE 3



.... WHY IS IT USEFUL?

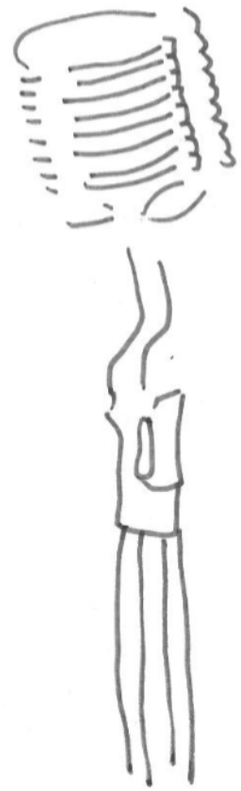


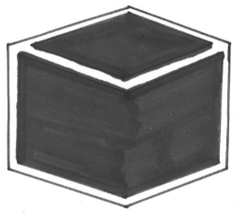
.... WHY IS IT USEFUL?



.... WHY IS IT USEFUL?

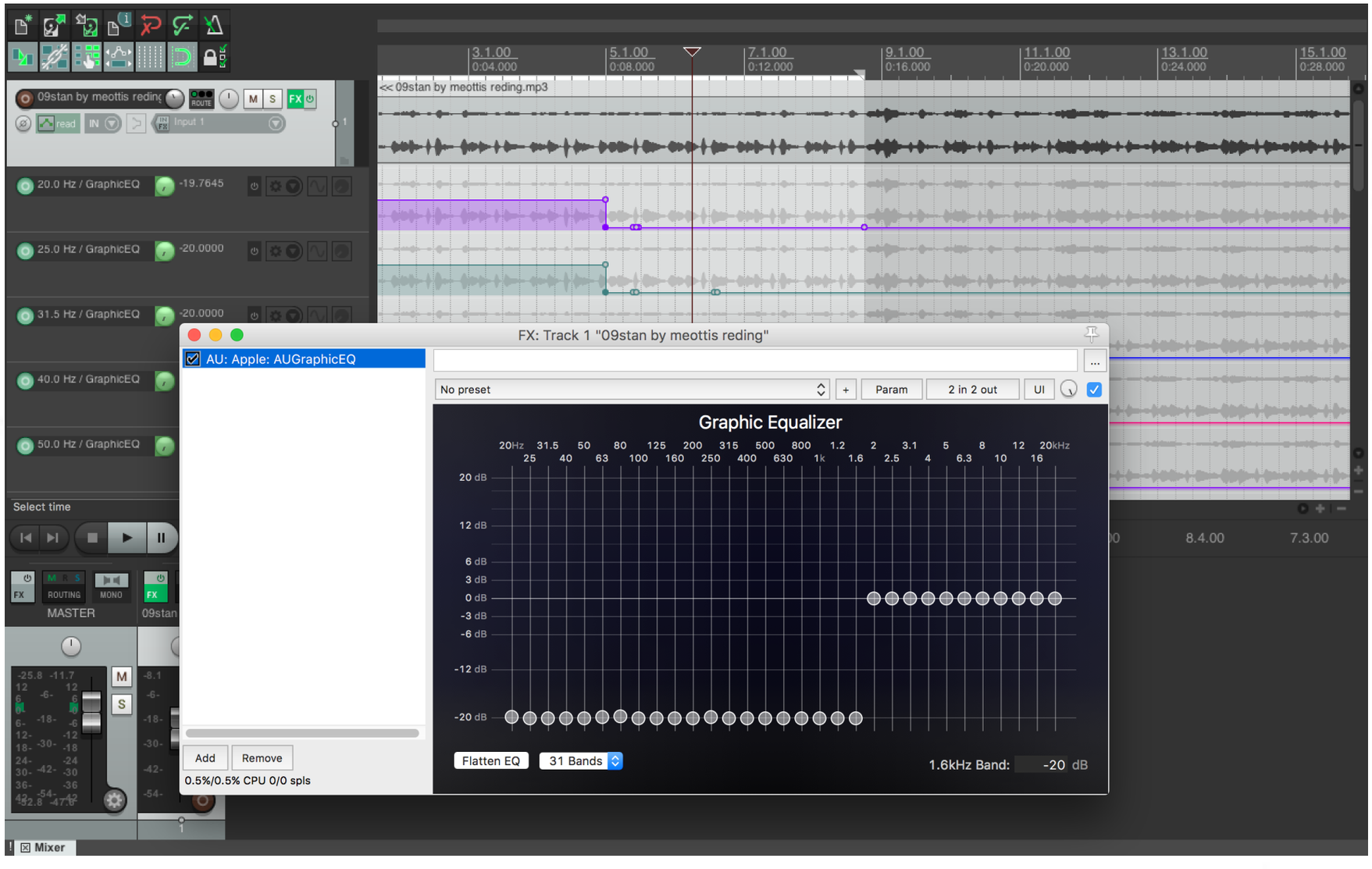
- UP MIXING
- SIGNAL DENOISING
- AUDIO RESTORATION
- KARAOKE
- MUSICAL FEATURE EXTRACTION
- AUTOMATIC TRANSCRIPTION
- ....

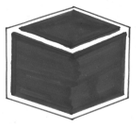




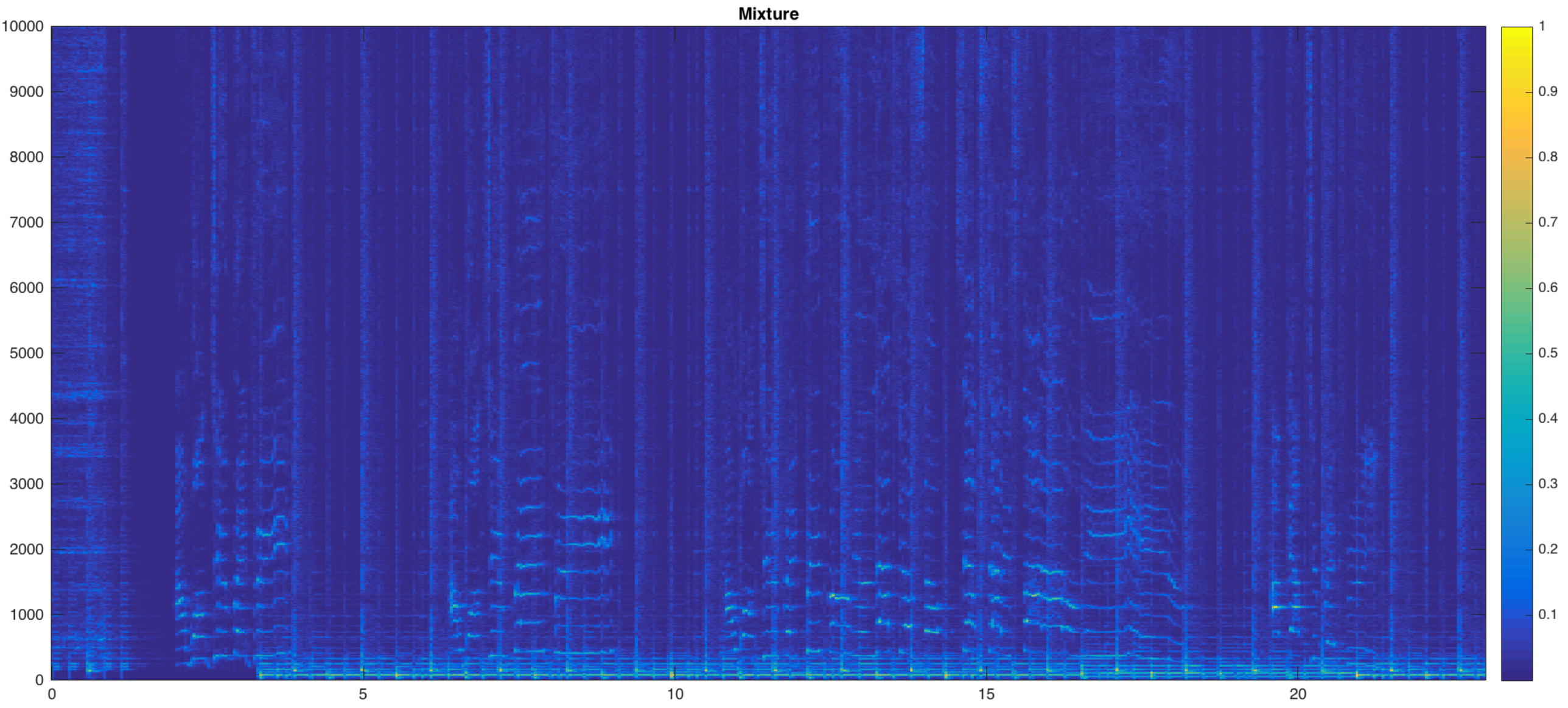
.... HOW DOES IT WORK?

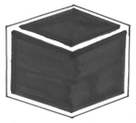
.... A SIMPLE EXAMPLE : NO OVERLAP



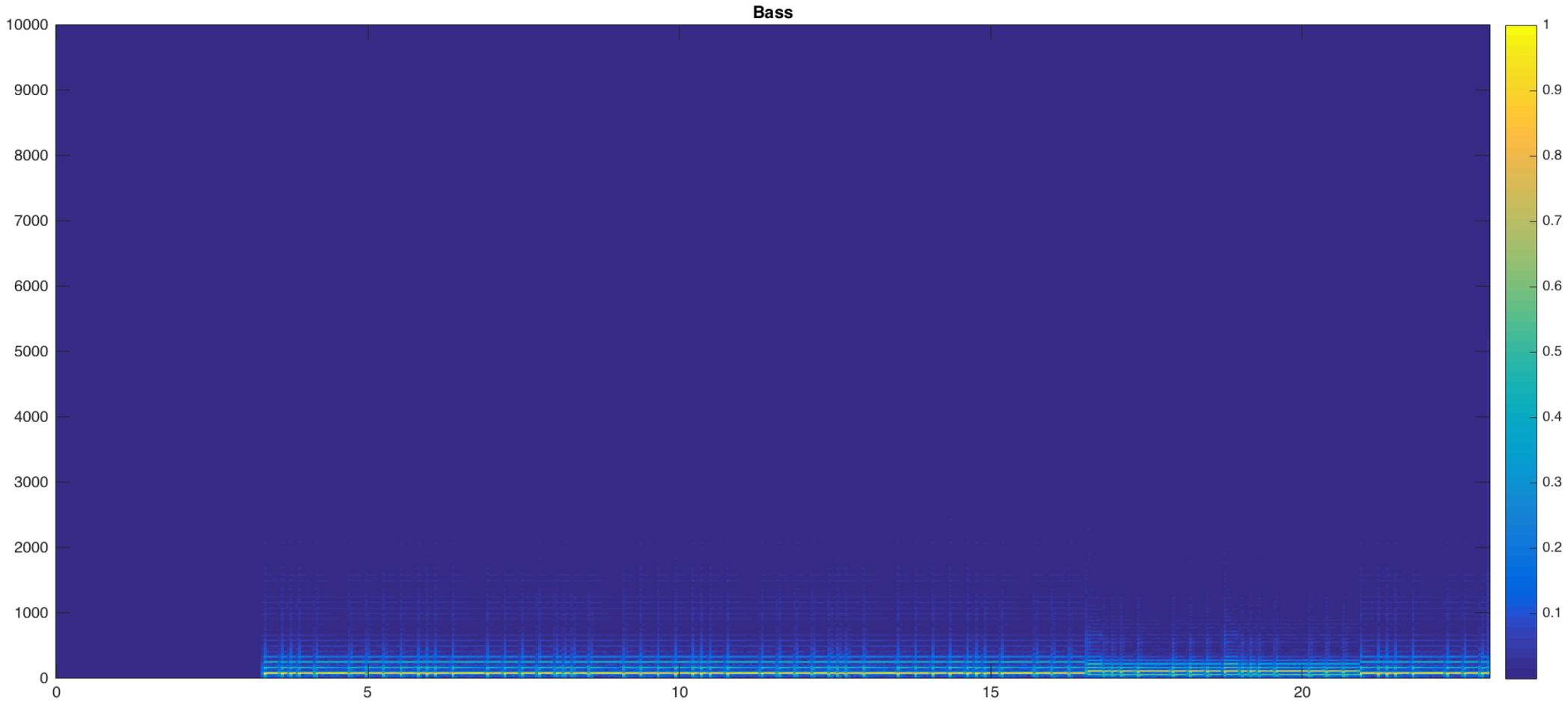


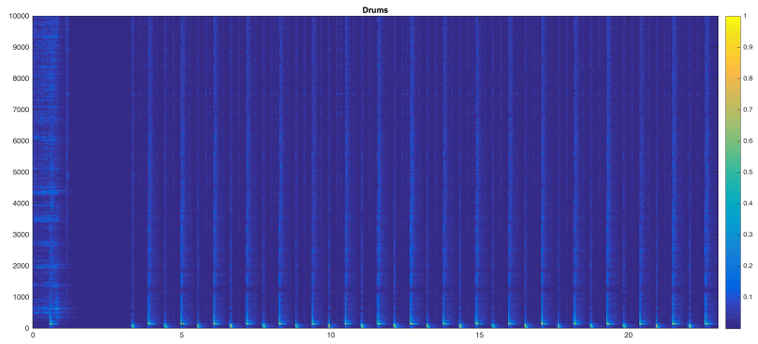
.... MOST POPULAR SONGS: SOURCES DO OVERLAP

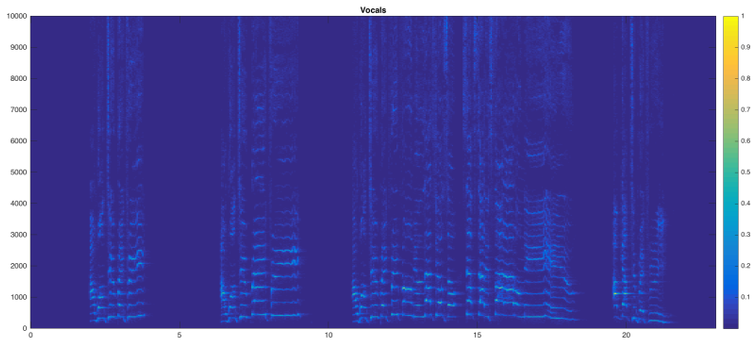


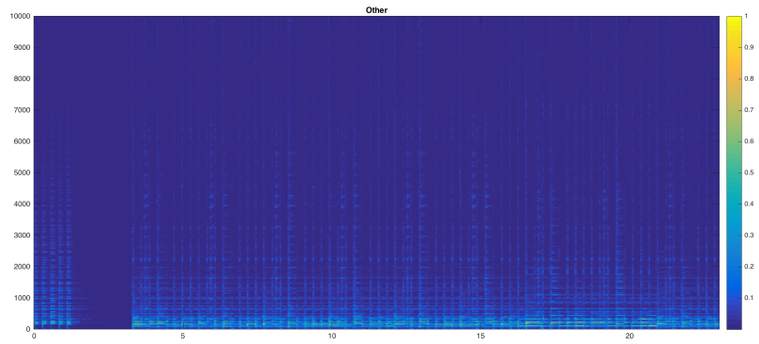


. . . . WHAT WE ARE AIMING FOR

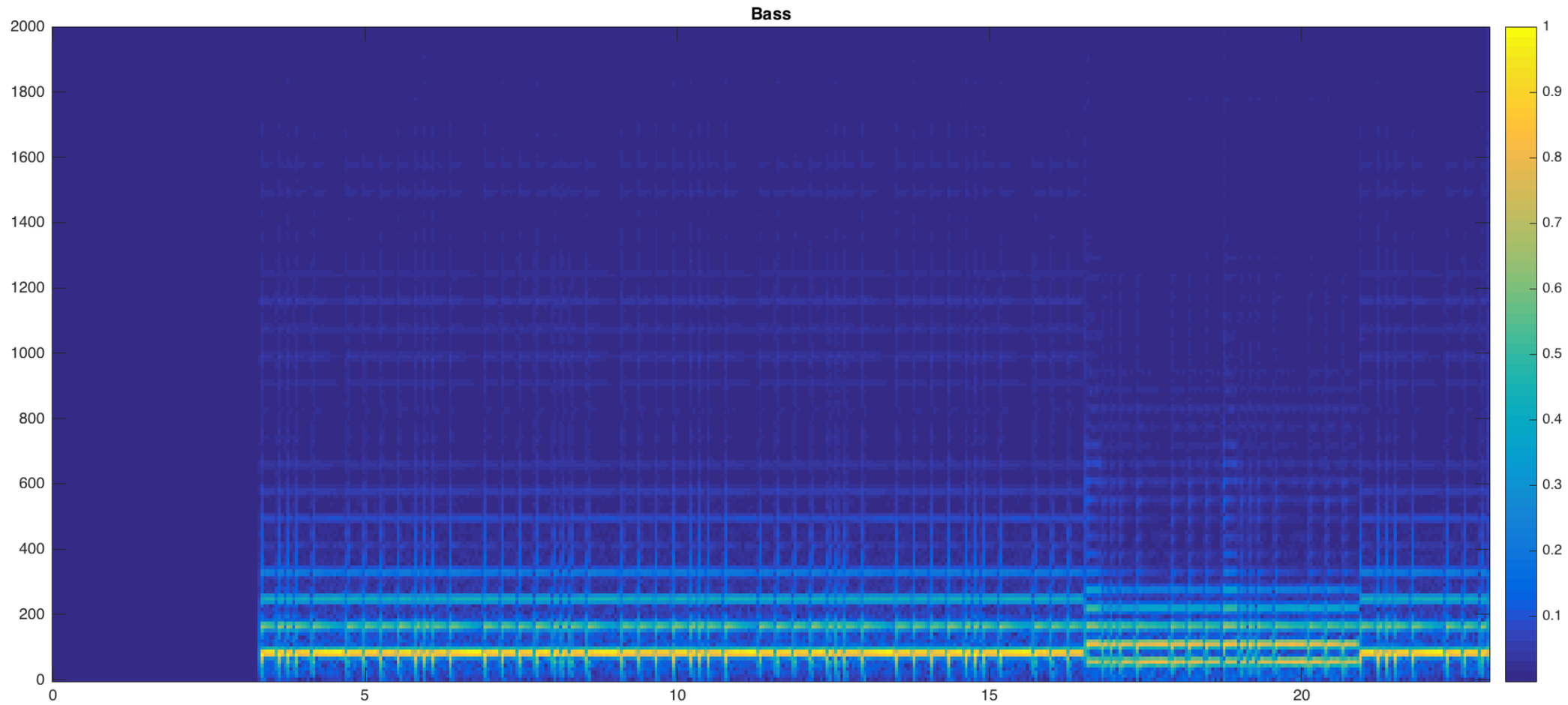


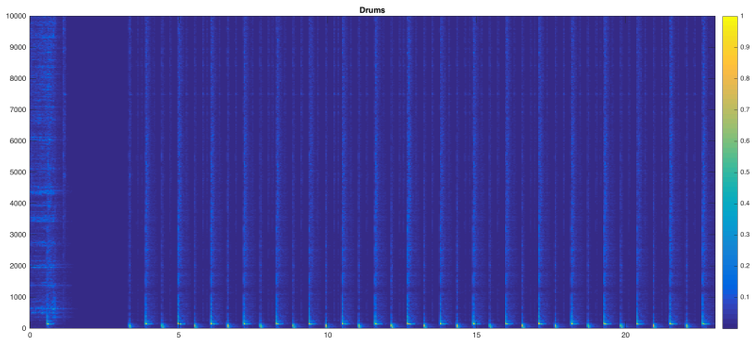


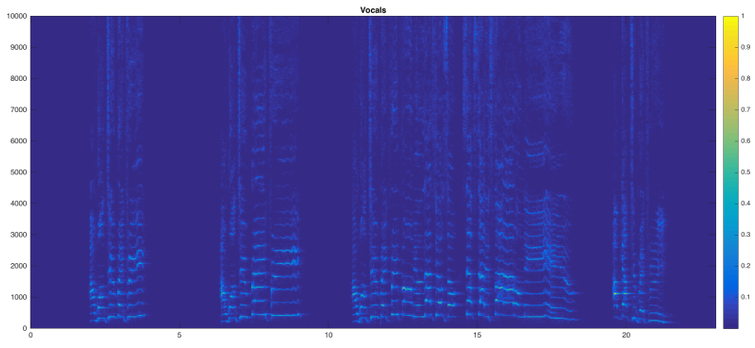


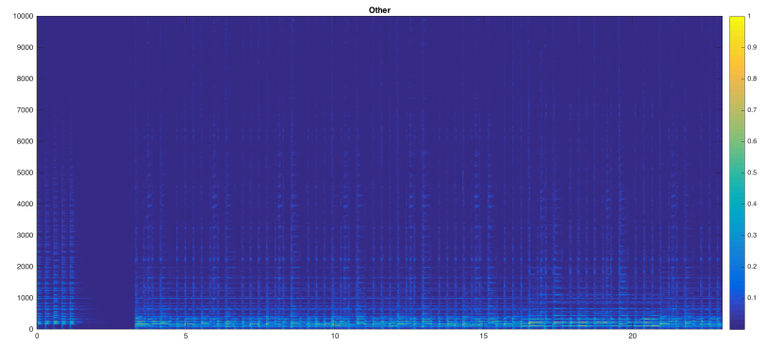


. . . . WHAT WE ARE AIMING FOR

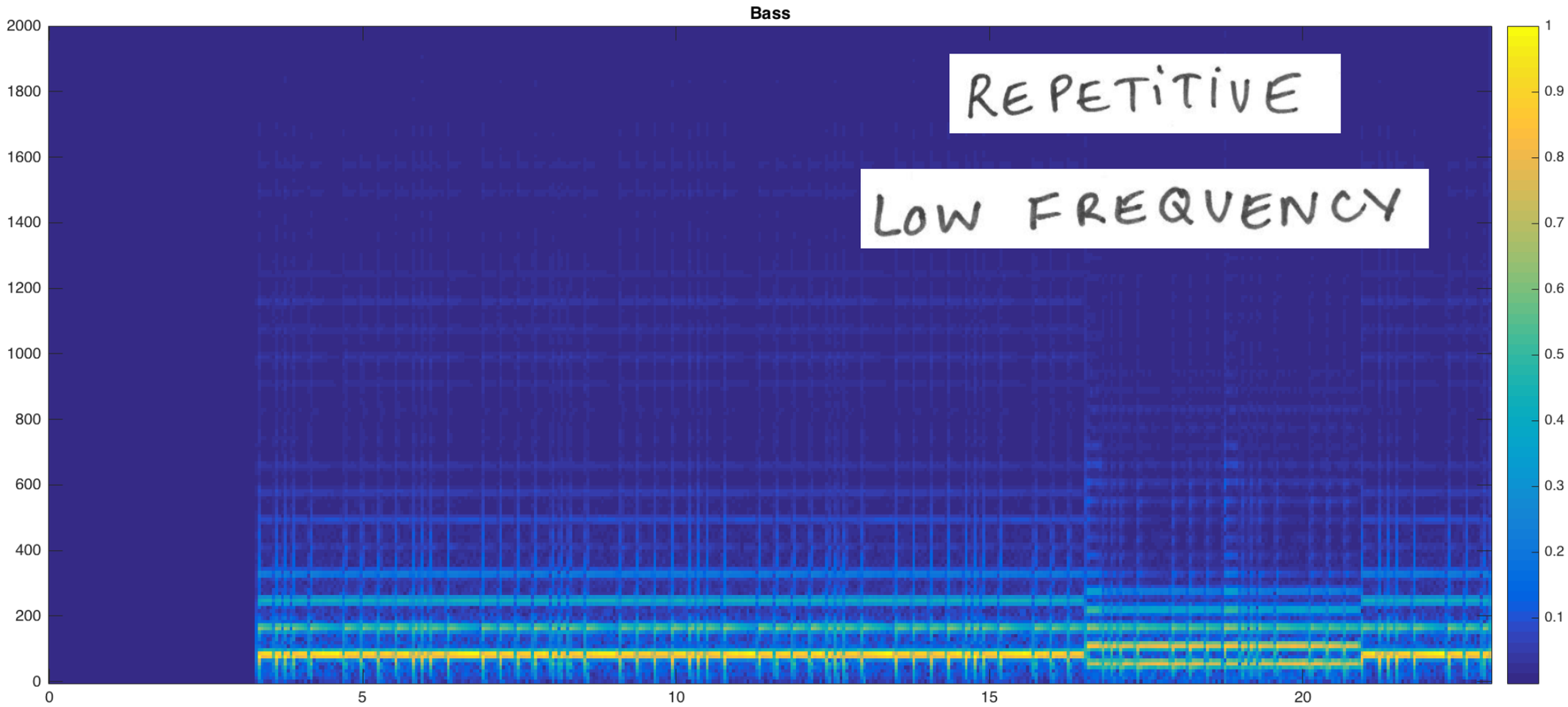


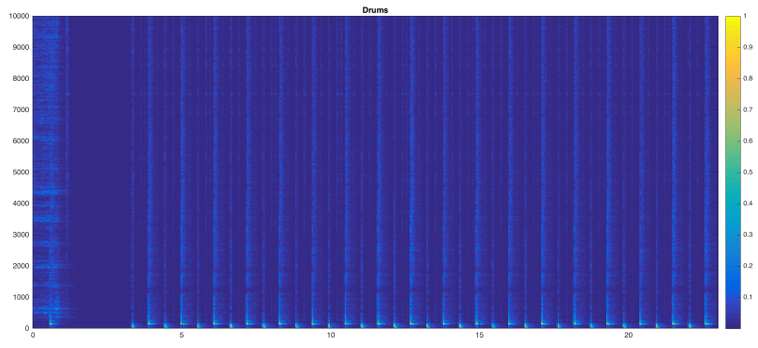


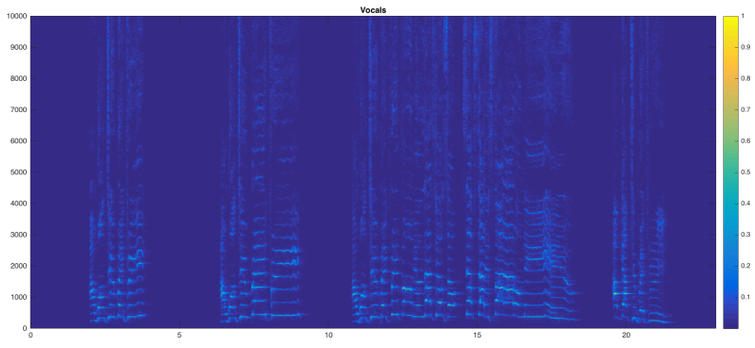


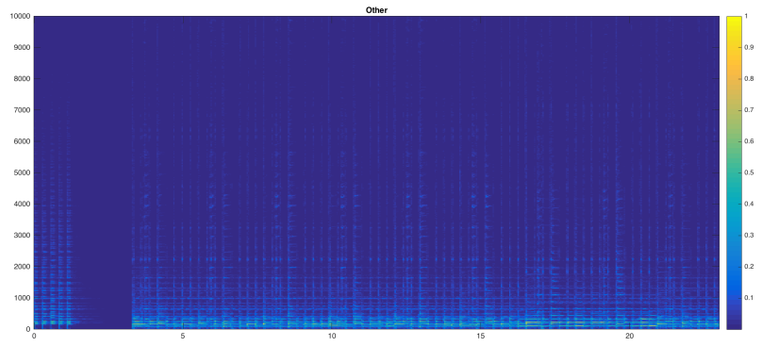


. . . . WHAT WE ARE AIMING FOR

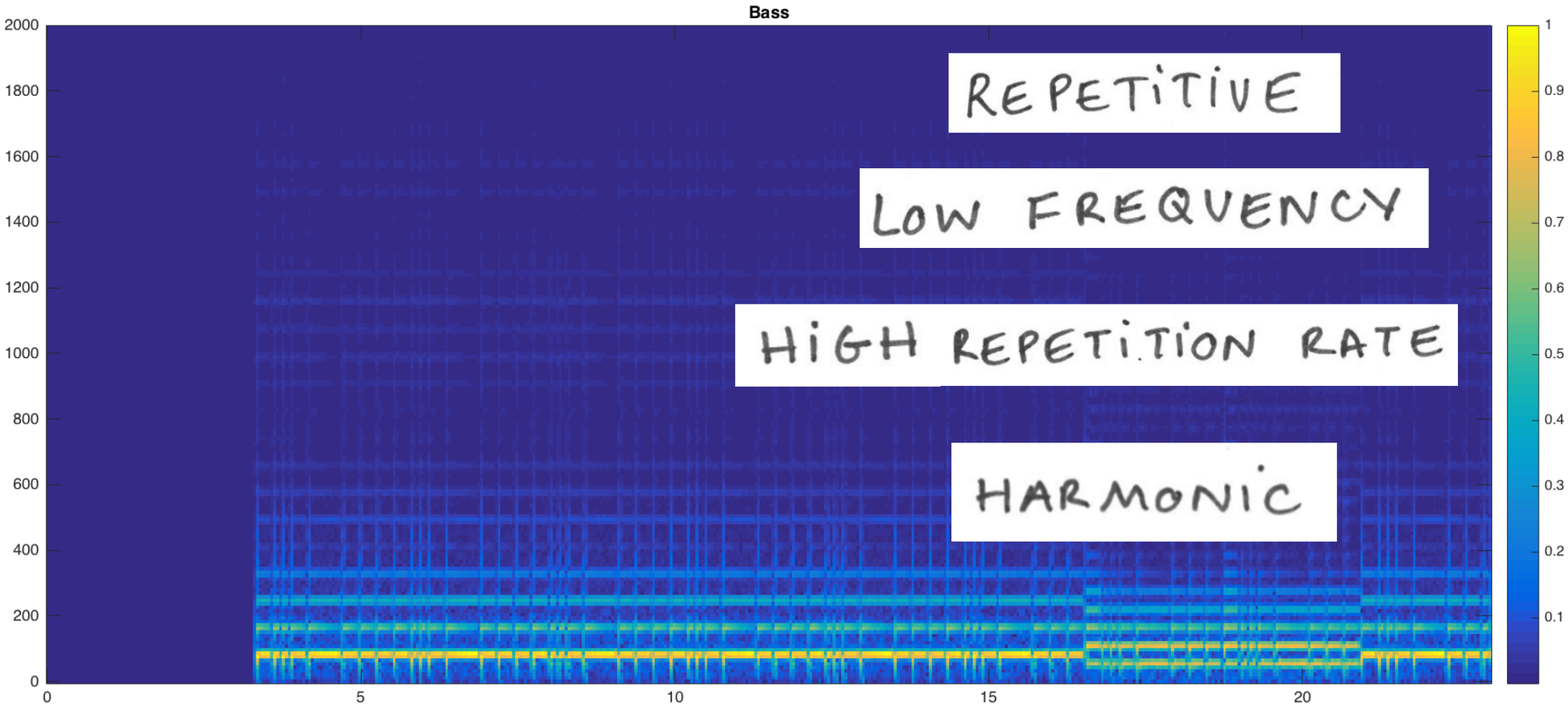


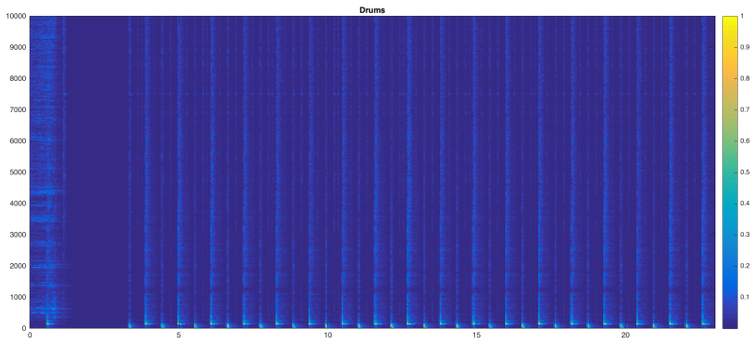


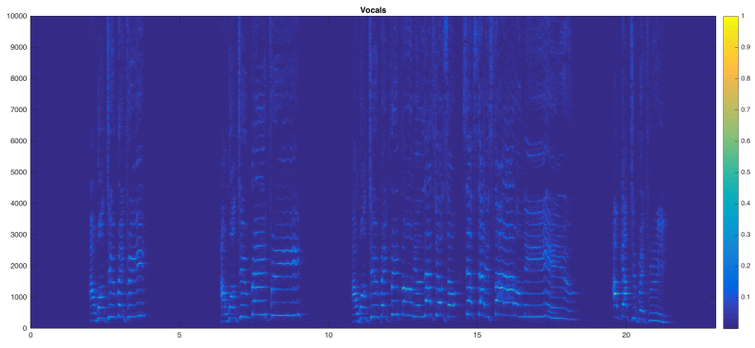


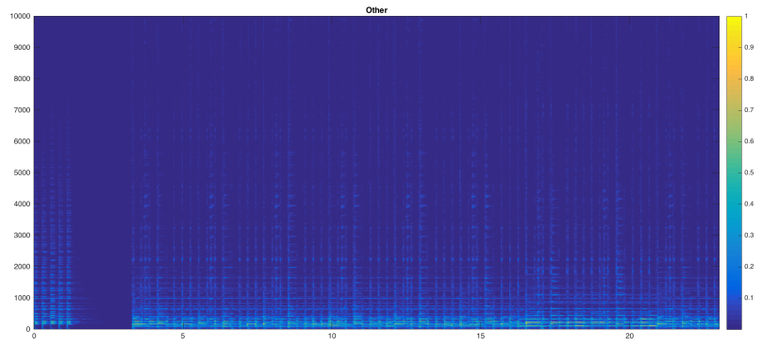


. . . . WHAT WE ARE AIMING FOR

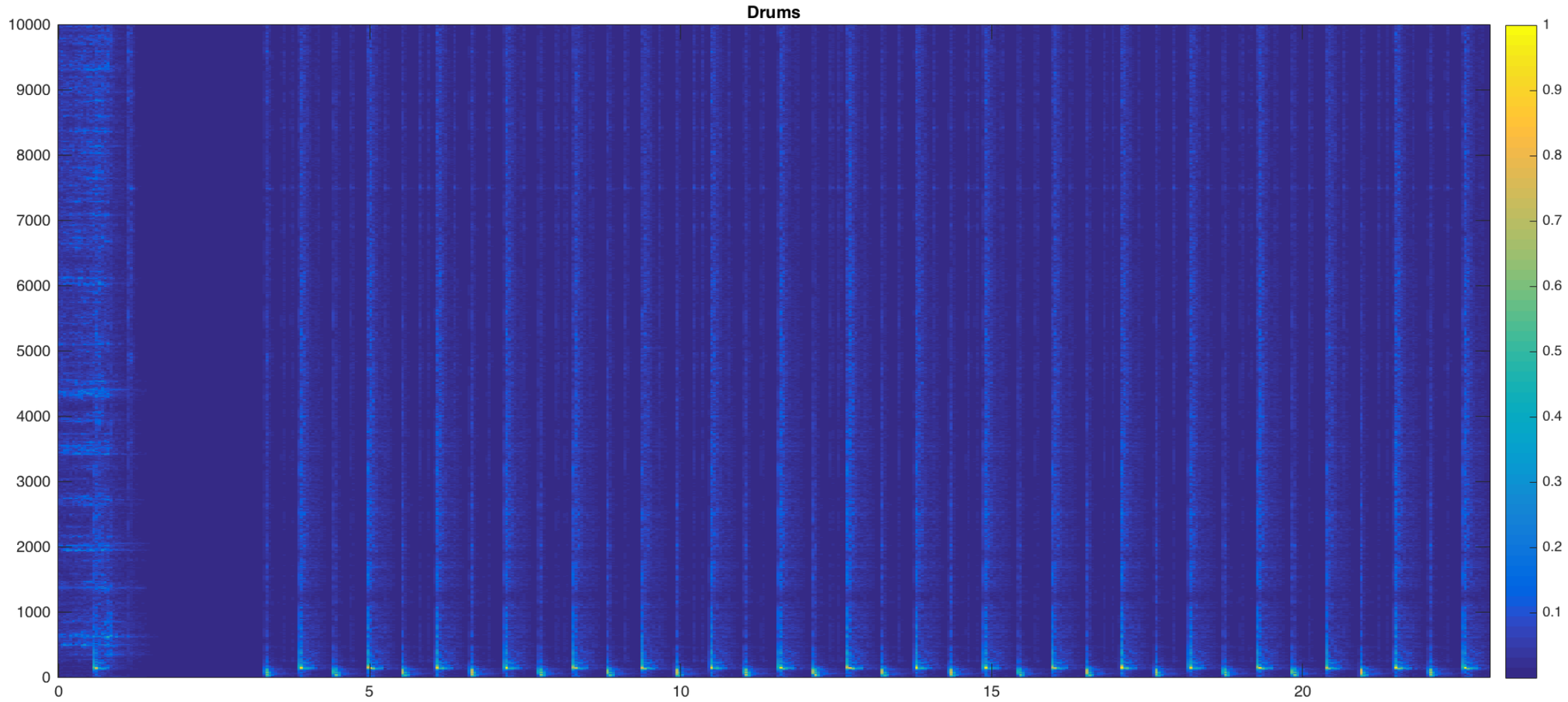


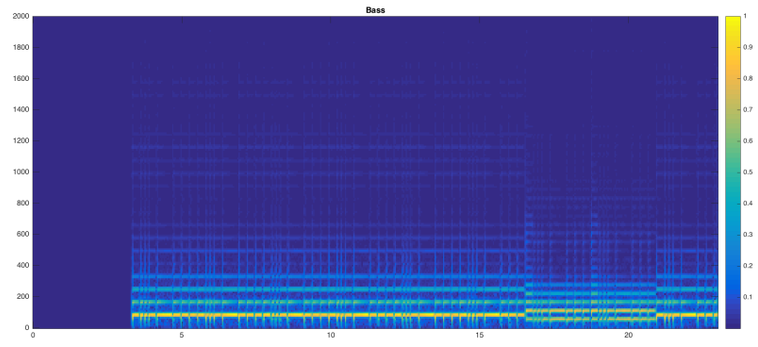


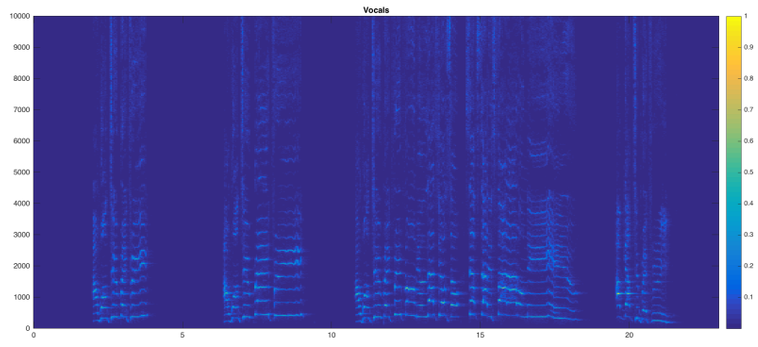


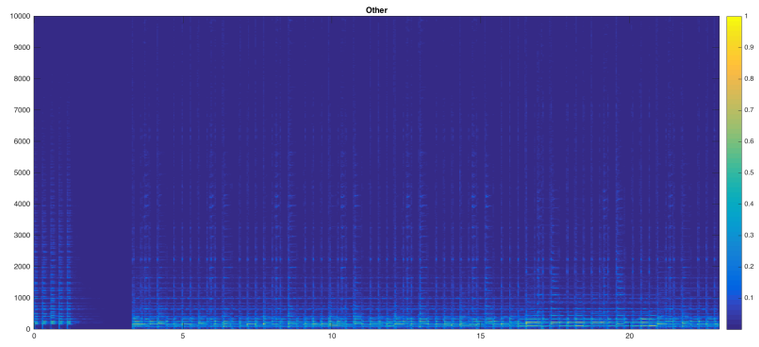


. . . . WHAT WE ARE AIMING FOR









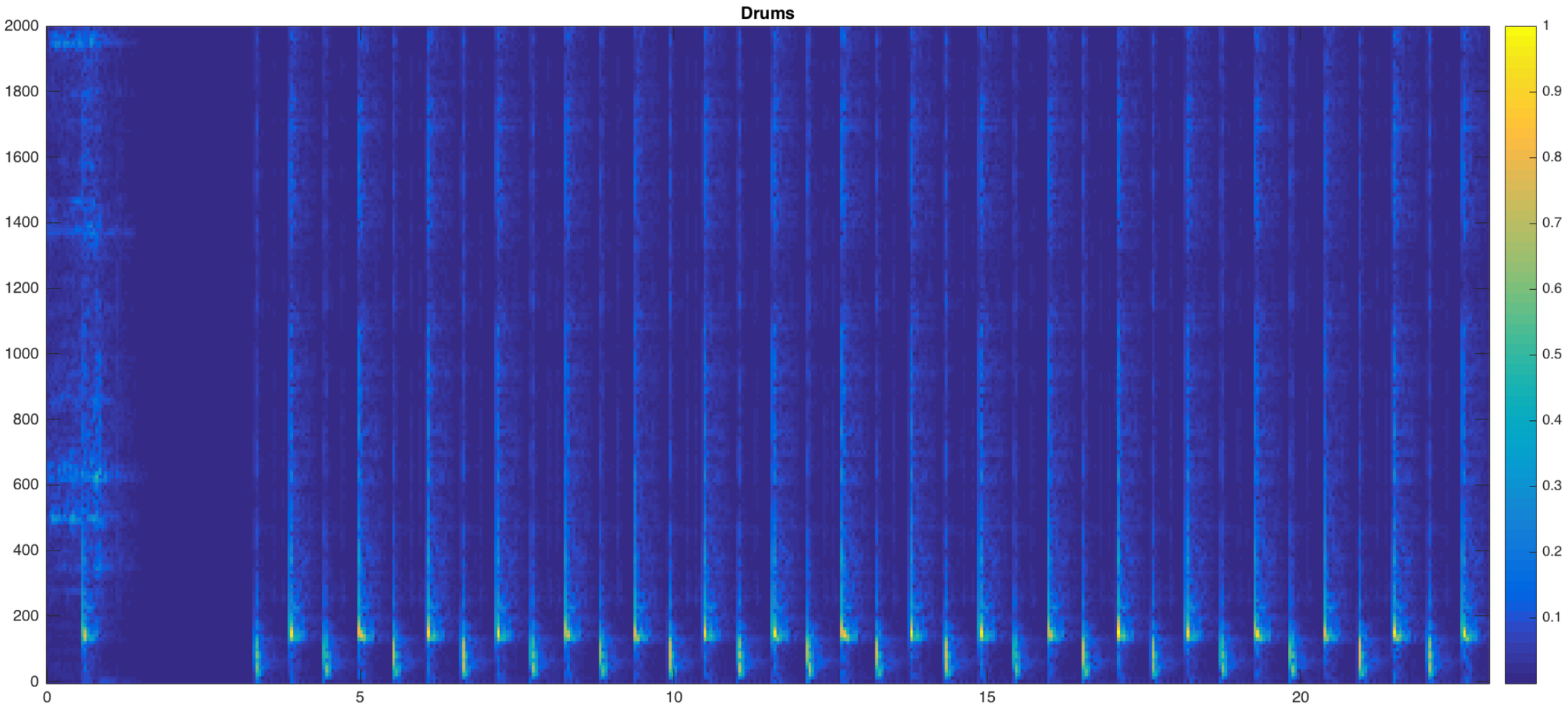
REPETITIVE

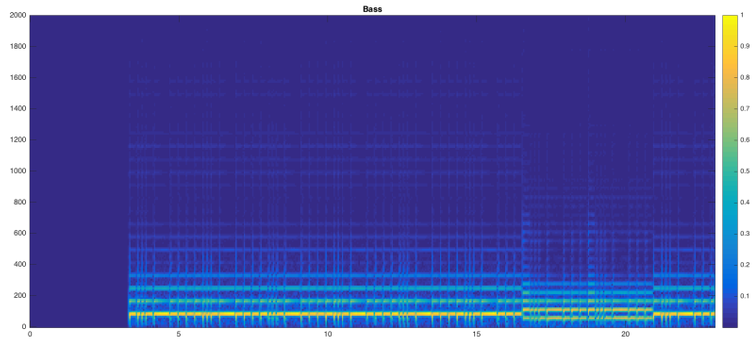
LOW FREQUENCY

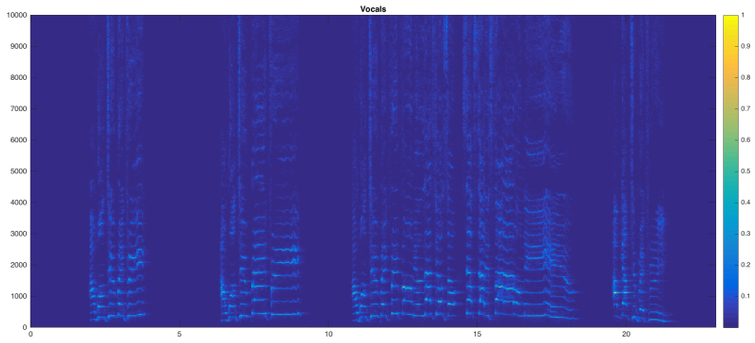
REPETITION RATE

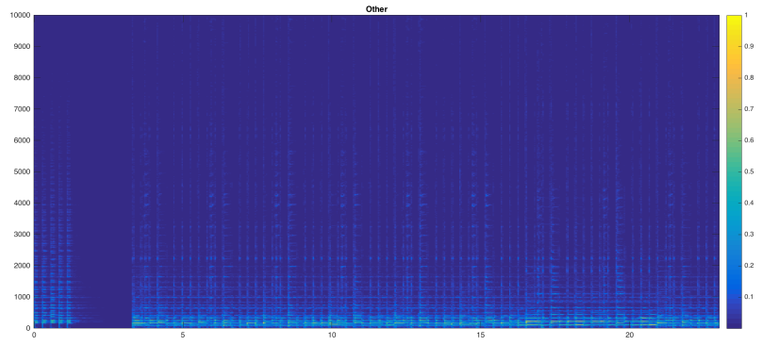
HARMONIC

. . . . WHAT WE ARE AIMING FOR









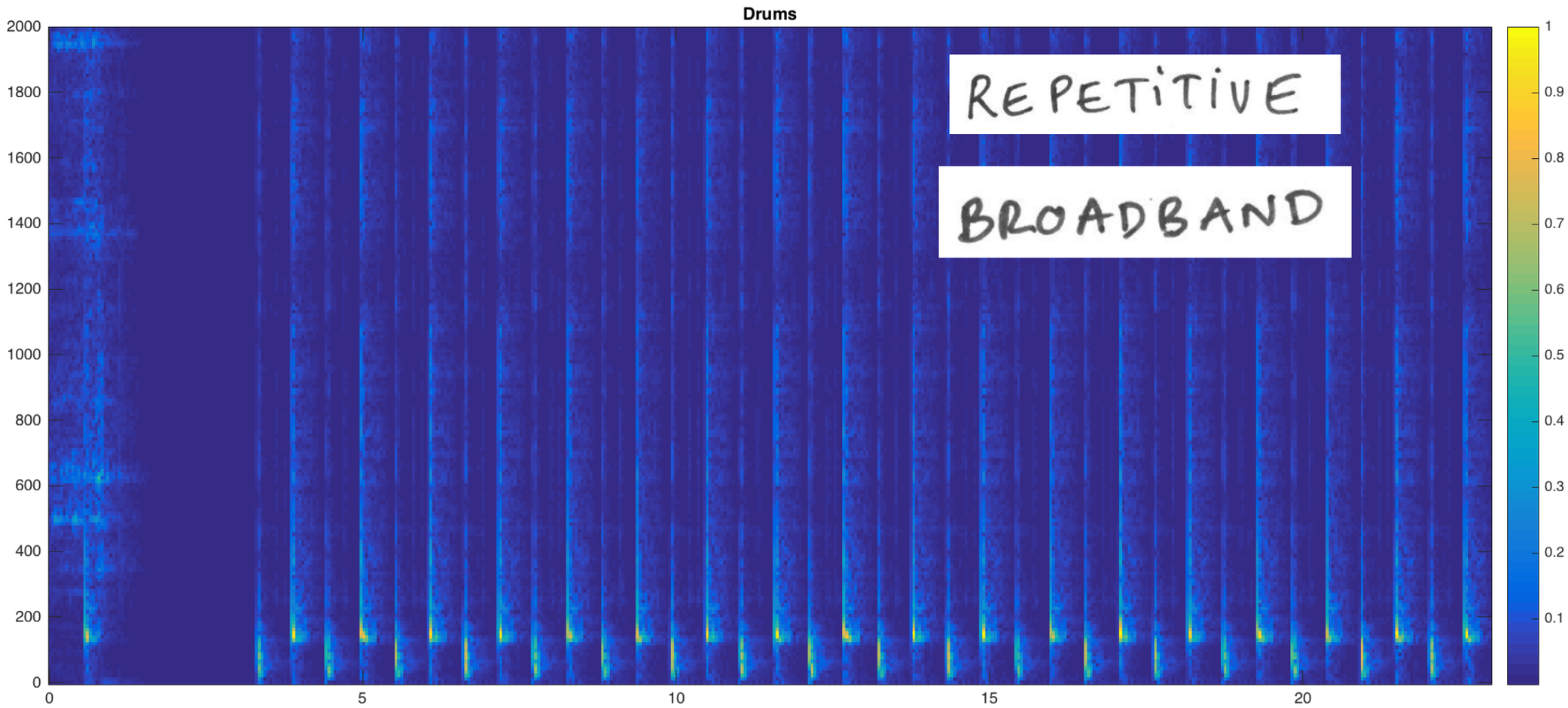
REPETITIVE

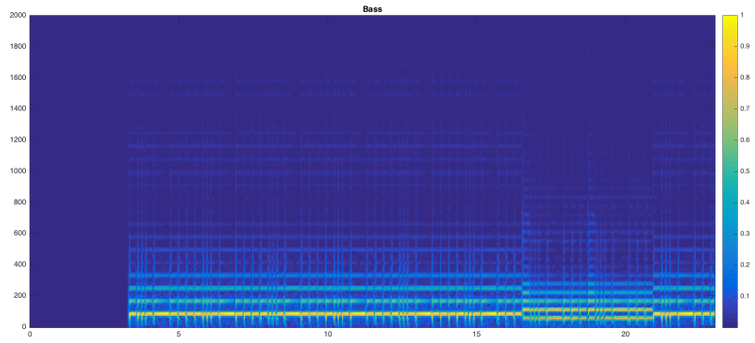
LOW FREQUENCY

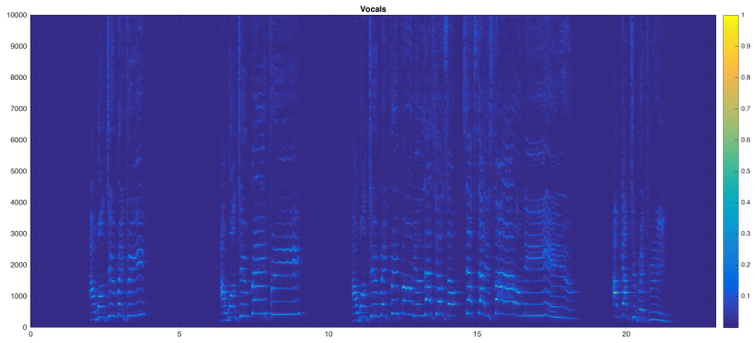
REPETITION RATE

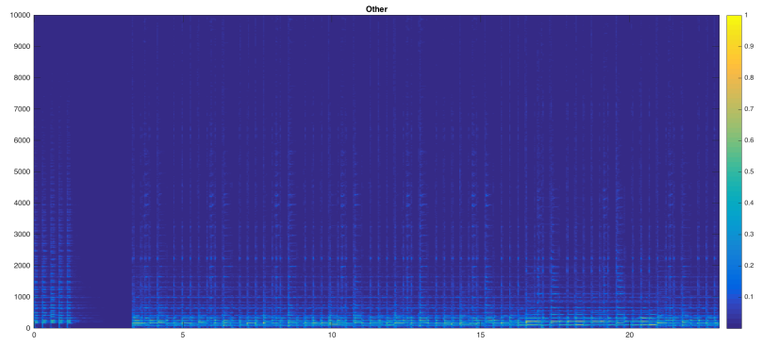
HARMONIC

. . . . WHAT WE ARE AIMING FOR









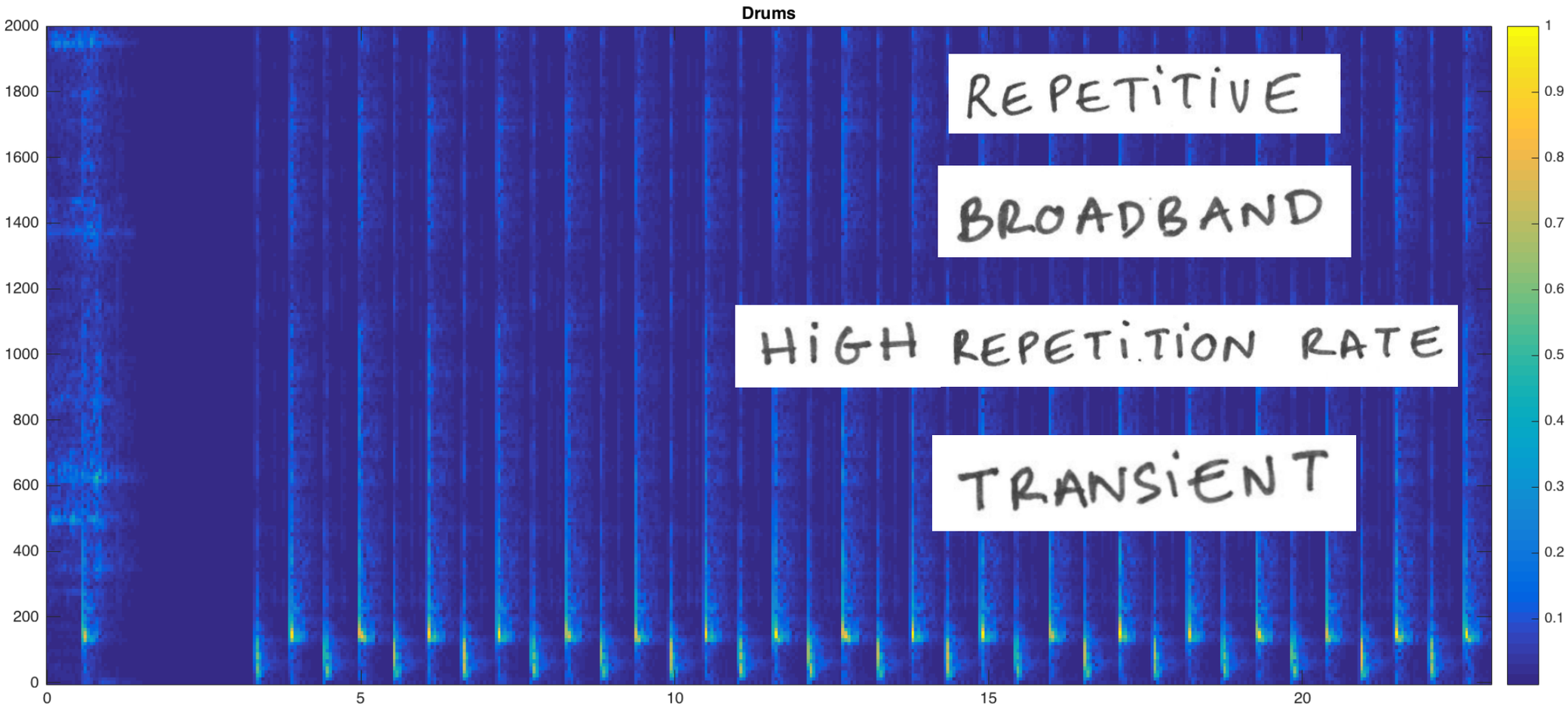
REPETITIVE

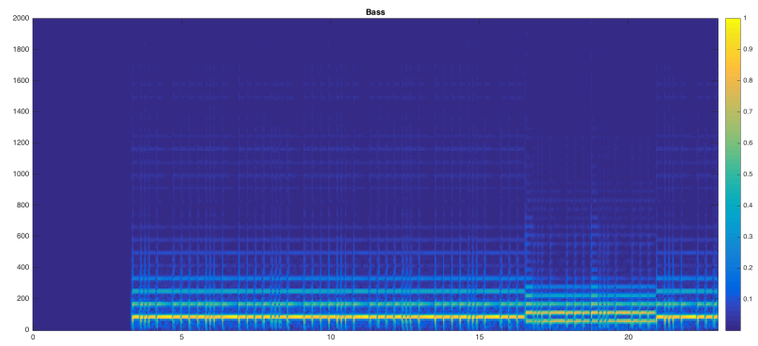
LOW FREQUENCY

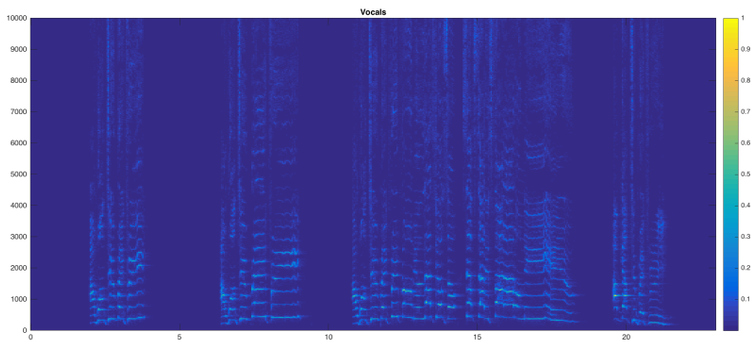
REPETITION RATE

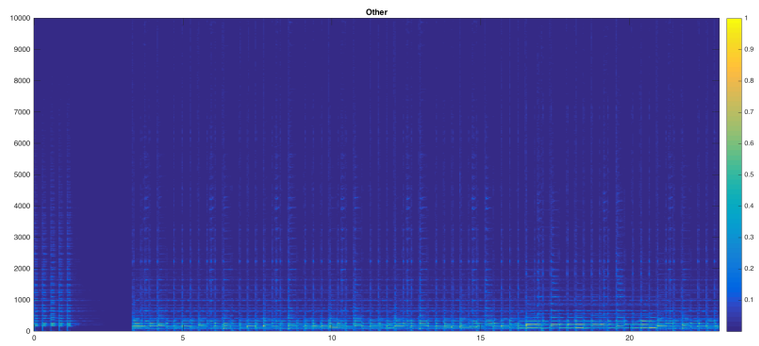
HARMONIC

. . . . WHAT WE ARE AIMING FOR









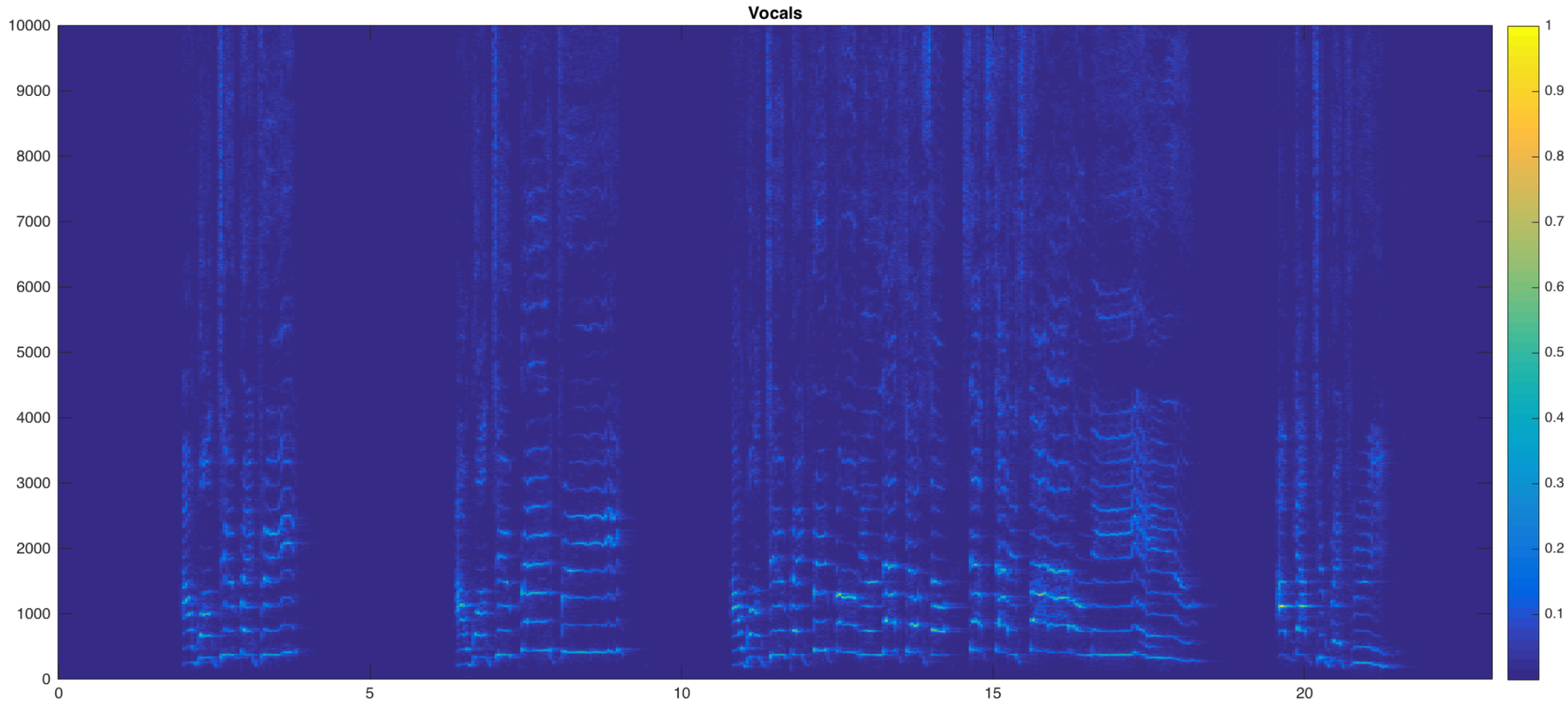
REPETITIVE

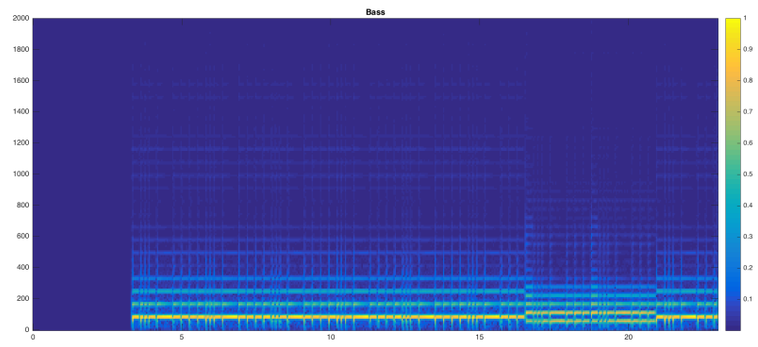
LOW FREQUENCY

REPETITION RATE

HARMONIC

. . . . WHAT WE ARE AIMING FOR



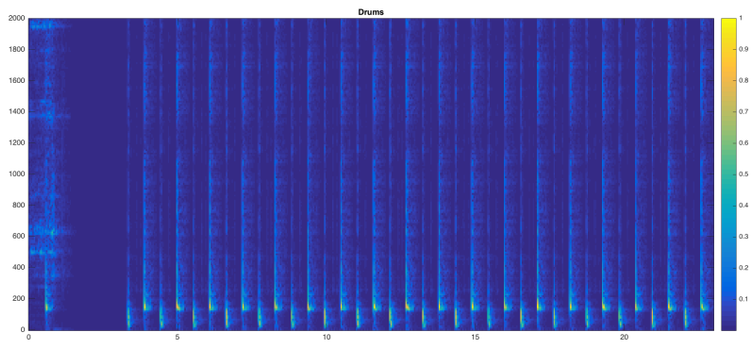


REPETITIVE

LOW FREQUENCY

REPETITION RATE

HARMONIC

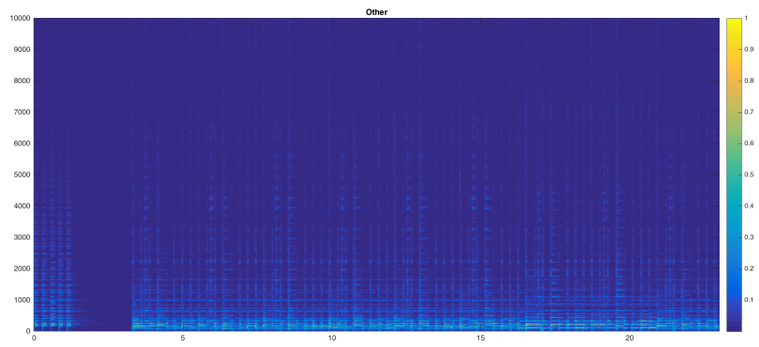


REPETITIVE

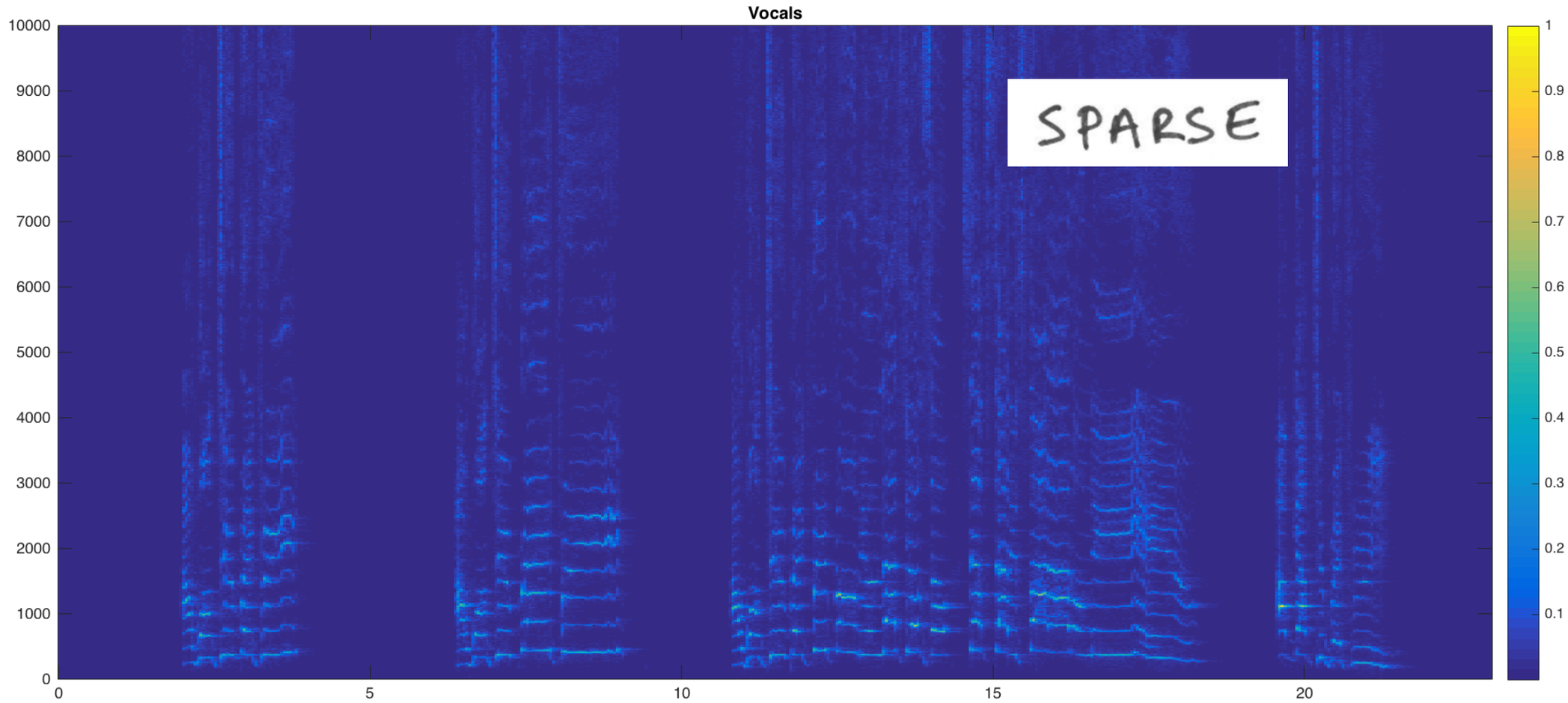
BROADBAND

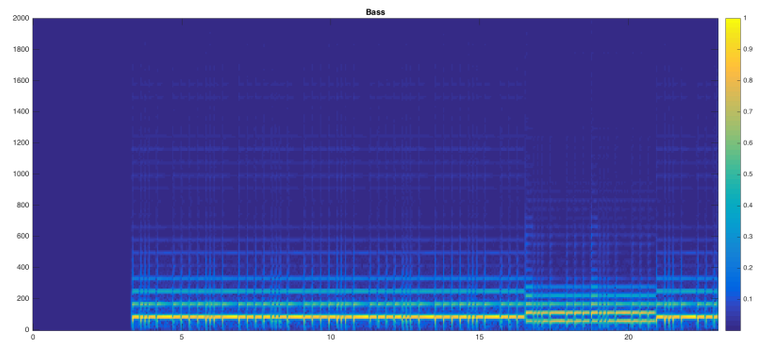
REPETITION RATE

TRANSIENT



. . . . WHAT WE ARE AIMING FOR



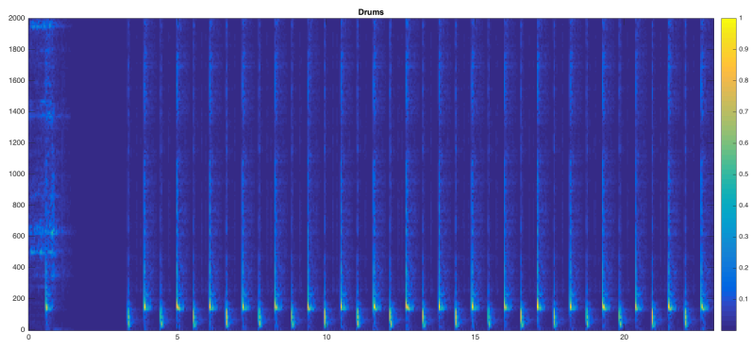


REPETITIVE

LOW FREQUENCY

REPETITION RATE

HARMONIC

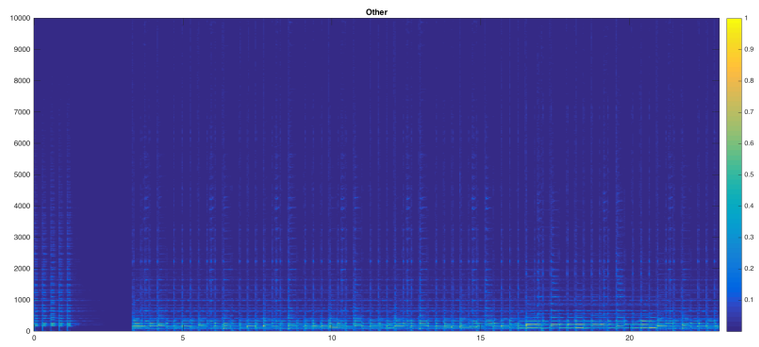


REPETITIVE

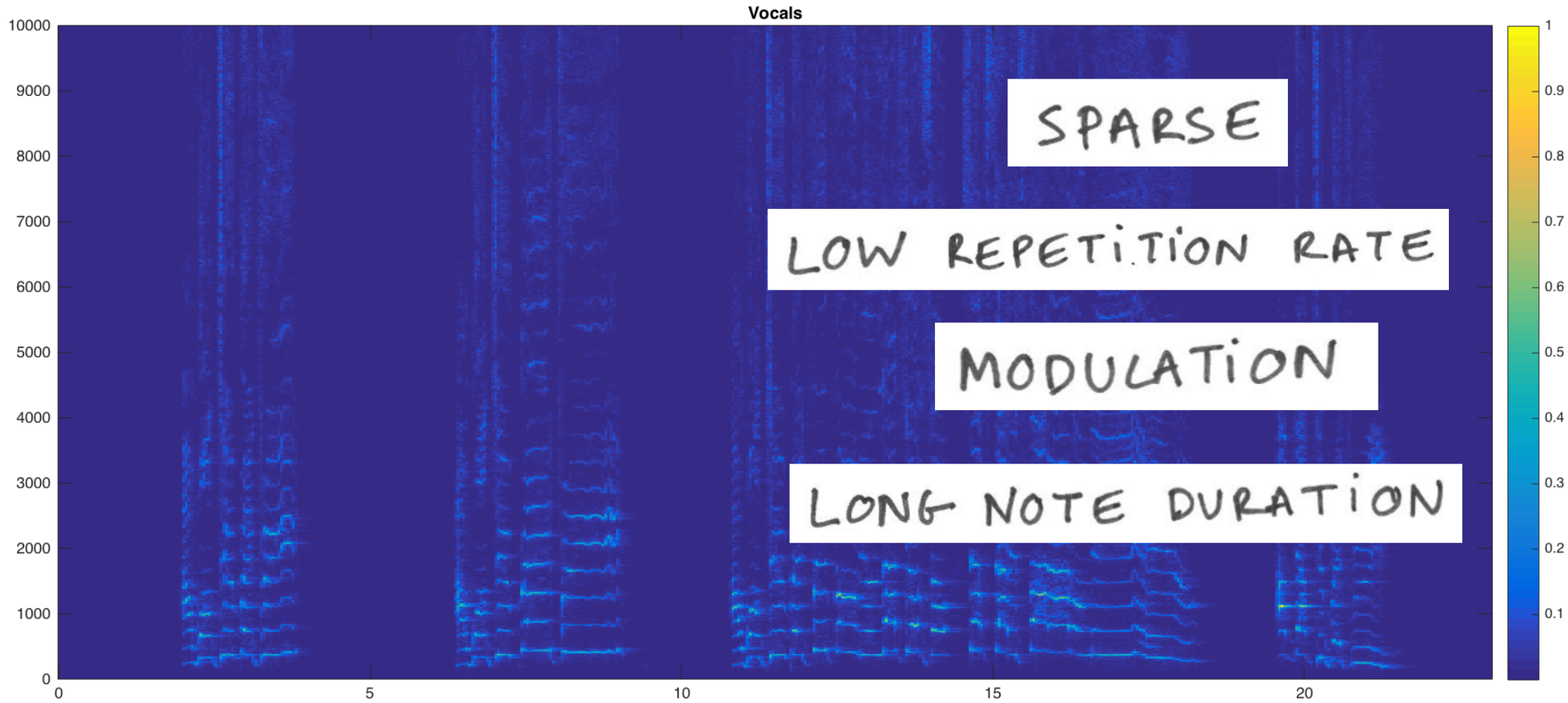
BROADBAND

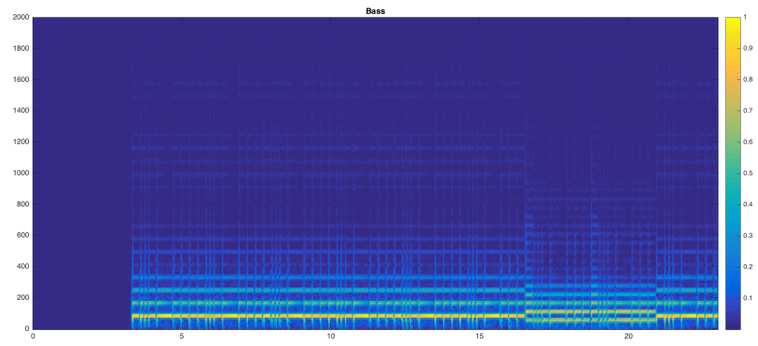
REPETITION RATE

TRANSIENT



. . . . WHAT WE ARE AIMING FOR



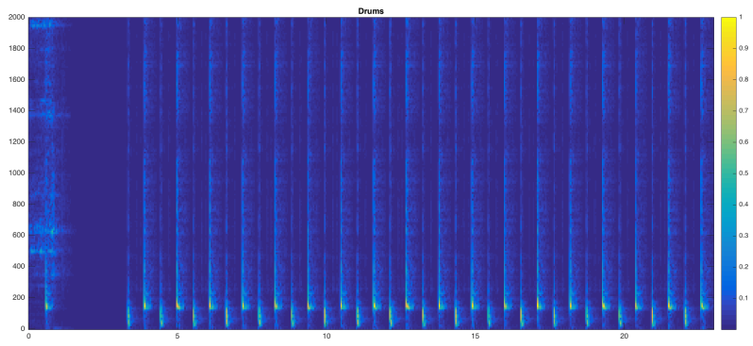


REPETITIVE

LOW FREQUENCY

REPETITION RATE

HARMONIC

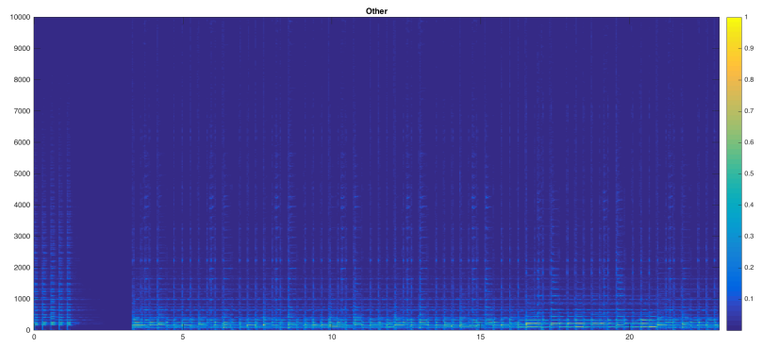


REPETITIVE

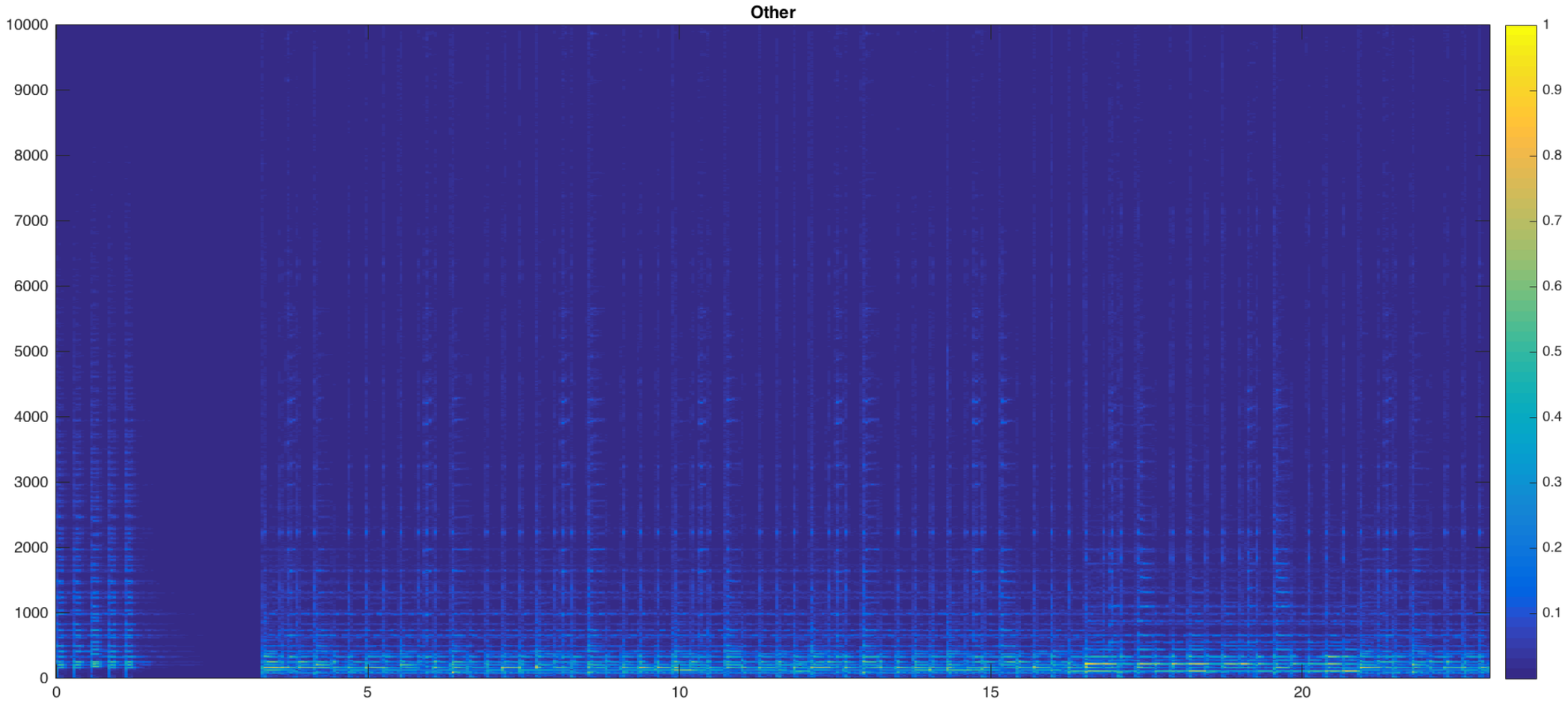
BROADBAND

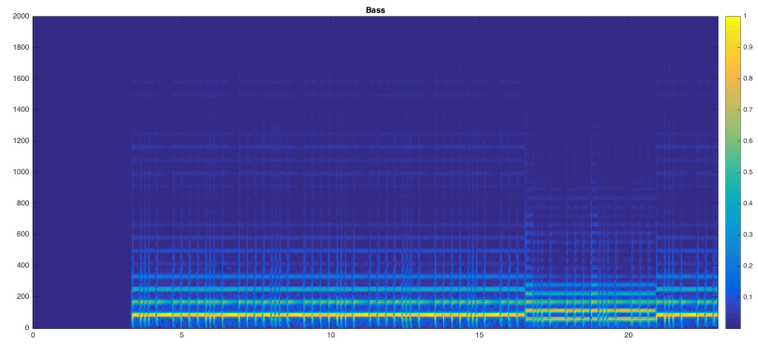
REPETITION RATE

TRANSIENT



. . . . WHAT WE ARE AIMING FOR



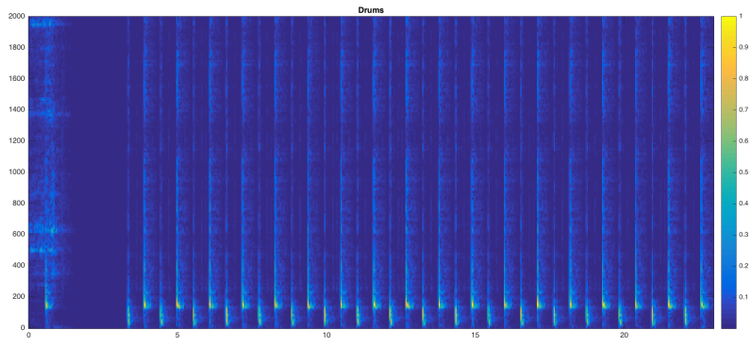


REPETITIVE

LOW FREQUENCY

REPETITION RATE

HARMONIC

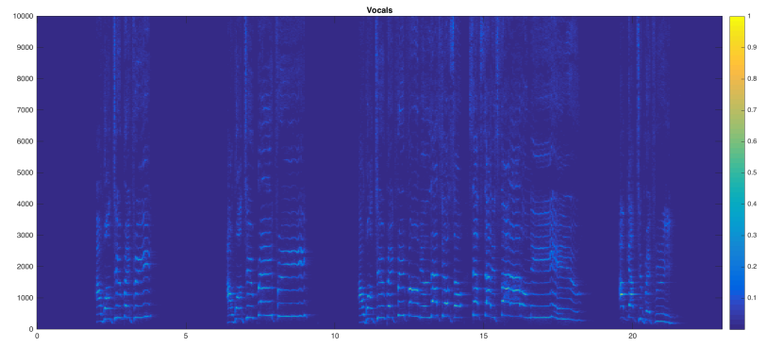


REPETITIVE

BROADBAND

REPETITION RATE

TRANSIENT



REPETITIVE

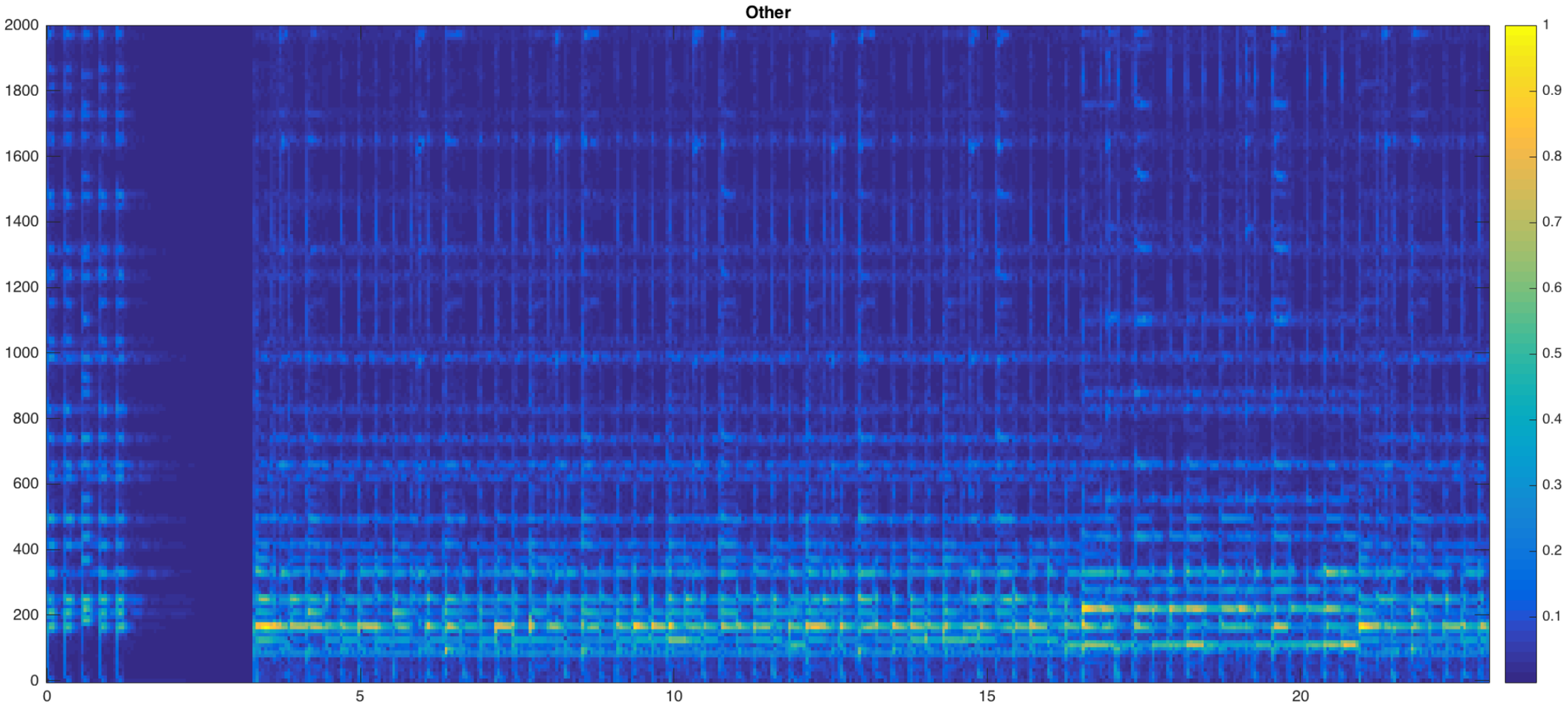
SPARSE

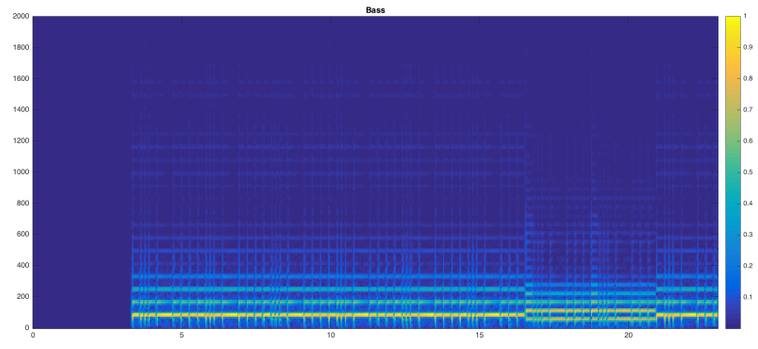
LOW REPETITION RATE

MODULATION

LONG NOTE DURATION

. . . . WHAT WE ARE AIMING FOR



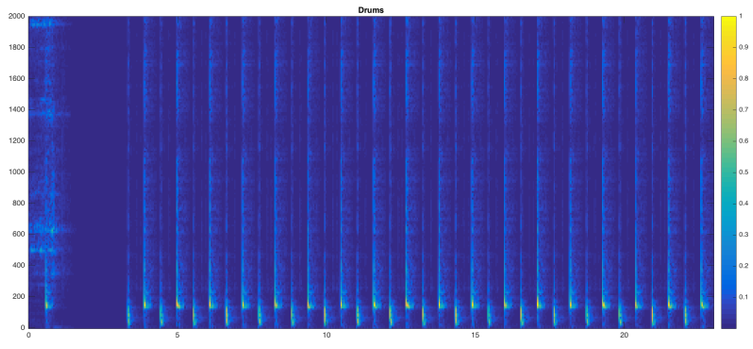


REPETITIVE

LOW FREQUENCY

REPETITION RATE

HARMONIC

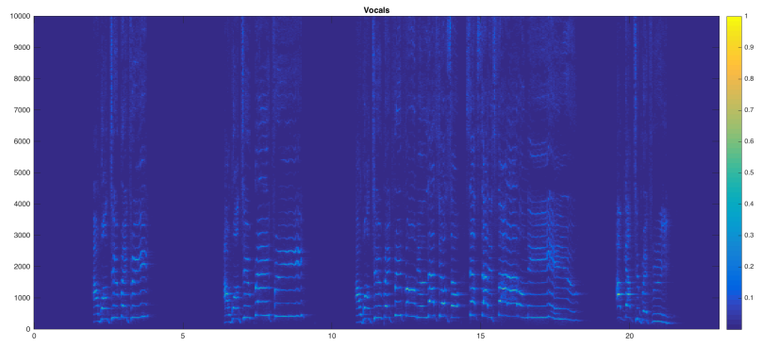


REPETITIVE

BROADBAND

REPETITION RATE

TRANSIENT



REPETITIVE

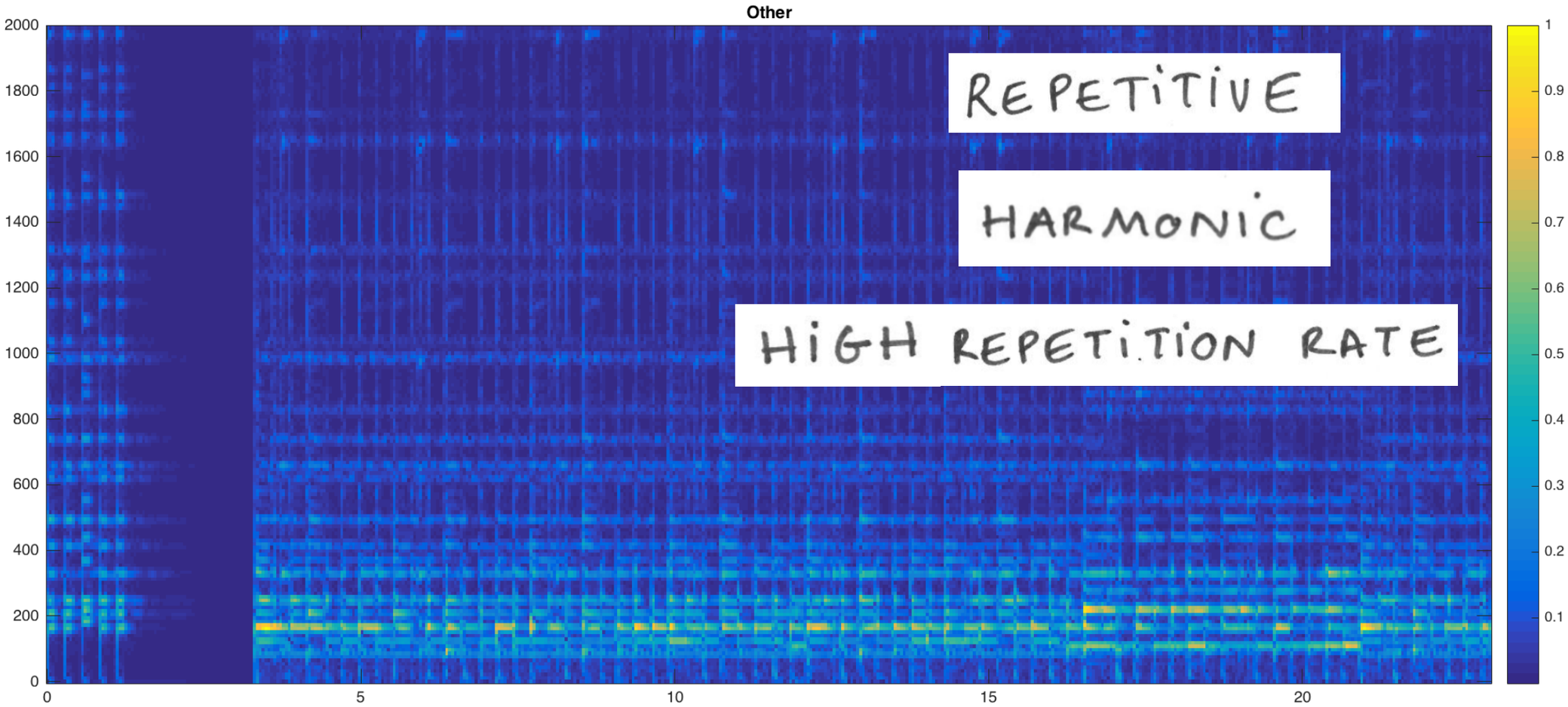
SPARSE

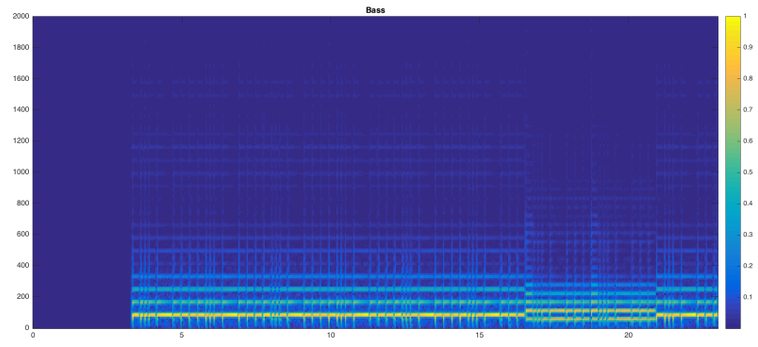
LOW REPETITION RATE

MODULATION

LONG NOTE DURATION

. . . . WHAT WE ARE AIMING FOR



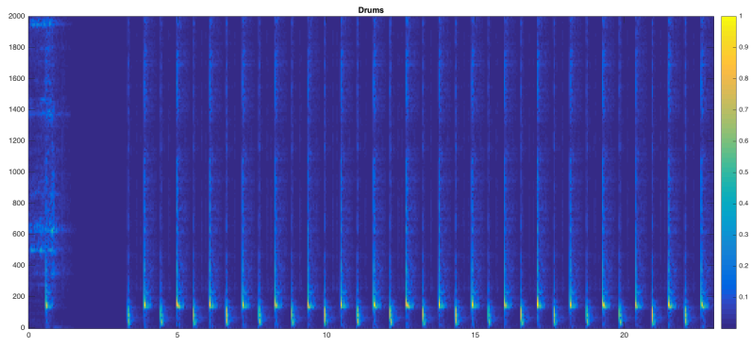


REPETITIVE

LOW FREQUENCY

REPETITION RATE

HARMONIC

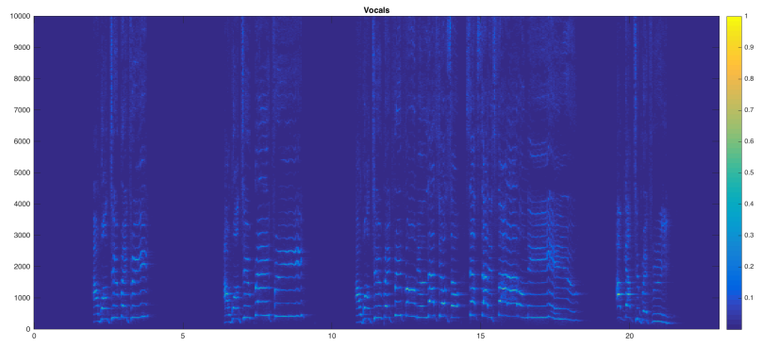


REPETITIVE

BROADBAND

REPETITION RATE

TRANSIENT



REPETITIVE

SPARSE

LOW REPETITION RATE

MODULATION

LONG NOTE DURATION

BASS

REPETITIVE

LOW FREQUENCY

HIGH REPETITION RATE

HARMONIC

DRUMS

REPETITIVE

BROADBAND

HIGH REPETITION RATE

TRANSIENT

VOCALS

SPARSE

LOW REPETITION RATE

MODULATION

LONG NOTE DURATION

OTHER

REPETITIVE

HARMONIC

HIGH REPETITION RATE

BASS

REPETITIVE

LOW FREQUENCY

HIGH REPETITION RATE

HARMONIC

DRUMS

REPETITIVE

BROADBAND

HIGH REPETITION RATE

TRANSIENT

VOCALS

SPARSE

LOW REPETITION RATE

MODULATION

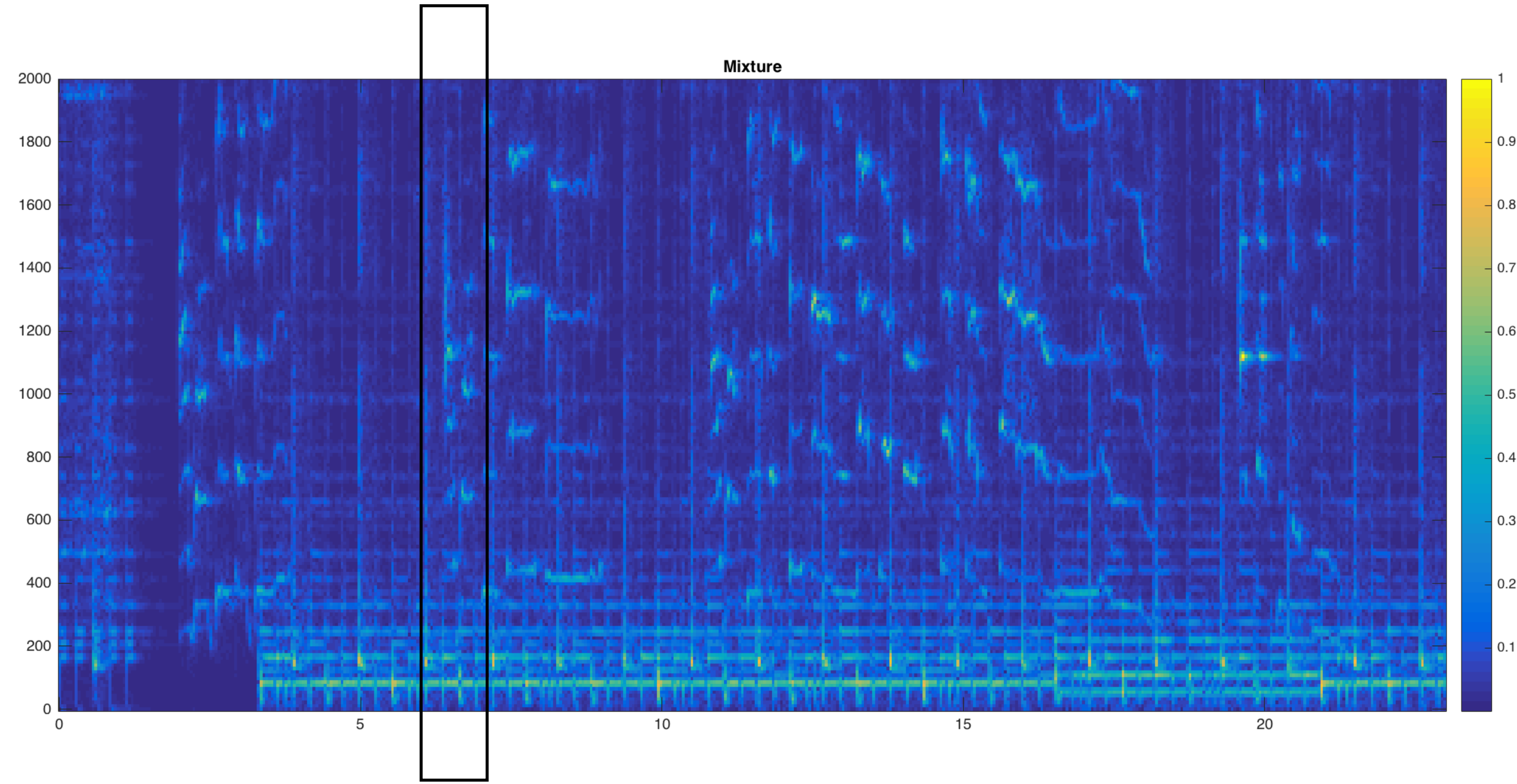
LONG NOTE DURATION

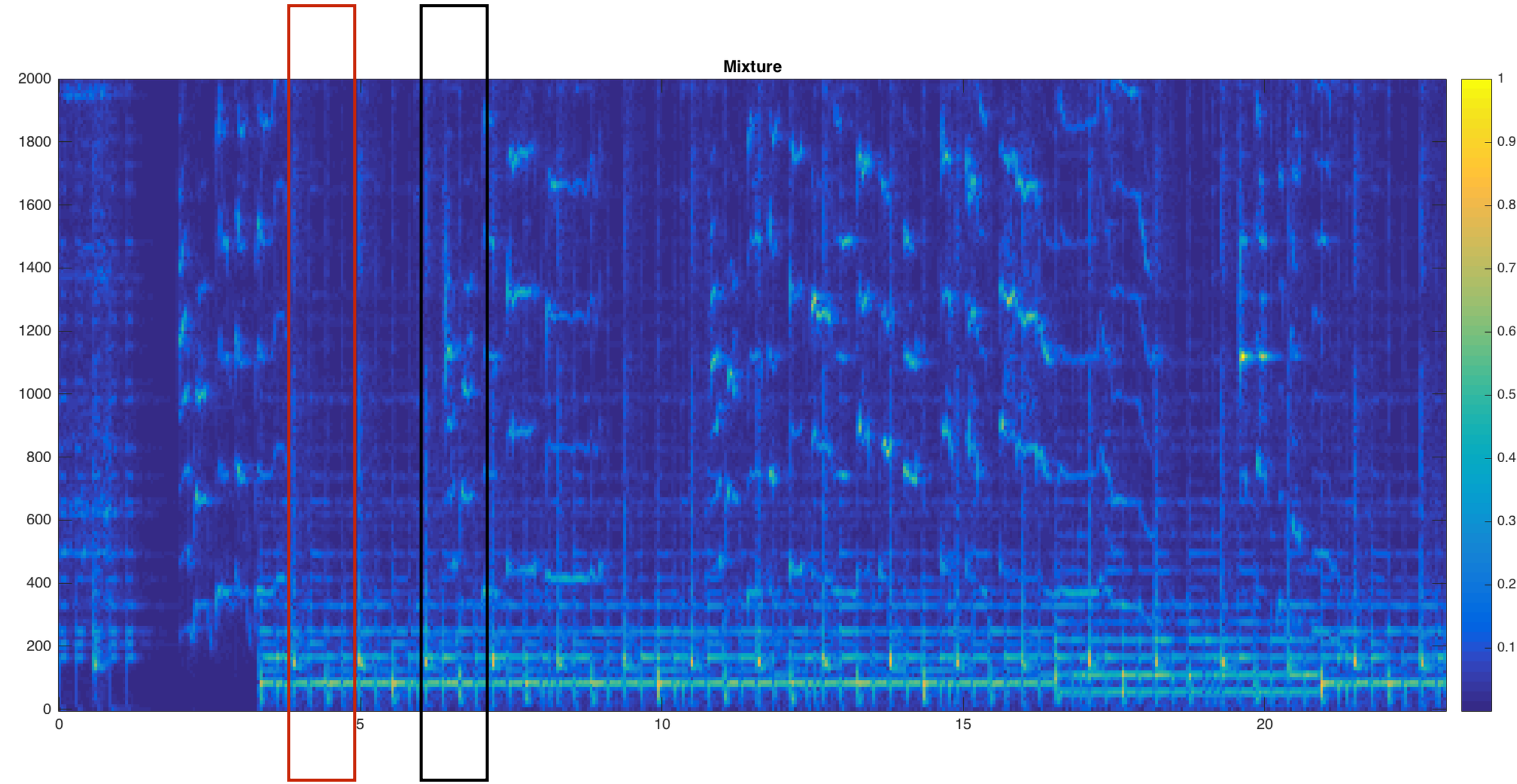
OTHER

REPETITIVE

HARMONIC

HIGH REPETITION RATE

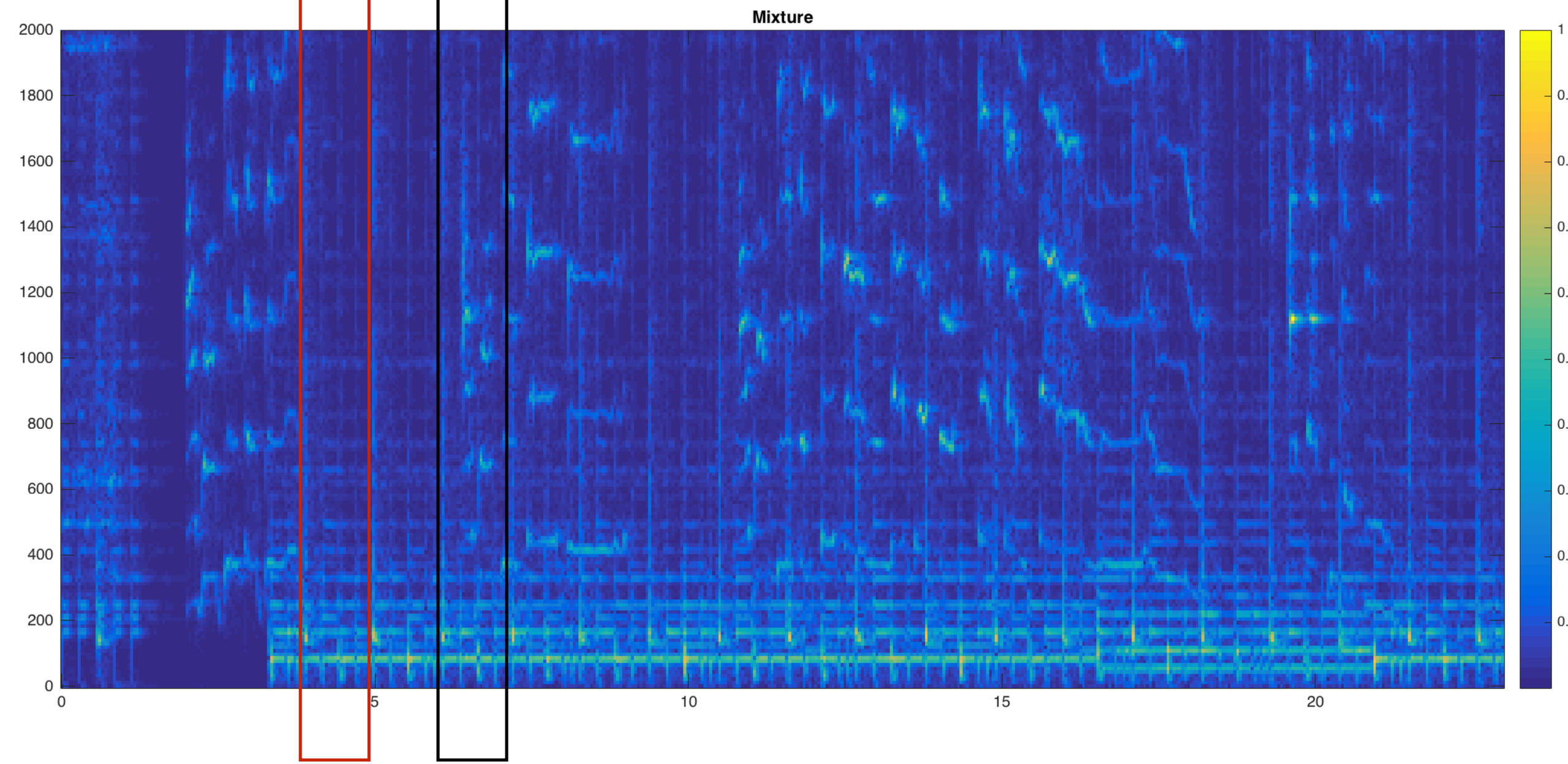

Mixture
2000
1800
1600
1400
1200
1000
800
600
400
200
0
0
5
10
15
20
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1


Mixture
2000
1800
1600
1400
1200
1000
800
600
400
200
0
0
5
10
15
20
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1

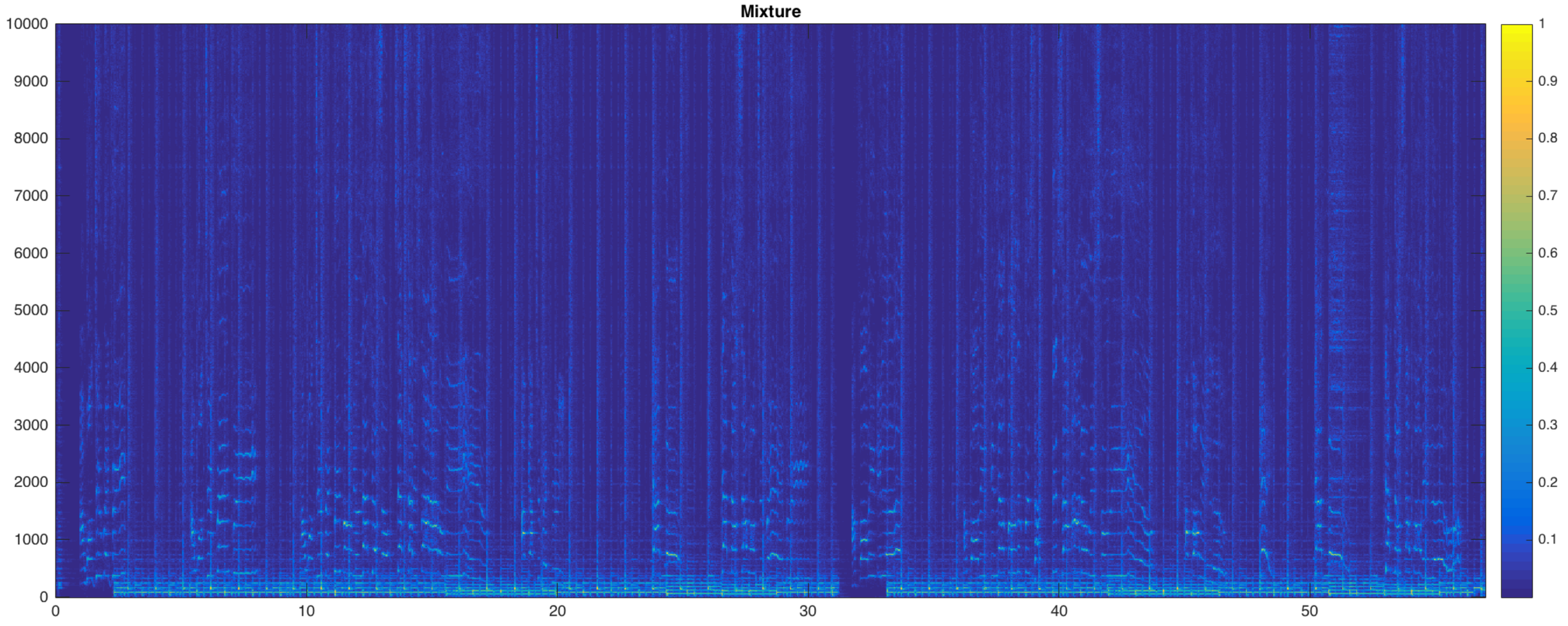
KERNEL ADDITIVE MODELLING

K.A.M.

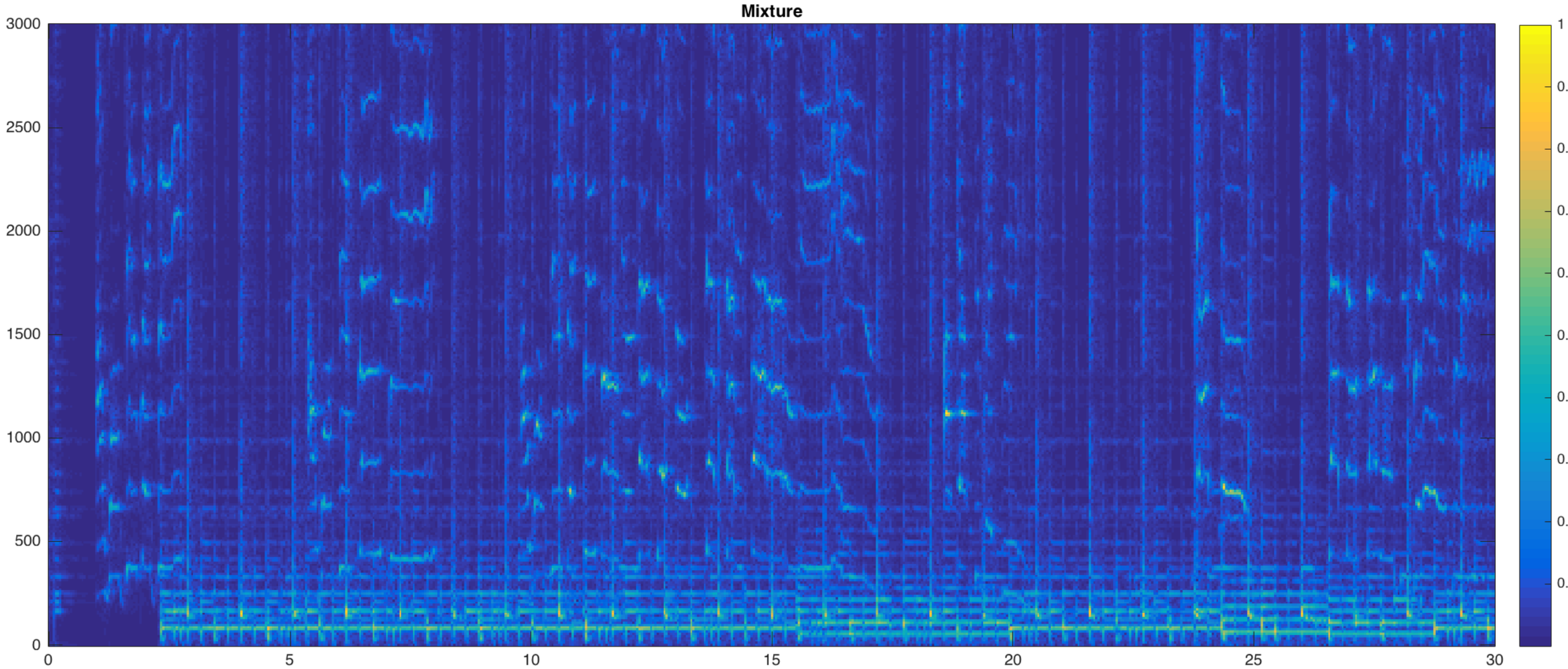
SIMILARITY KERNEL



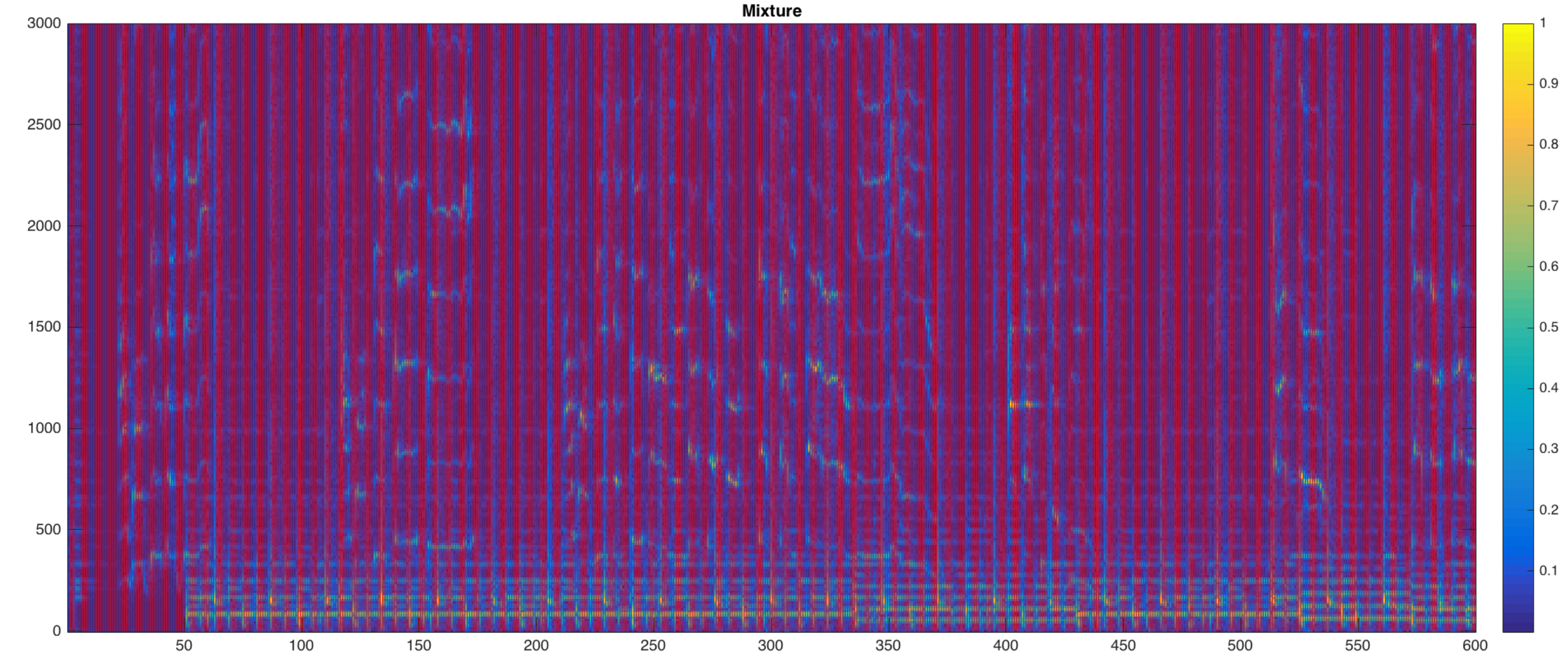
VOCAL SEPARATION WITH K.A.M



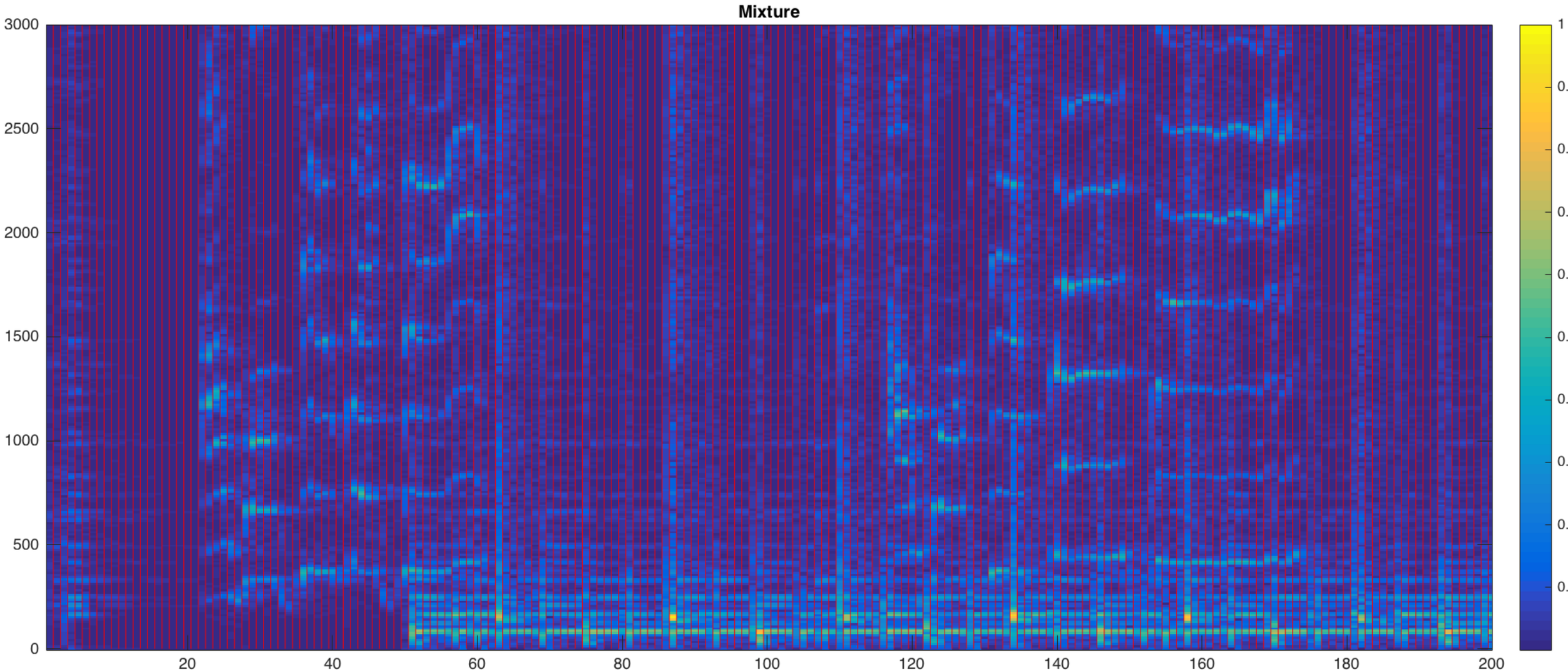
VOCAL SEPARATION WITH K.A.M



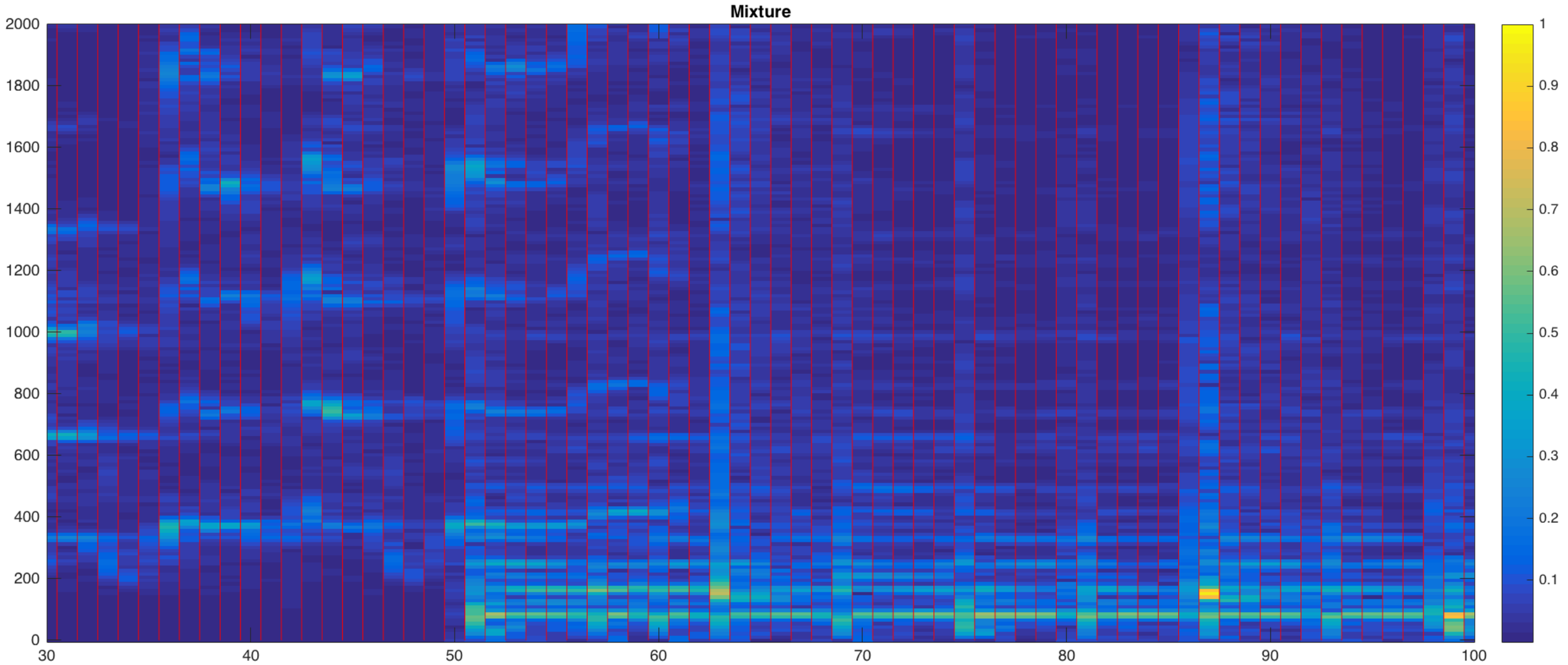
1. DIVIDE IN TIME FRAMES



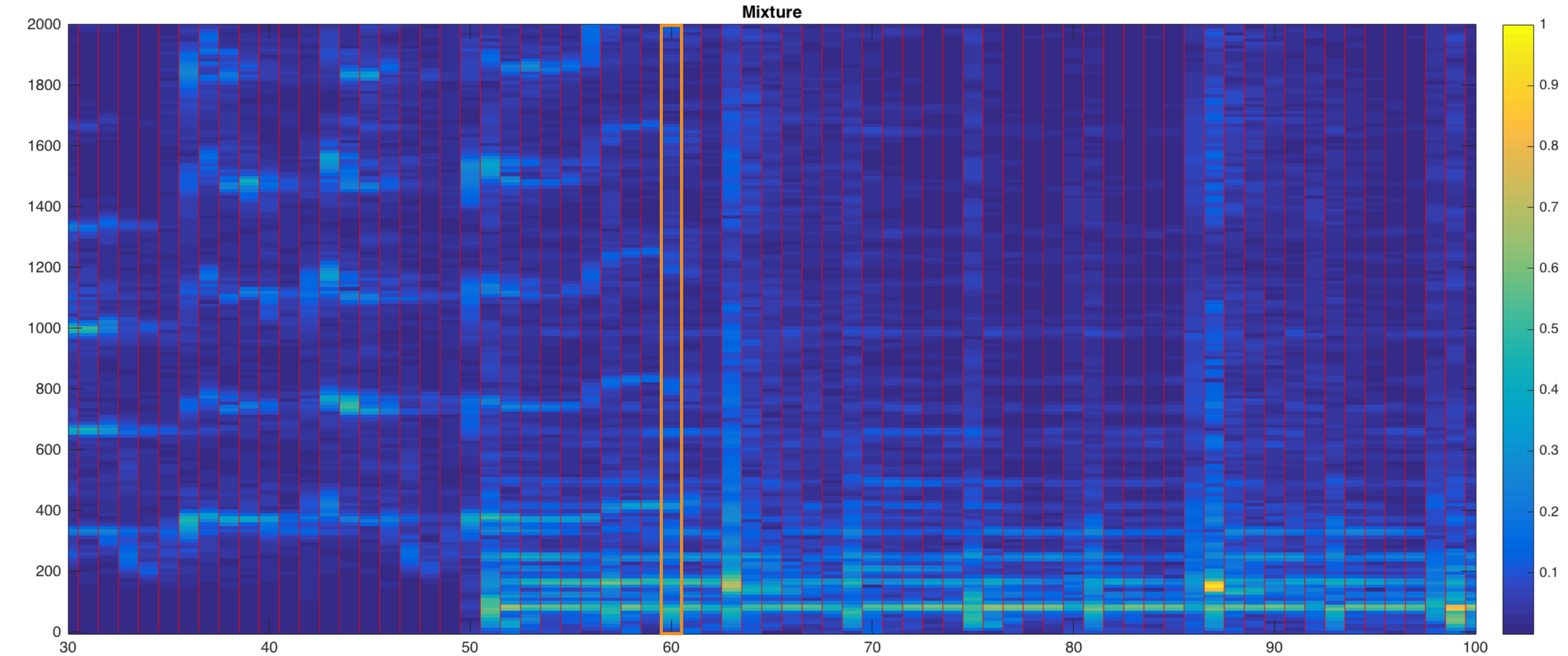
1. DIVIDE IN TIME FRAMES



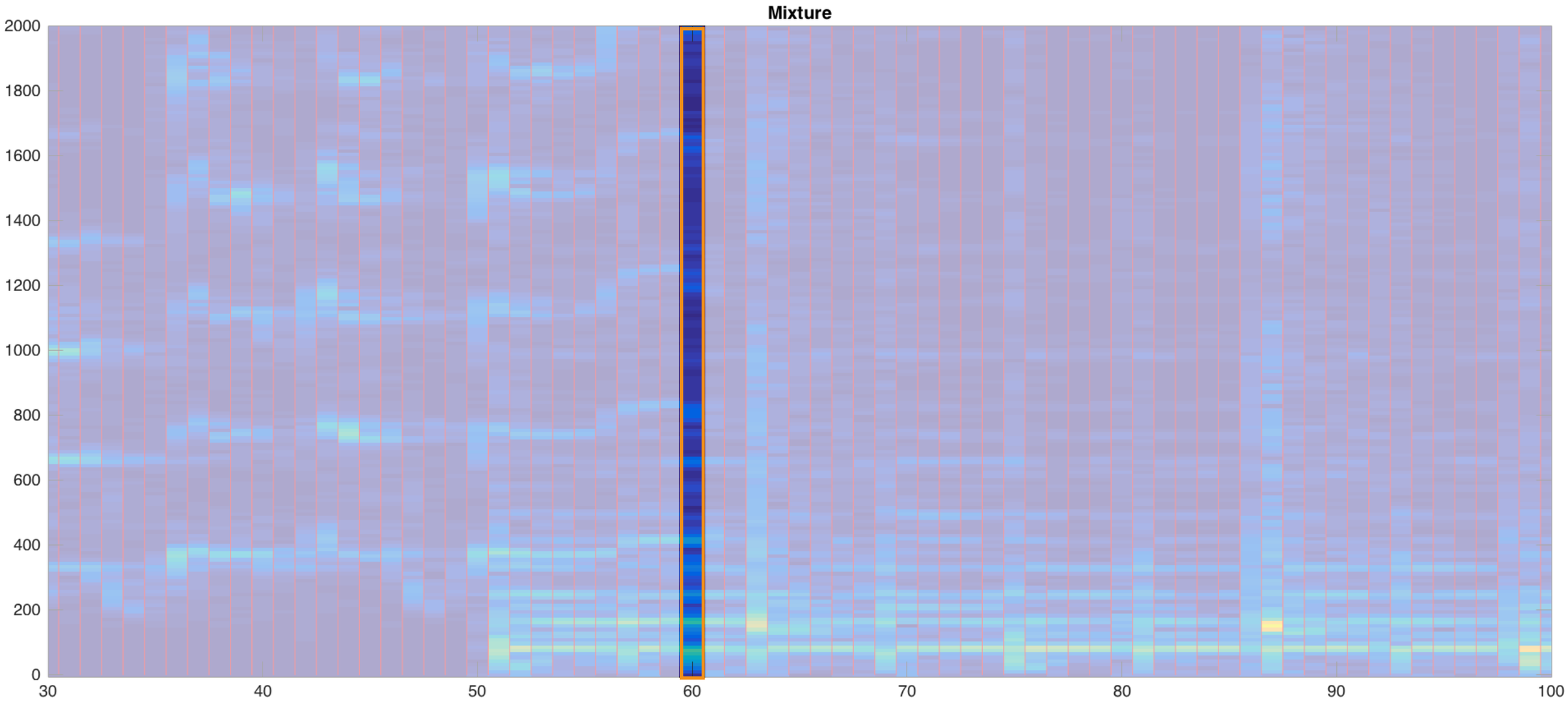
1. DIVIDE IN TIME FRAMES



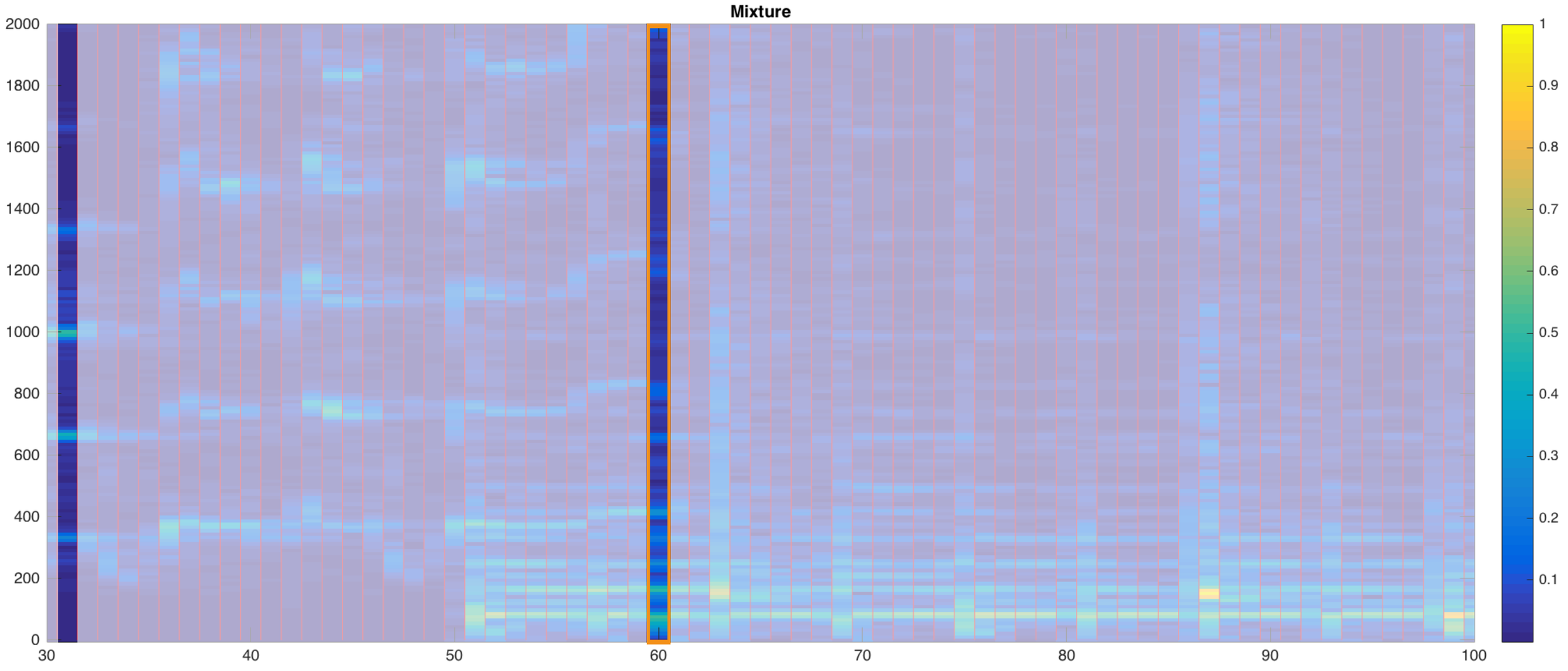
1. DIVIDE IN TIME FRAMES



2. LOOK FOR SIMILAR TIME FRAMES

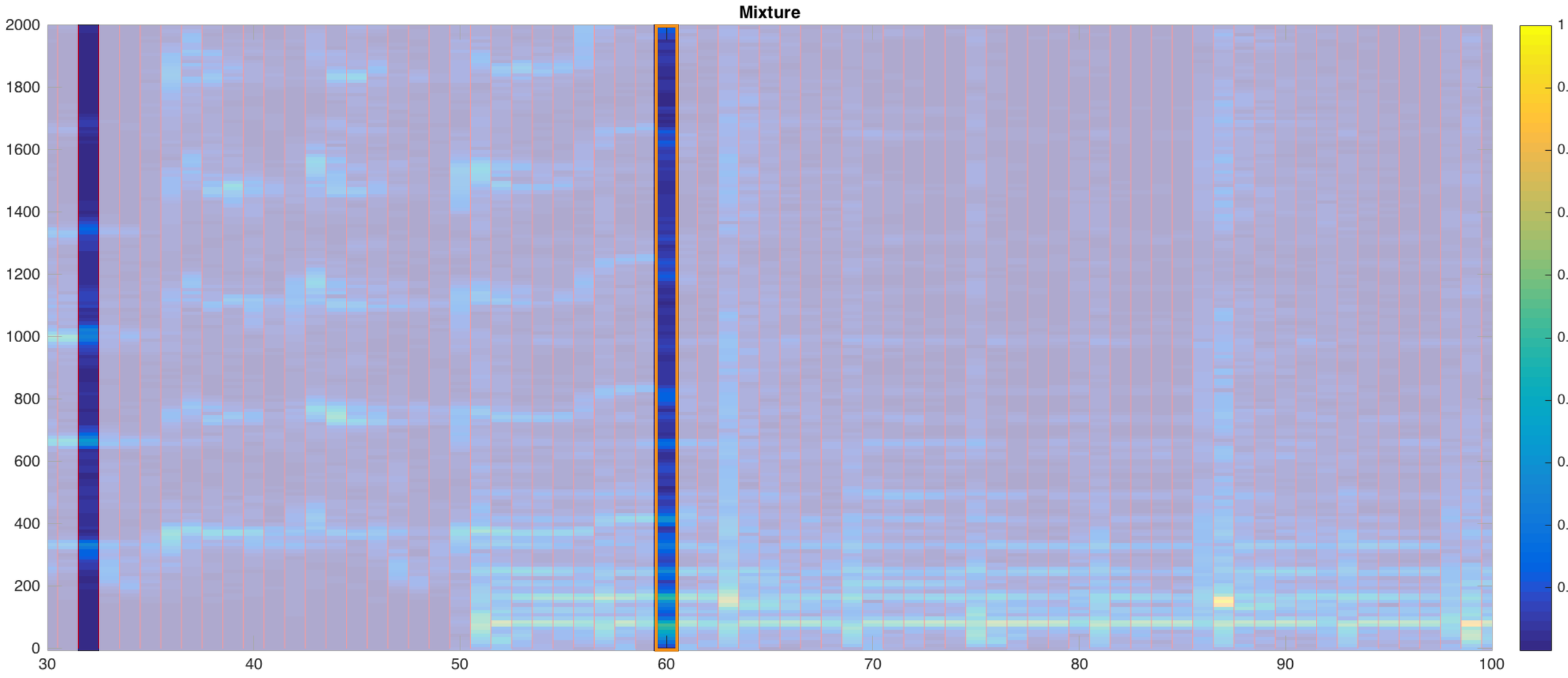


2. LOOK FOR SIMILAR TIME FRAMES



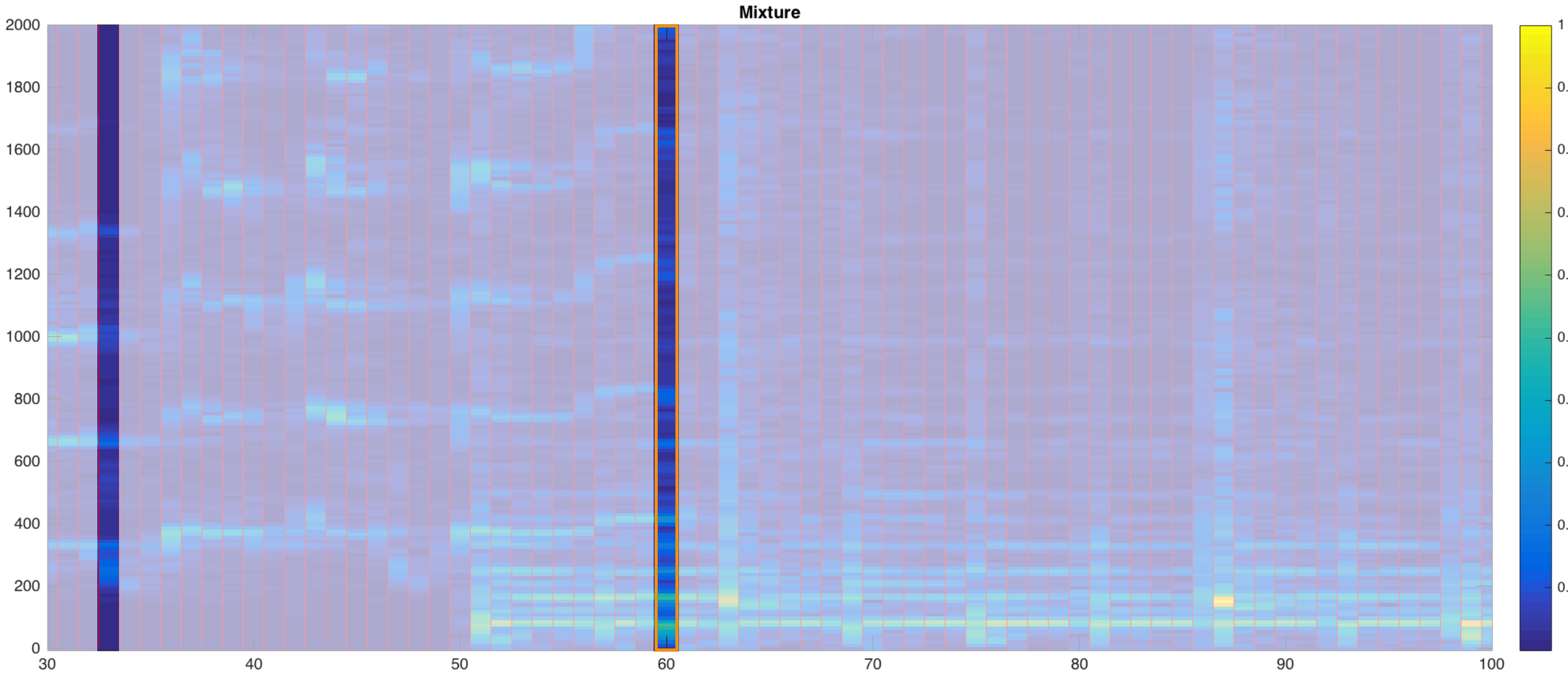
DISTANCE BETWEEN FRAMES

2. LOOK FOR SIMILAR TIME FRAMES



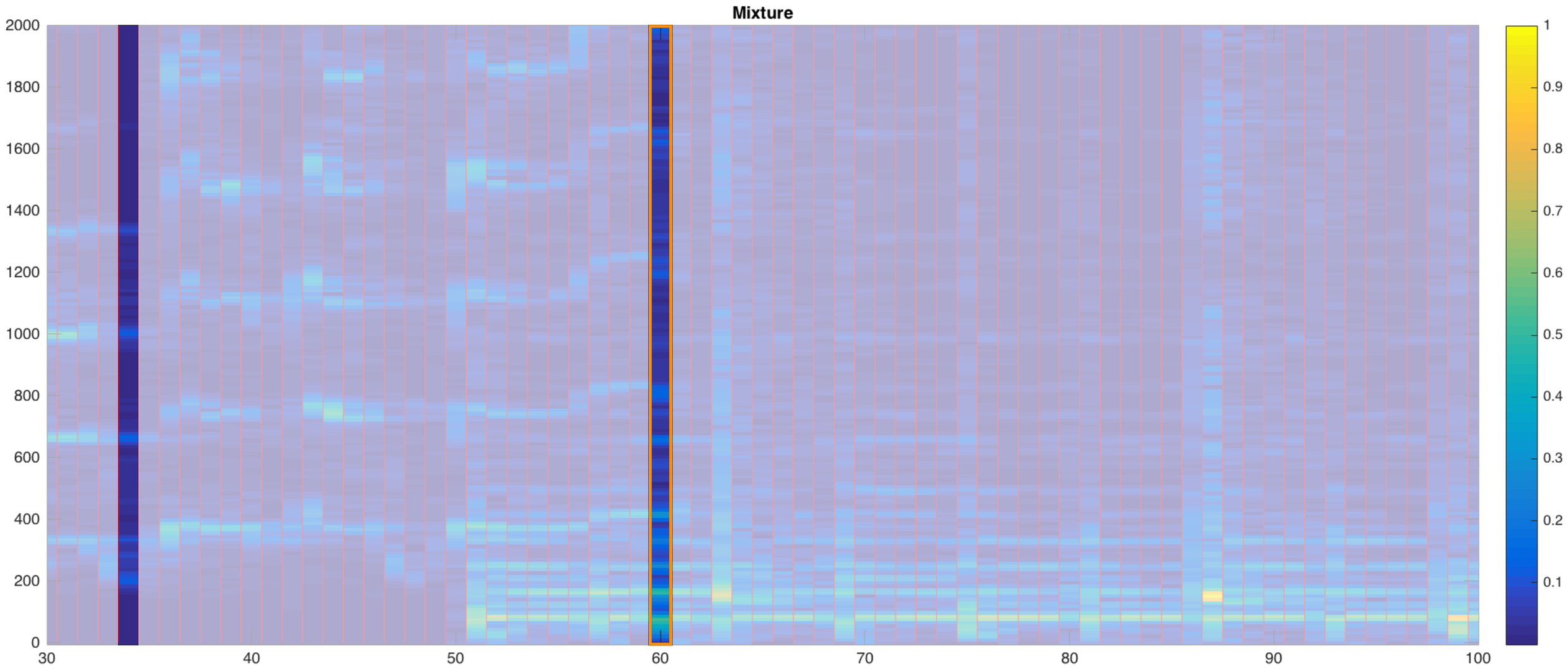
DISTANCE BETWEEN FRAMES

2. LOOK FOR SIMILAR TIME FRAMES



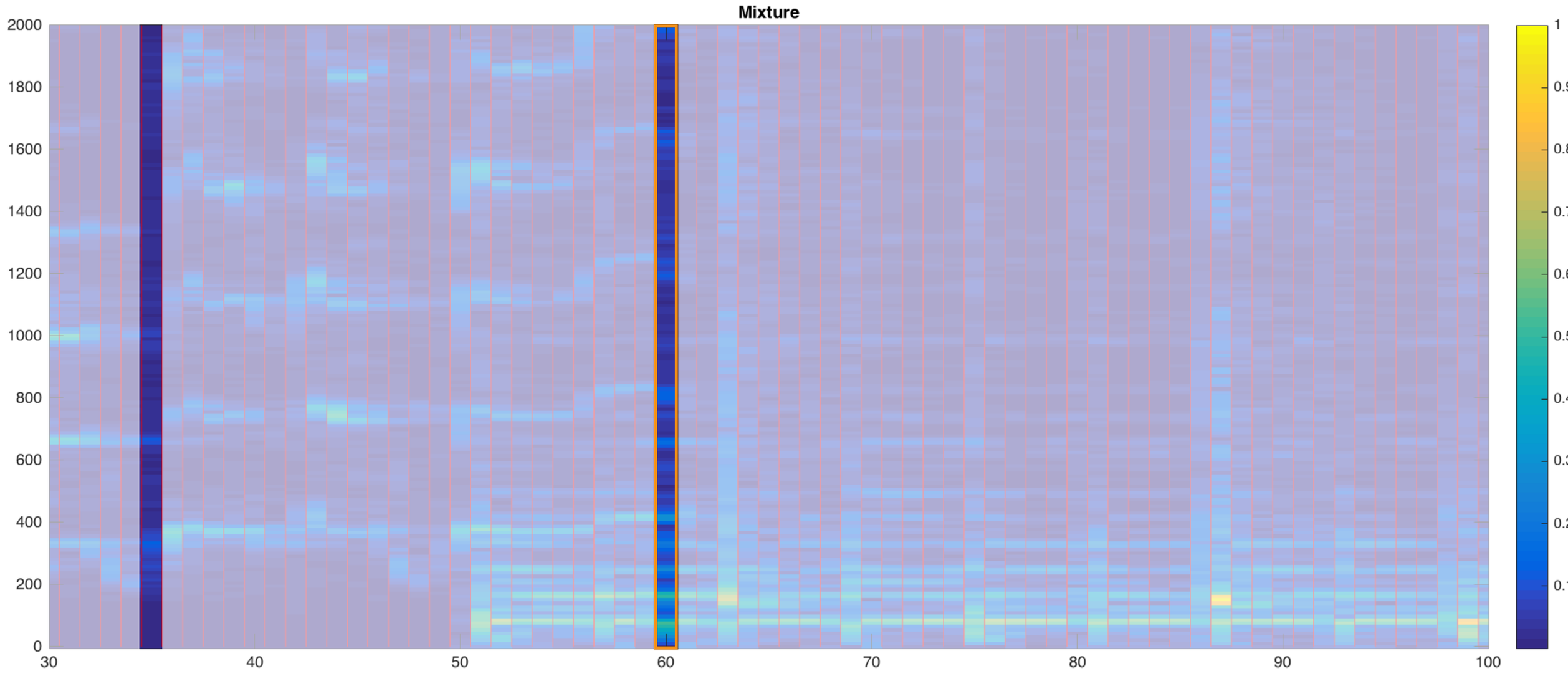
DISTANCE BETWEEN FRAMES

2. LOOK FOR SIMILAR TIME FRAMES



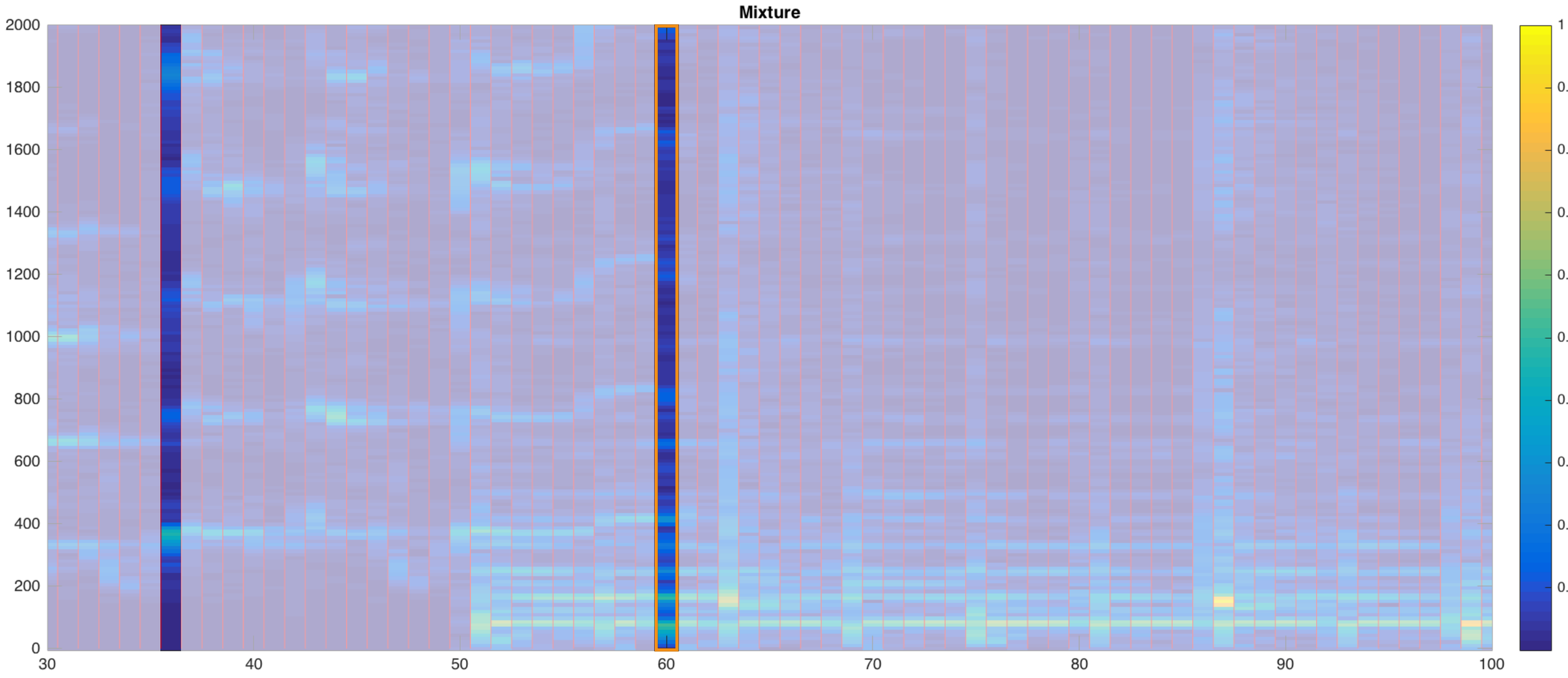
DISTANCE BETWEEN FRAMES

2. LOOK FOR SIMILAR TIME FRAMES



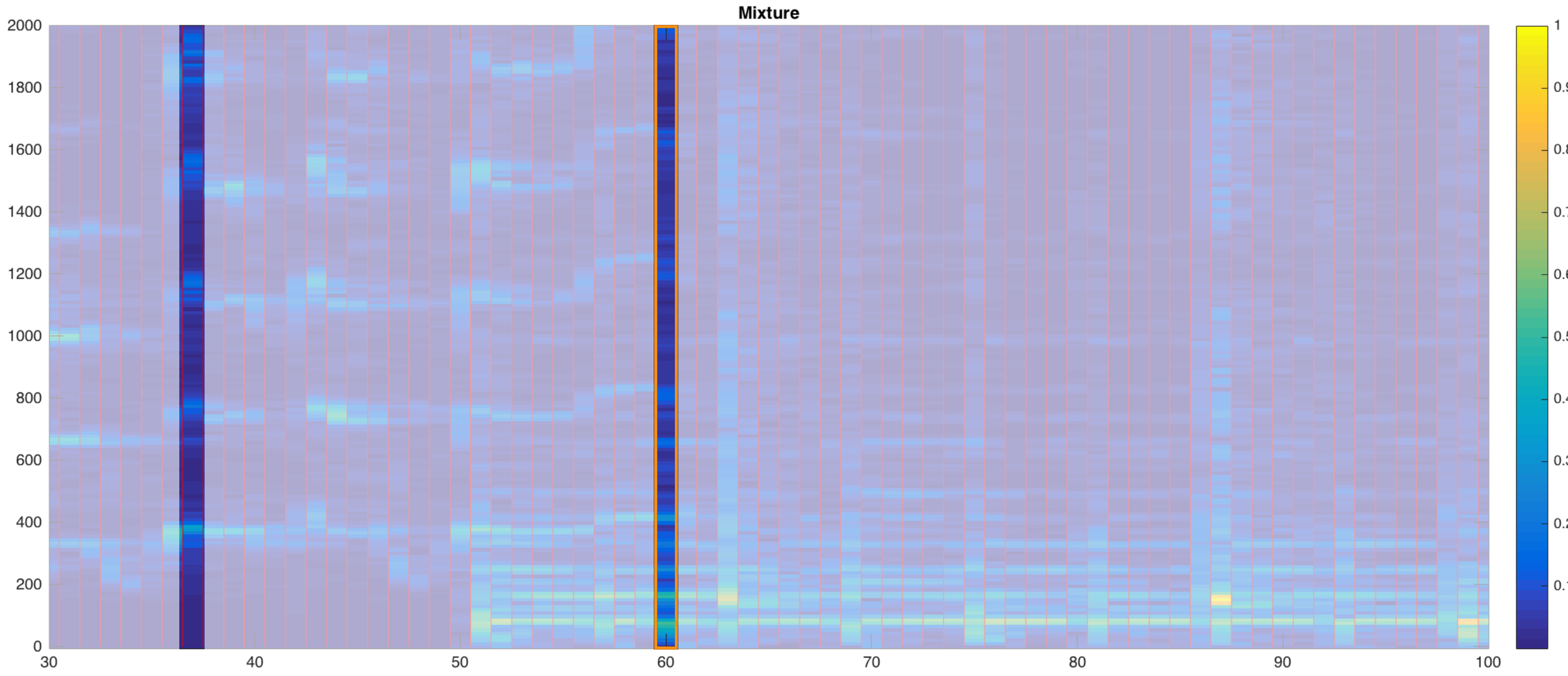
DISTANCE BETWEEN FRAMES

2. LOOK FOR SIMILAR TIME FRAMES



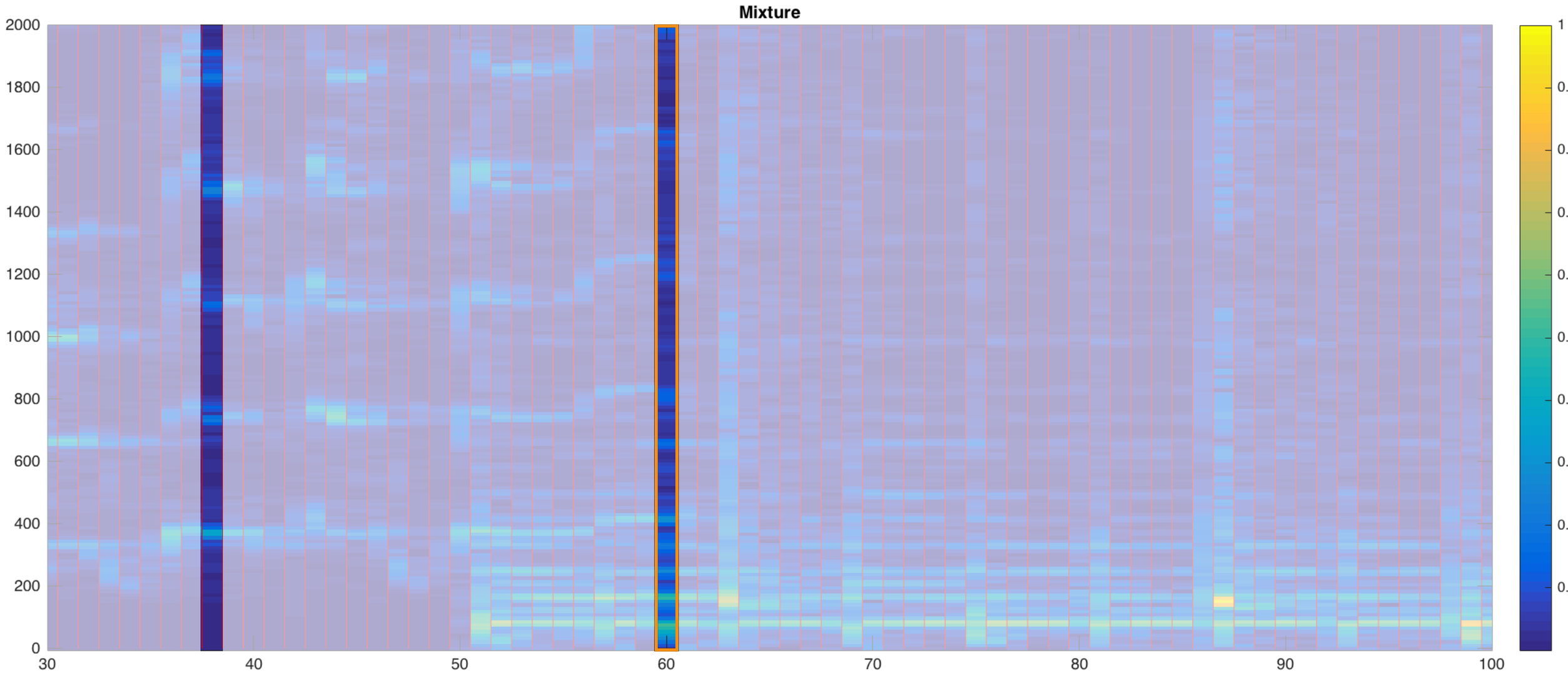
DISTANCE BETWEEN FRAMES

2. LOOK FOR SIMILAR TIME FRAMES



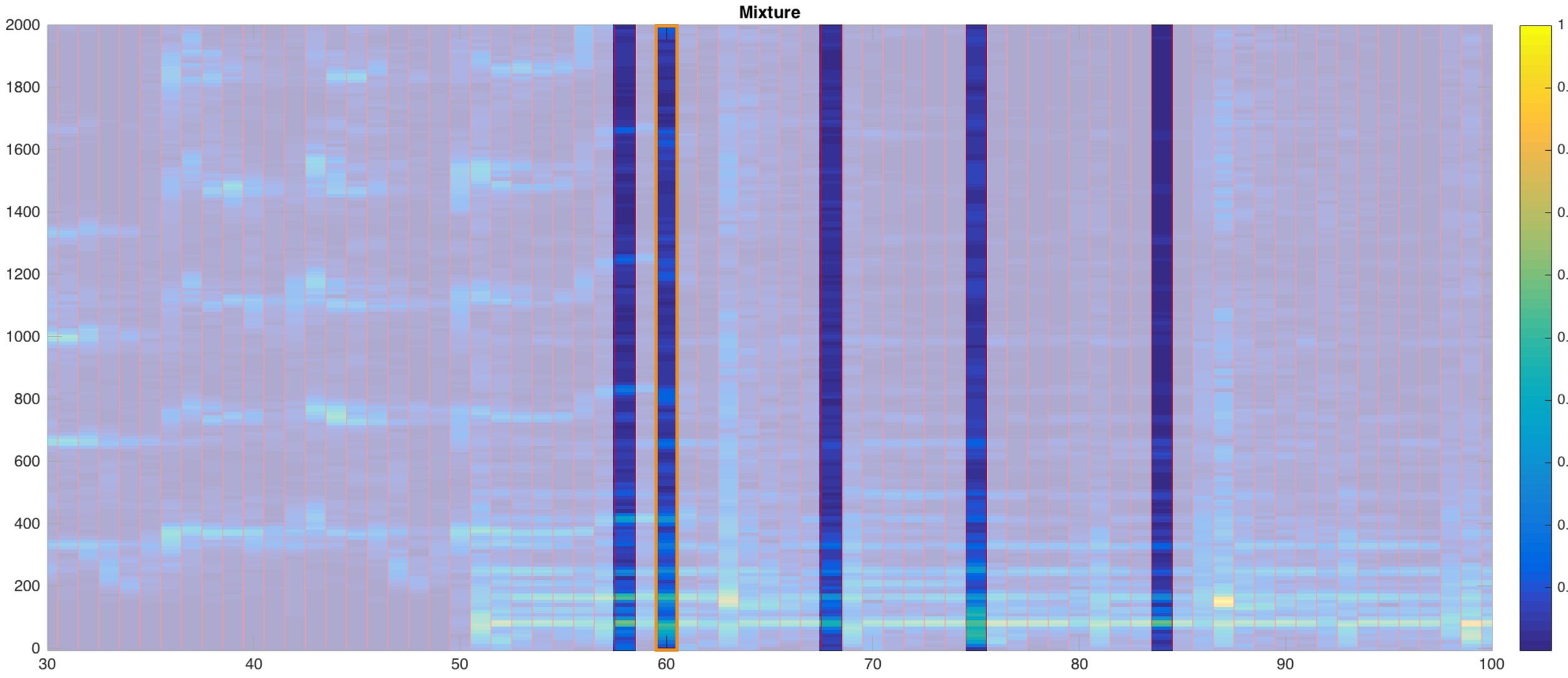
DISTANCE BETWEEN FRAMES

2. LOOK FOR SIMILAR TIME FRAMES

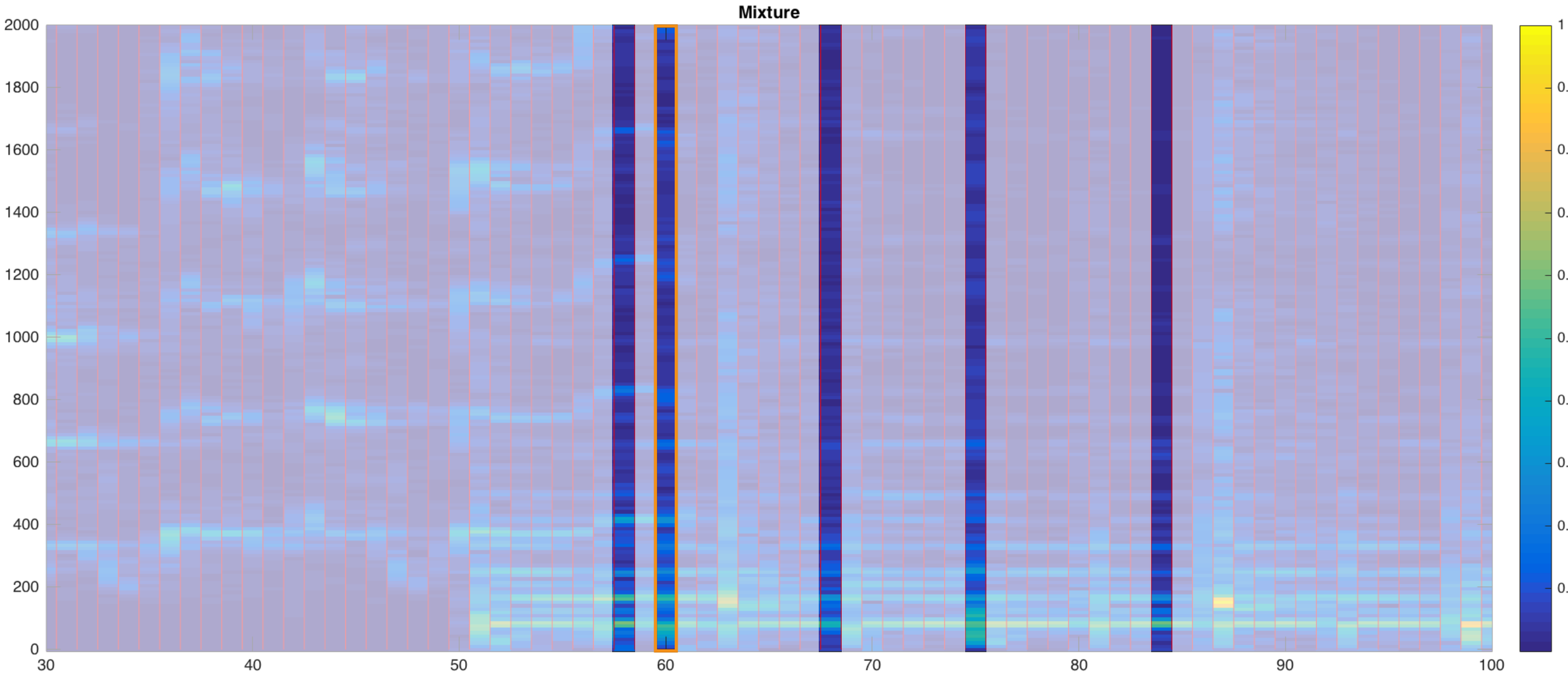


DISTANCE BETWEEN FRAMES

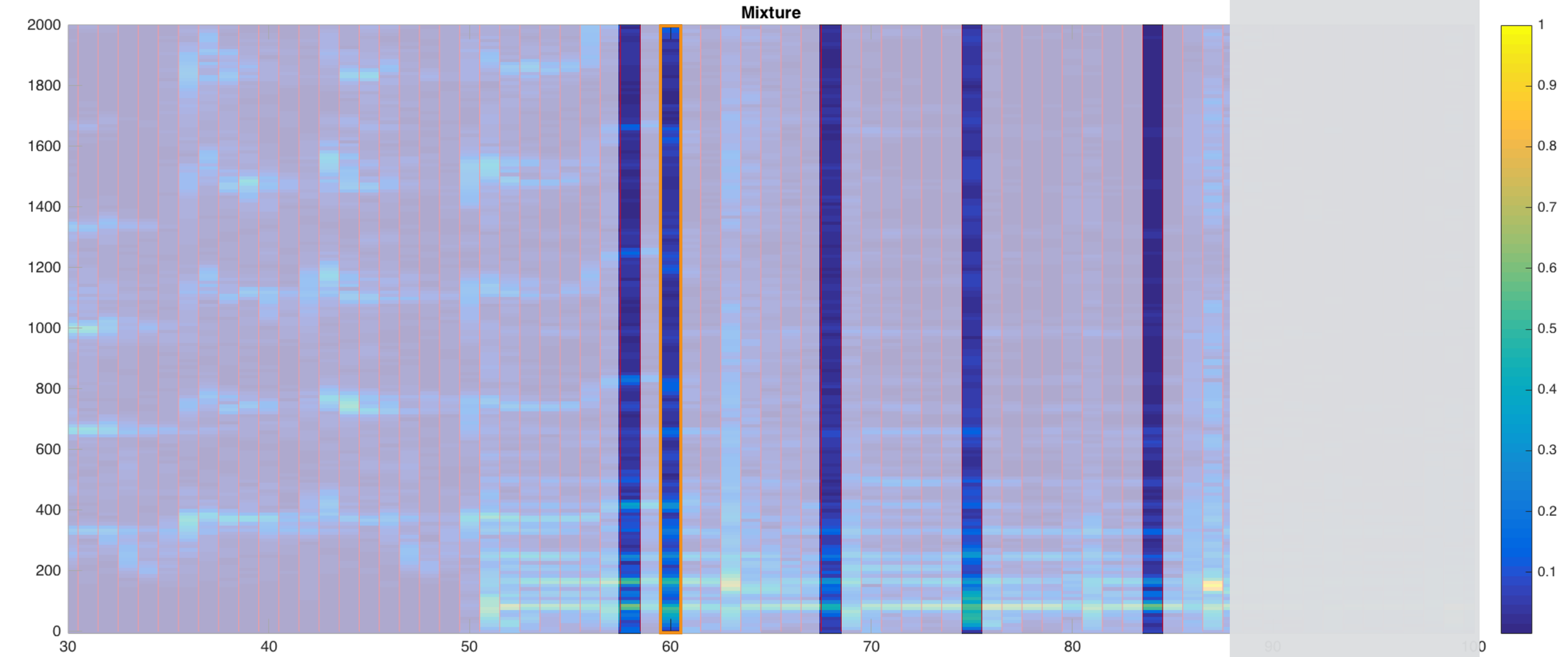
3. STORE THE N-CLOSEST TIME FRAMES



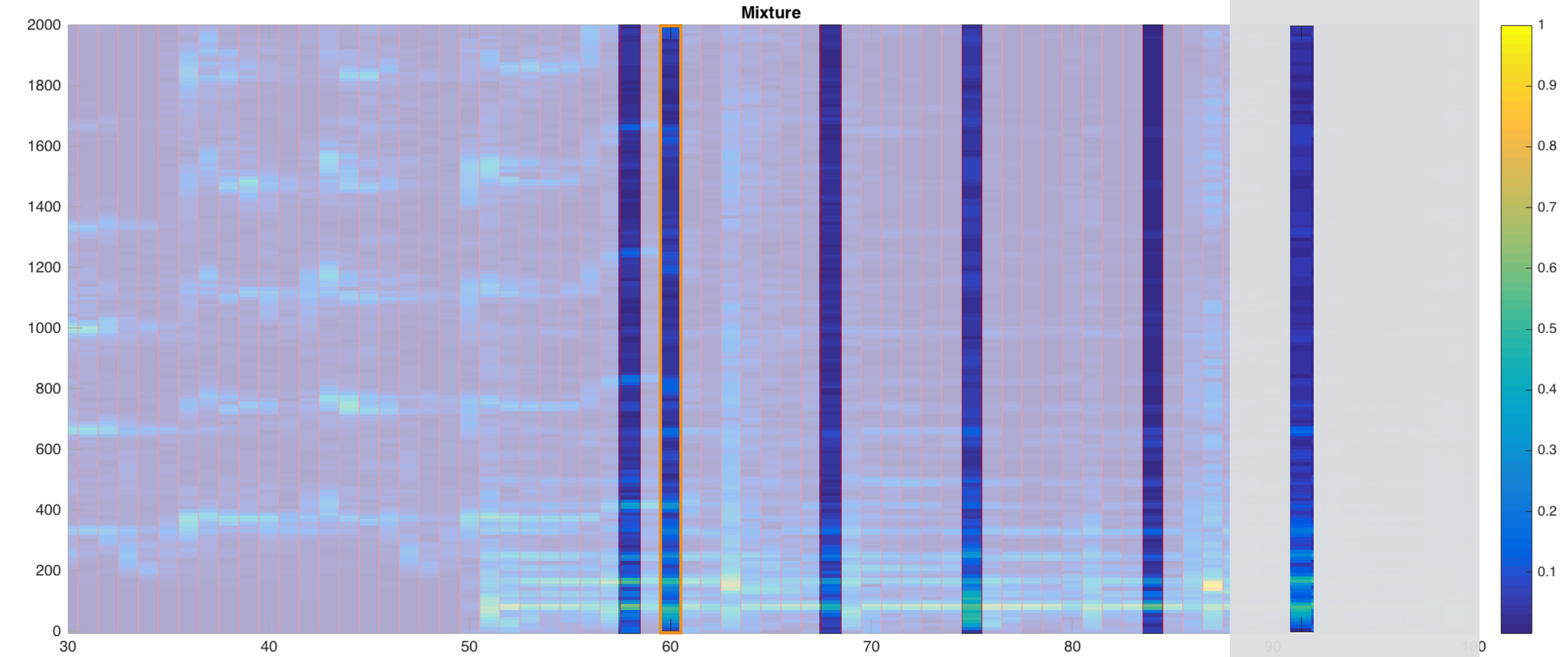
4. MEDIAN FILTERING



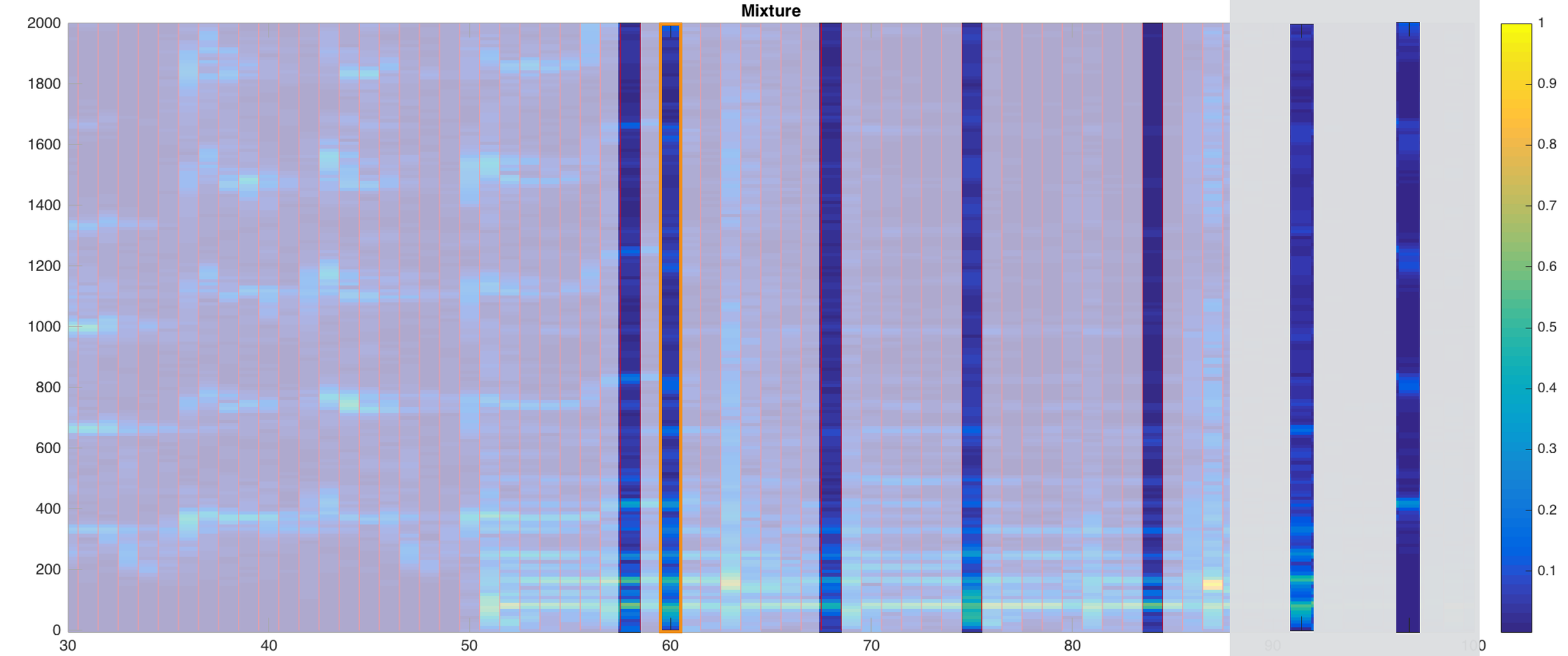
4. MEDIAN FILTERING

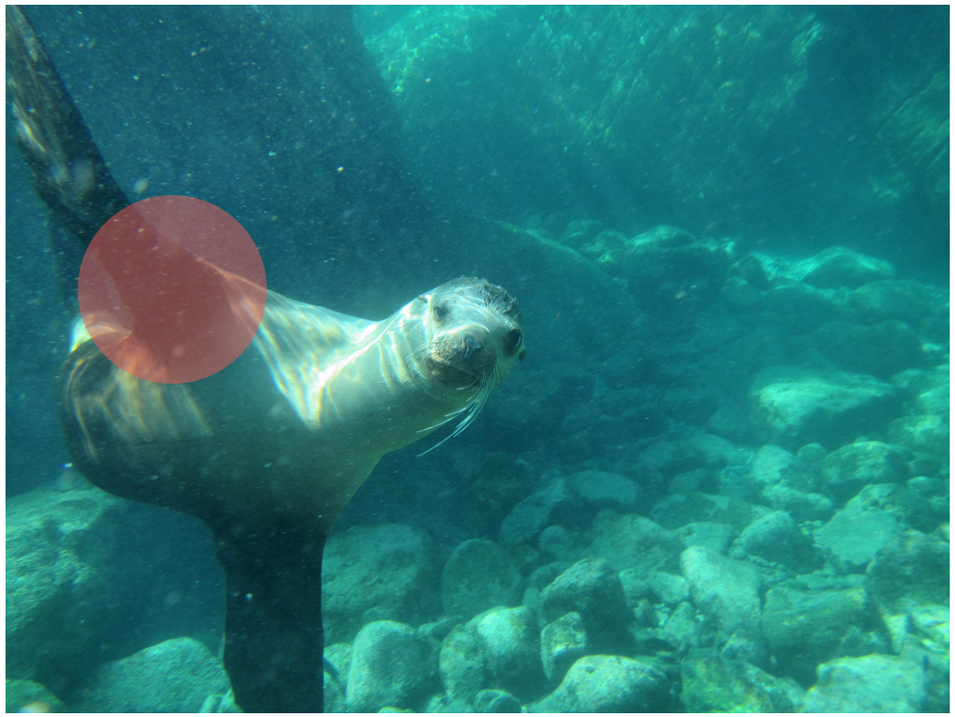


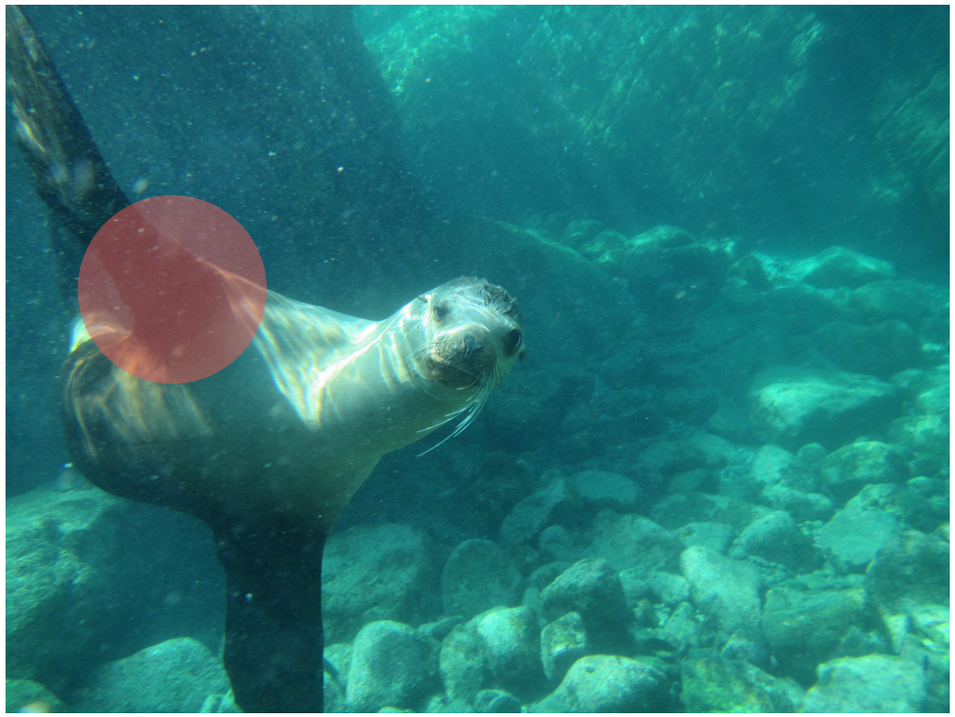
4. MEDIAN FILTERING

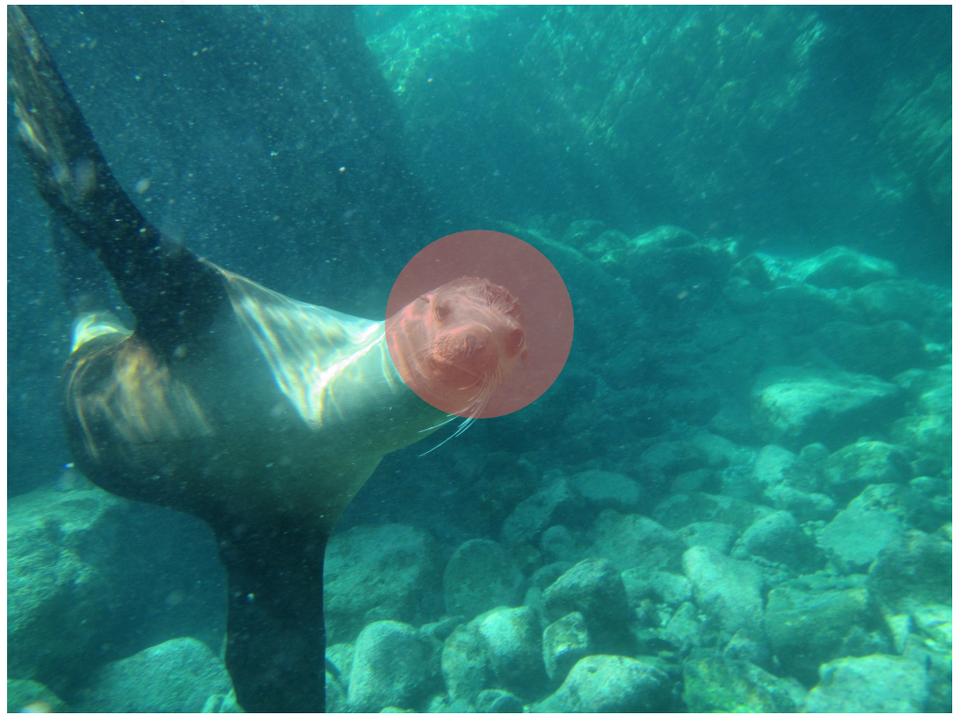


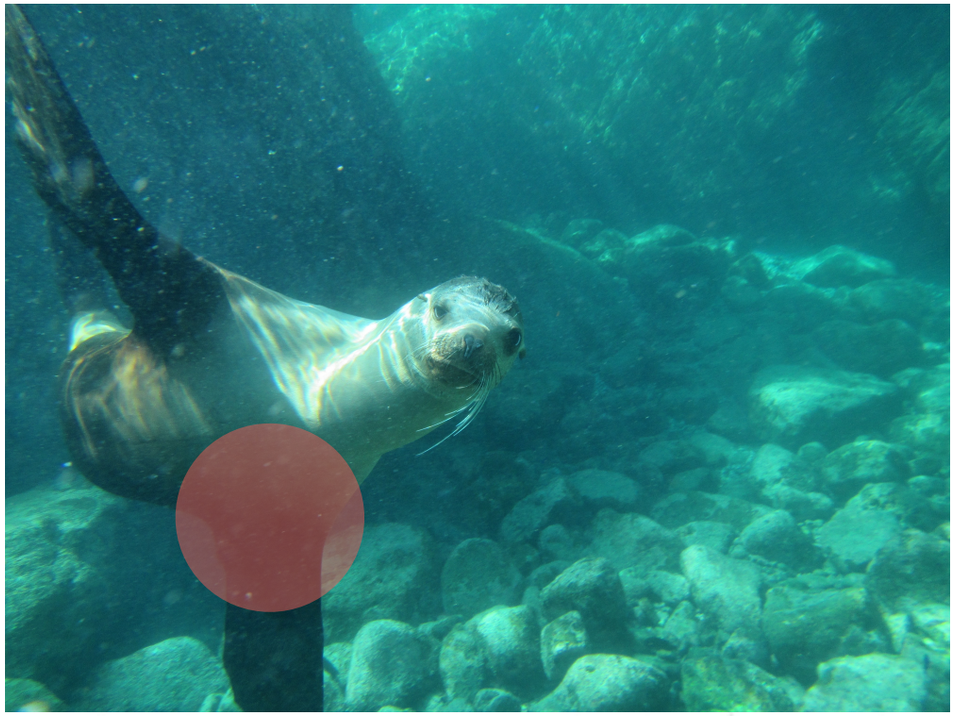
4. MEDIAN FILTERING

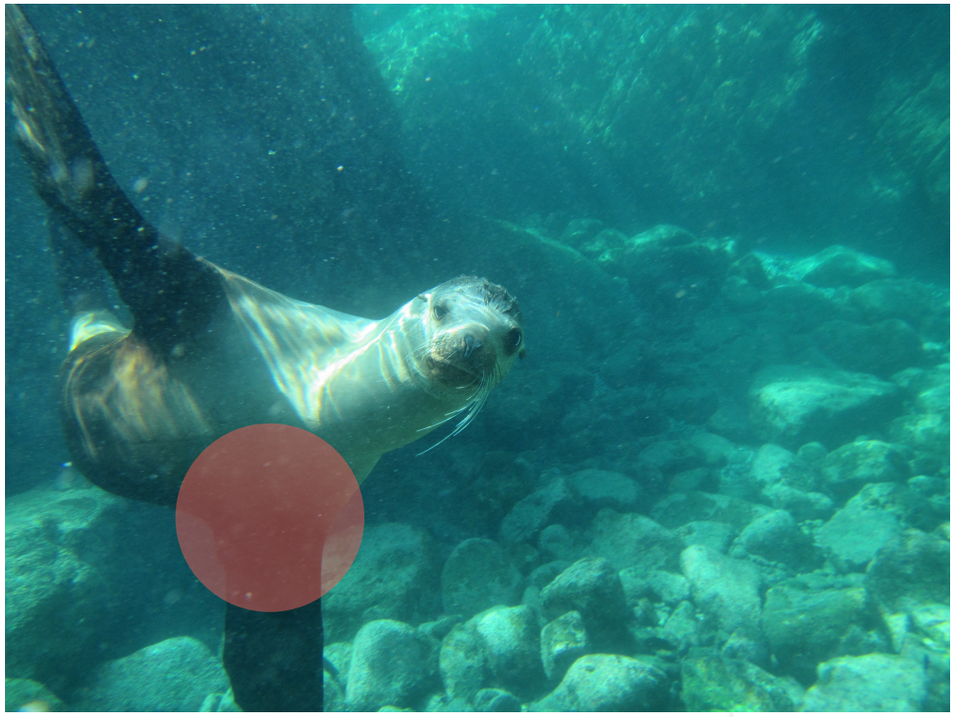


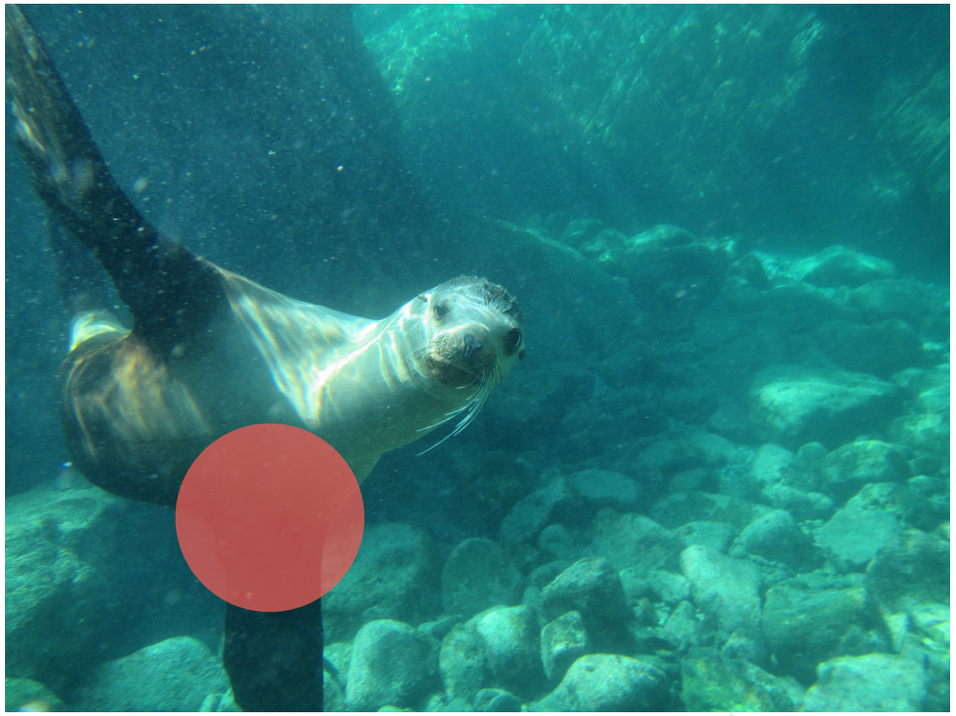




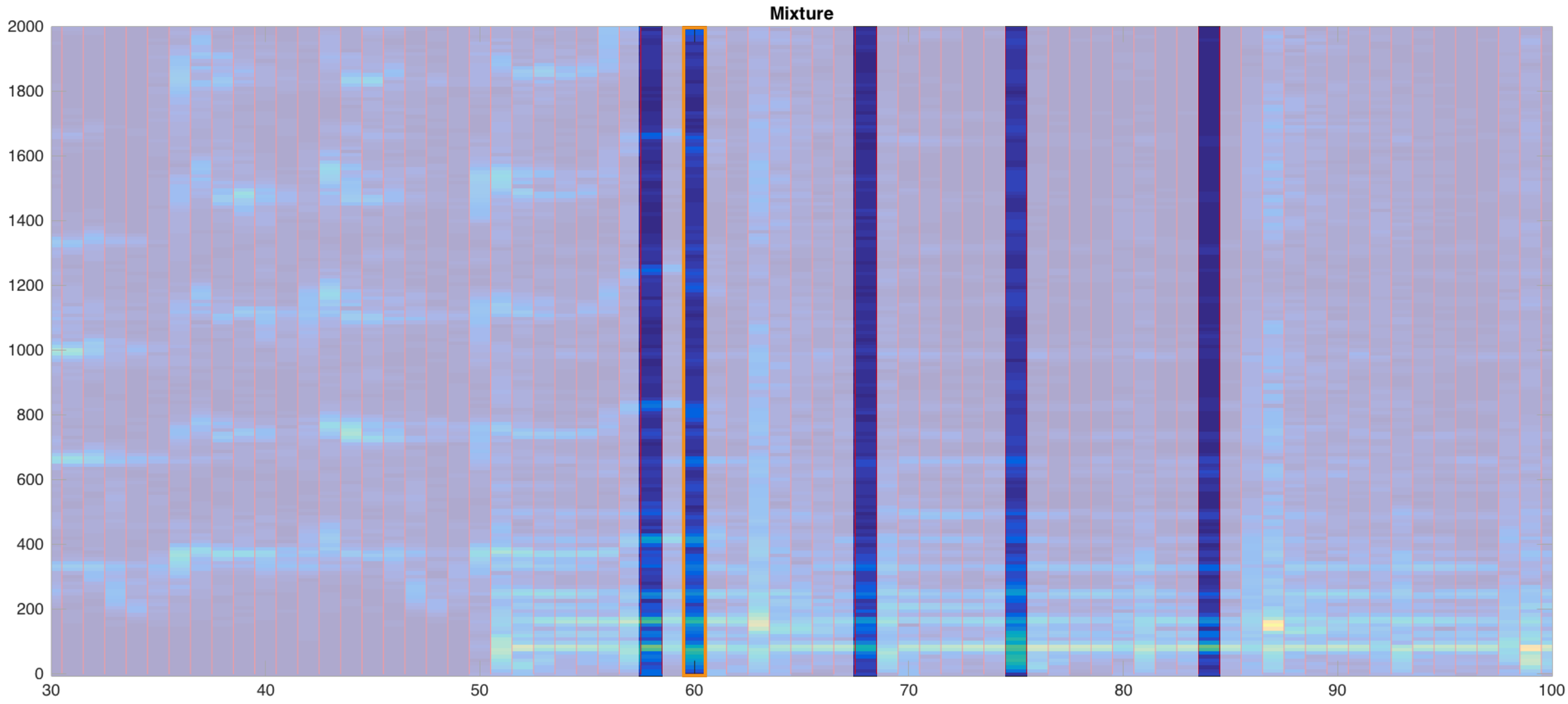


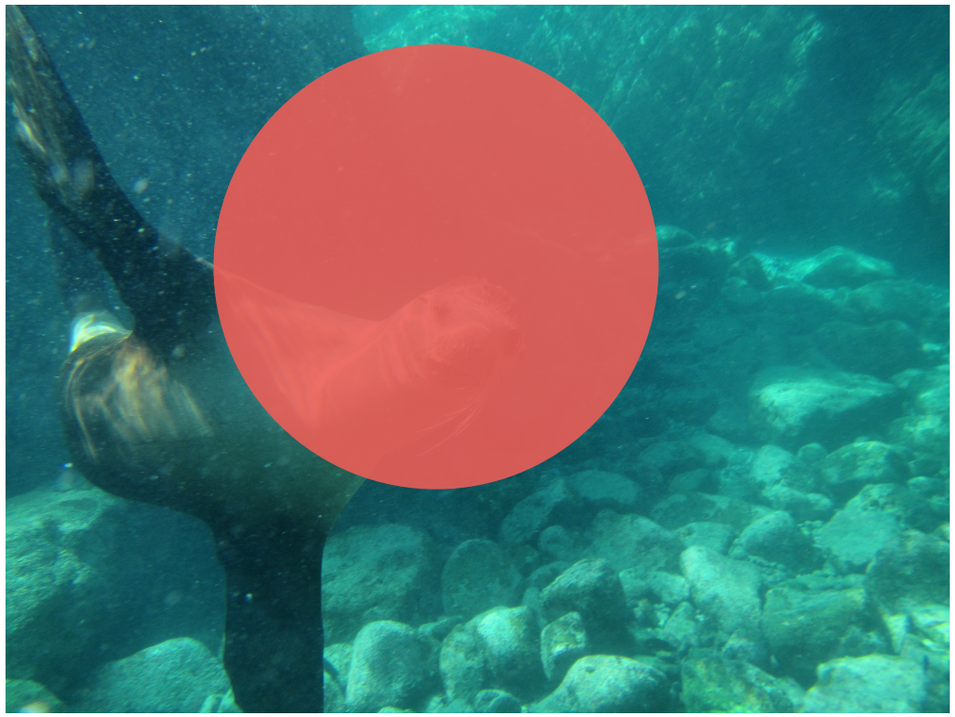


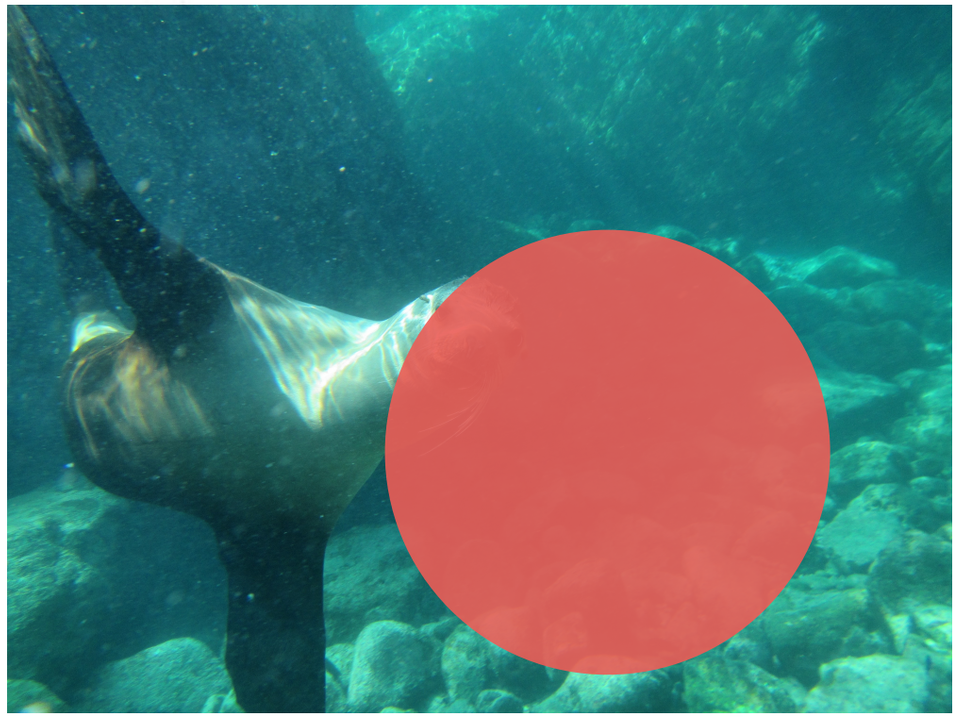


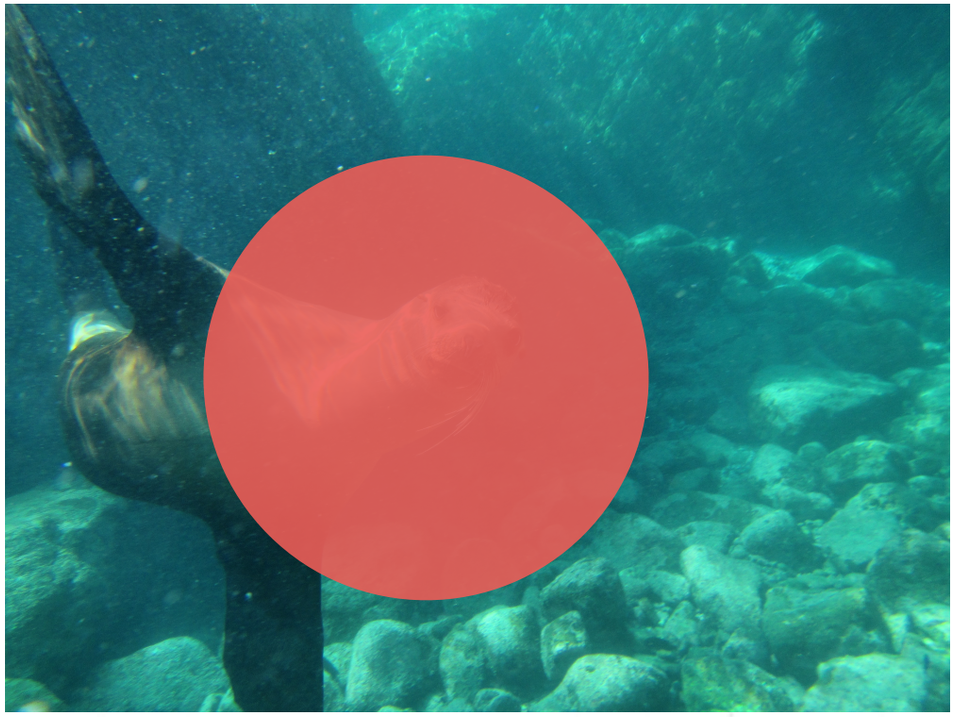


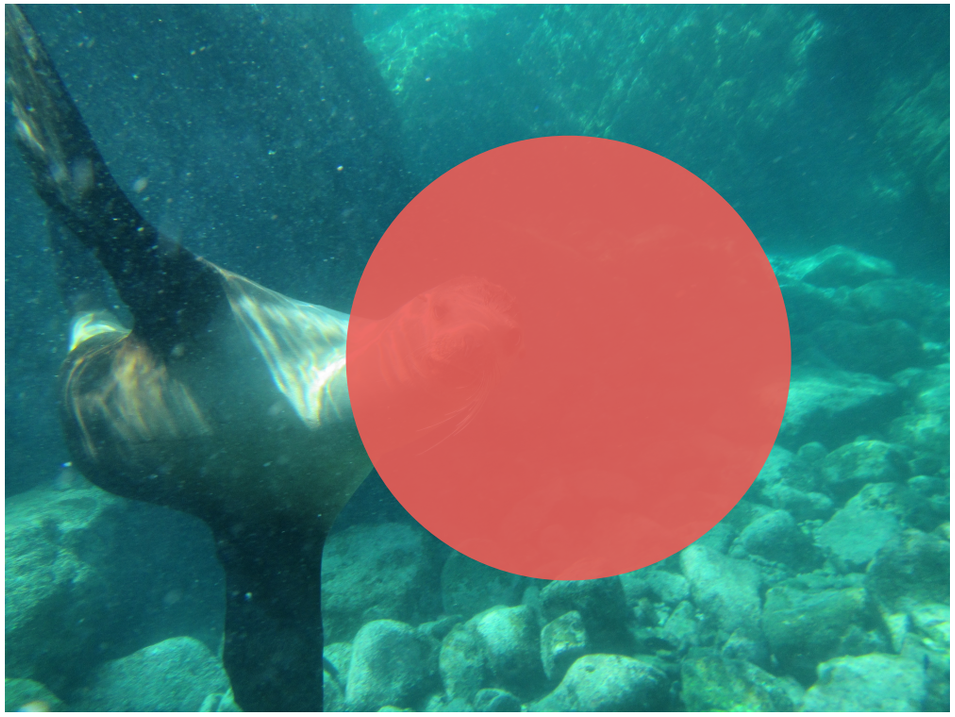
N-CLOSEST TIME FRAMES

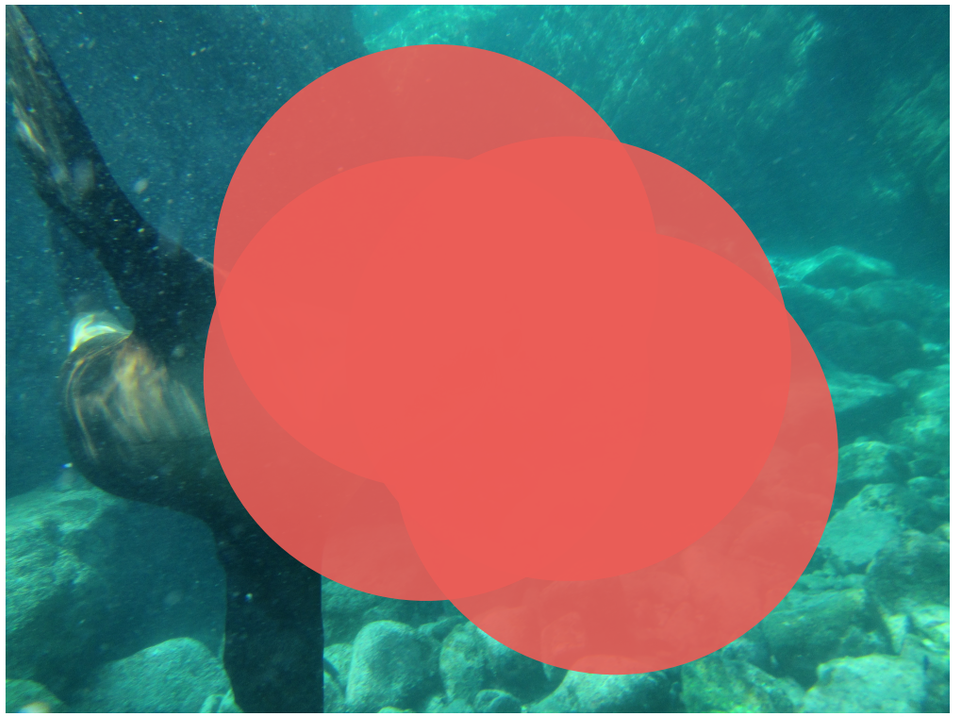




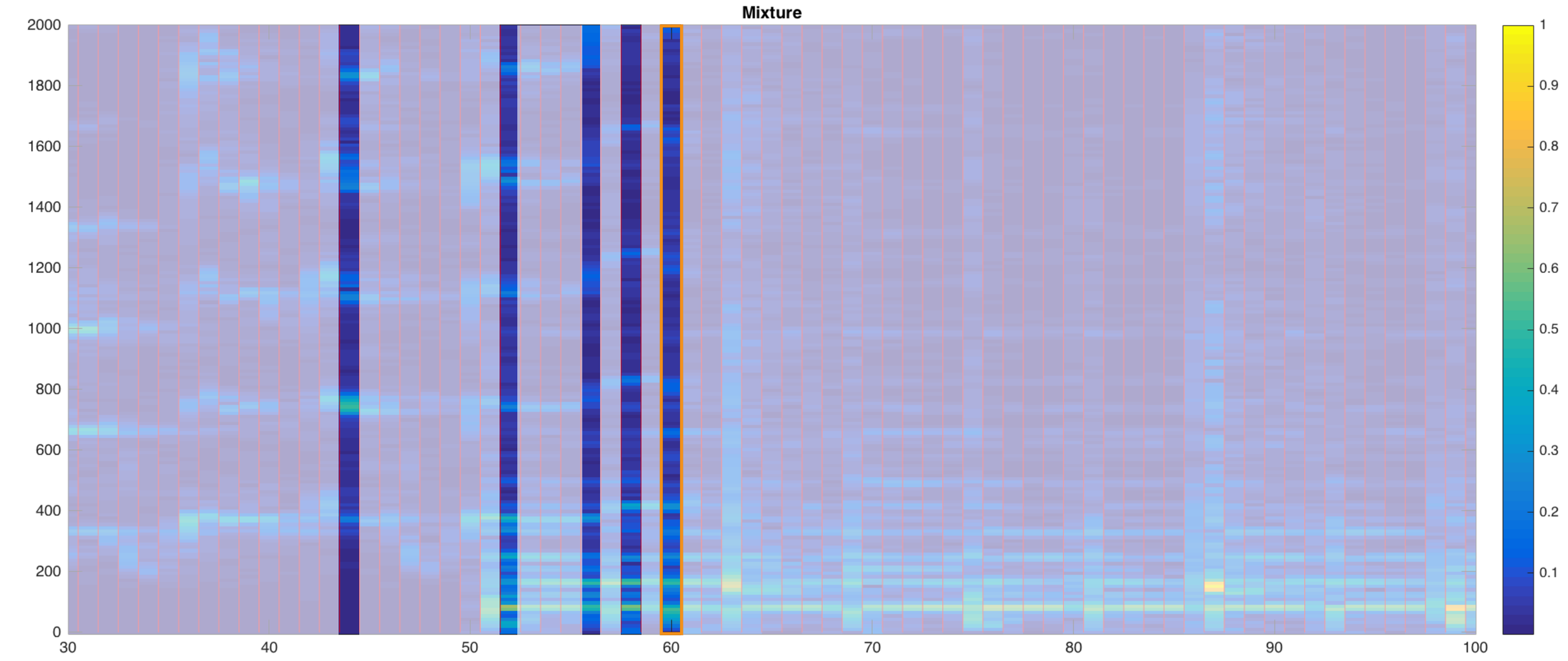




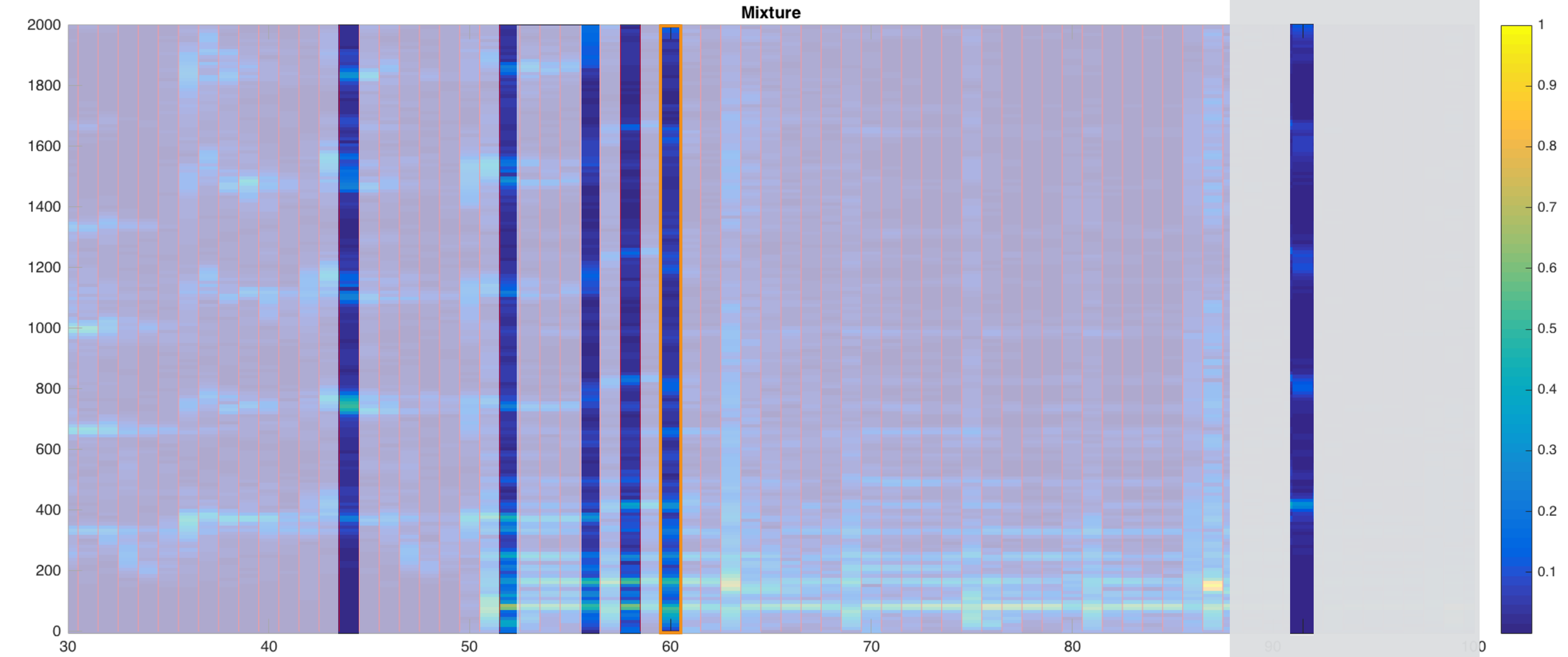




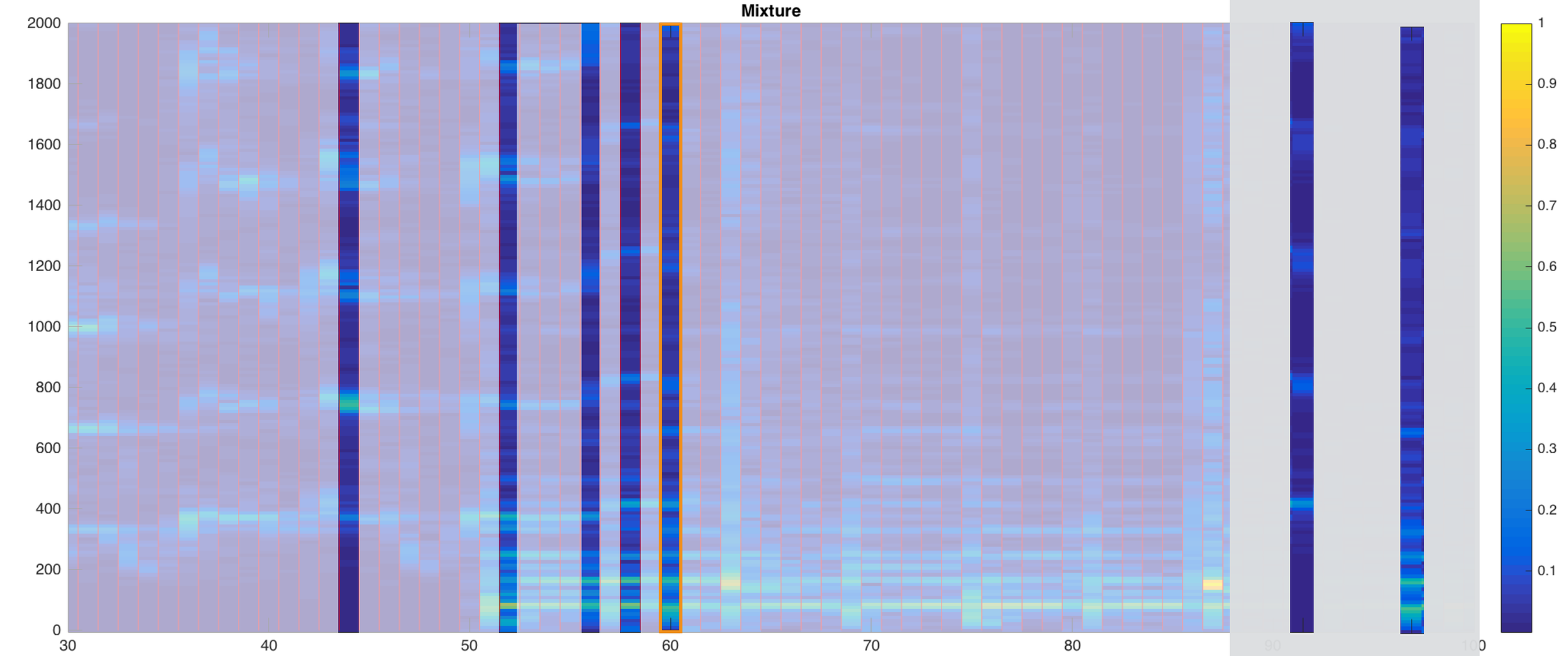
N-CLOSEST TIME FRAMES



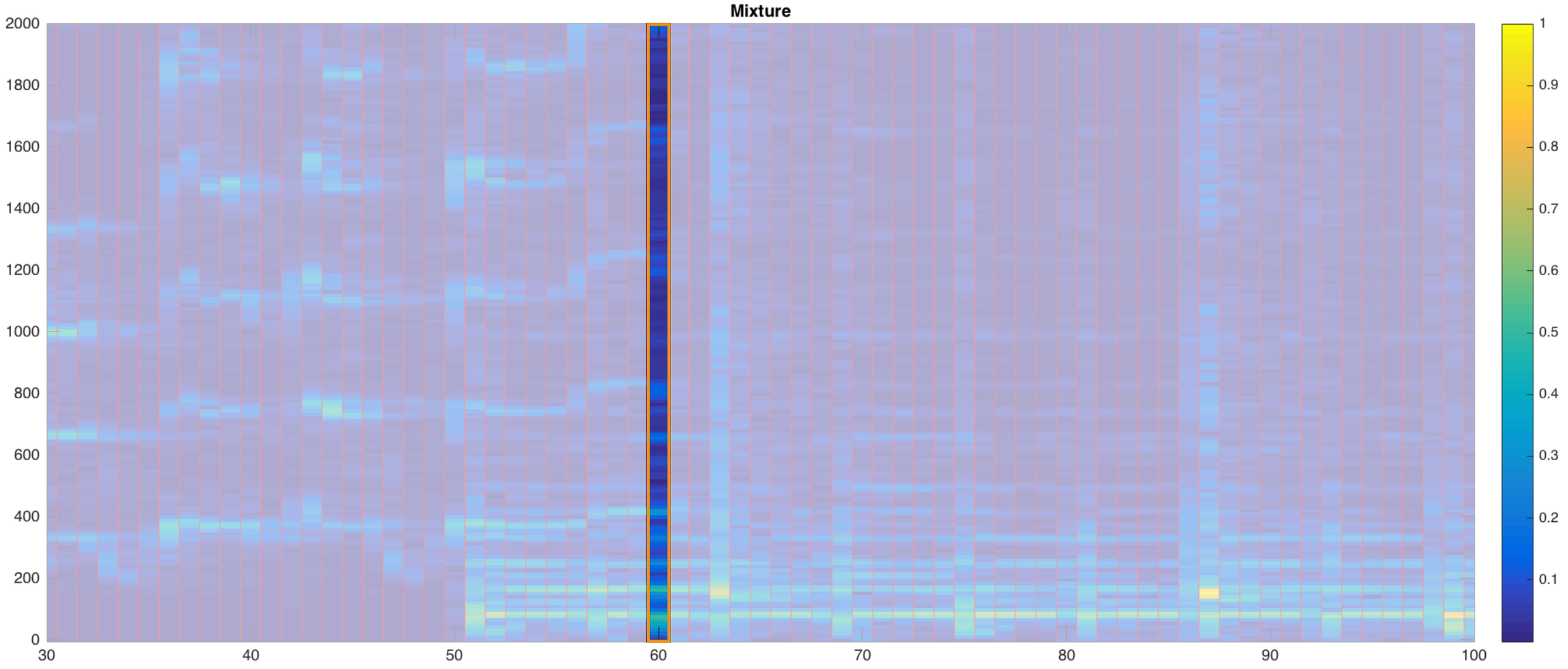
N-CLOSEST TIME FRAMES



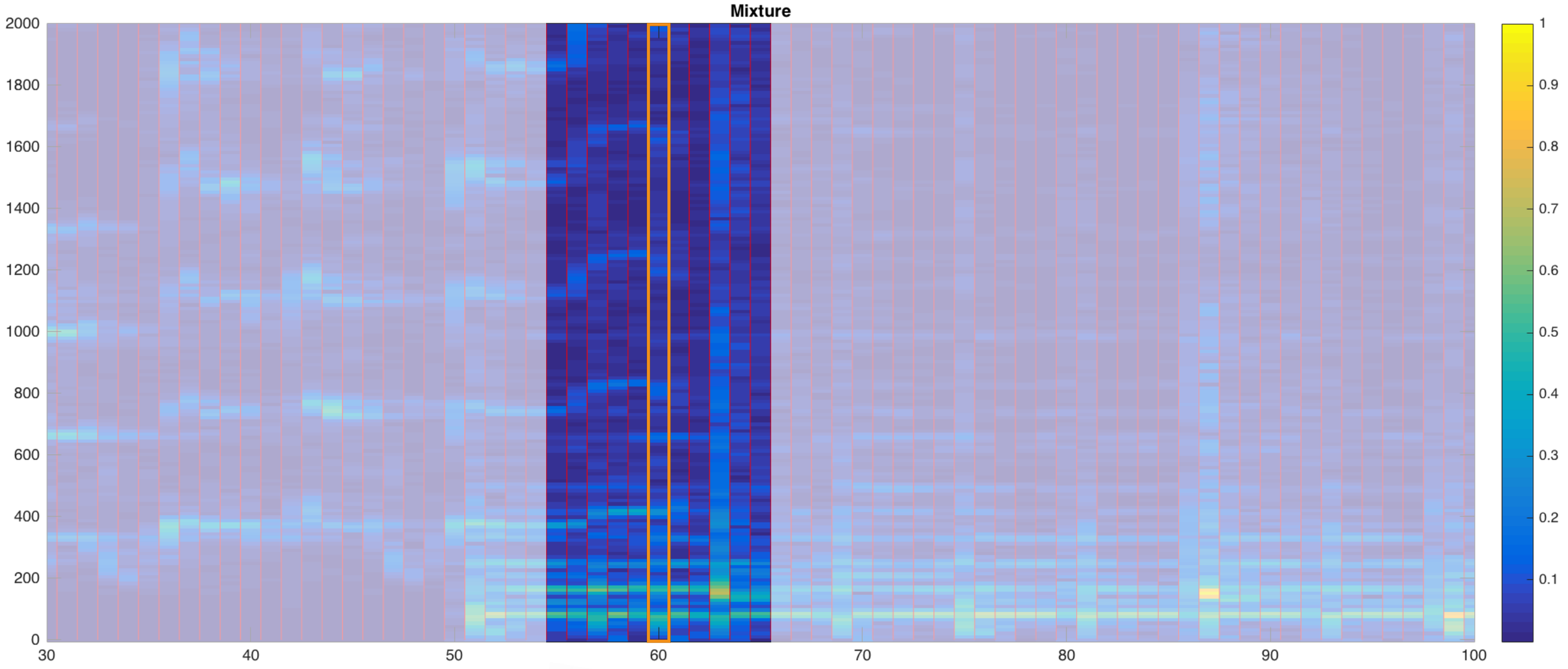
N-CLOSEST TIME FRAMES



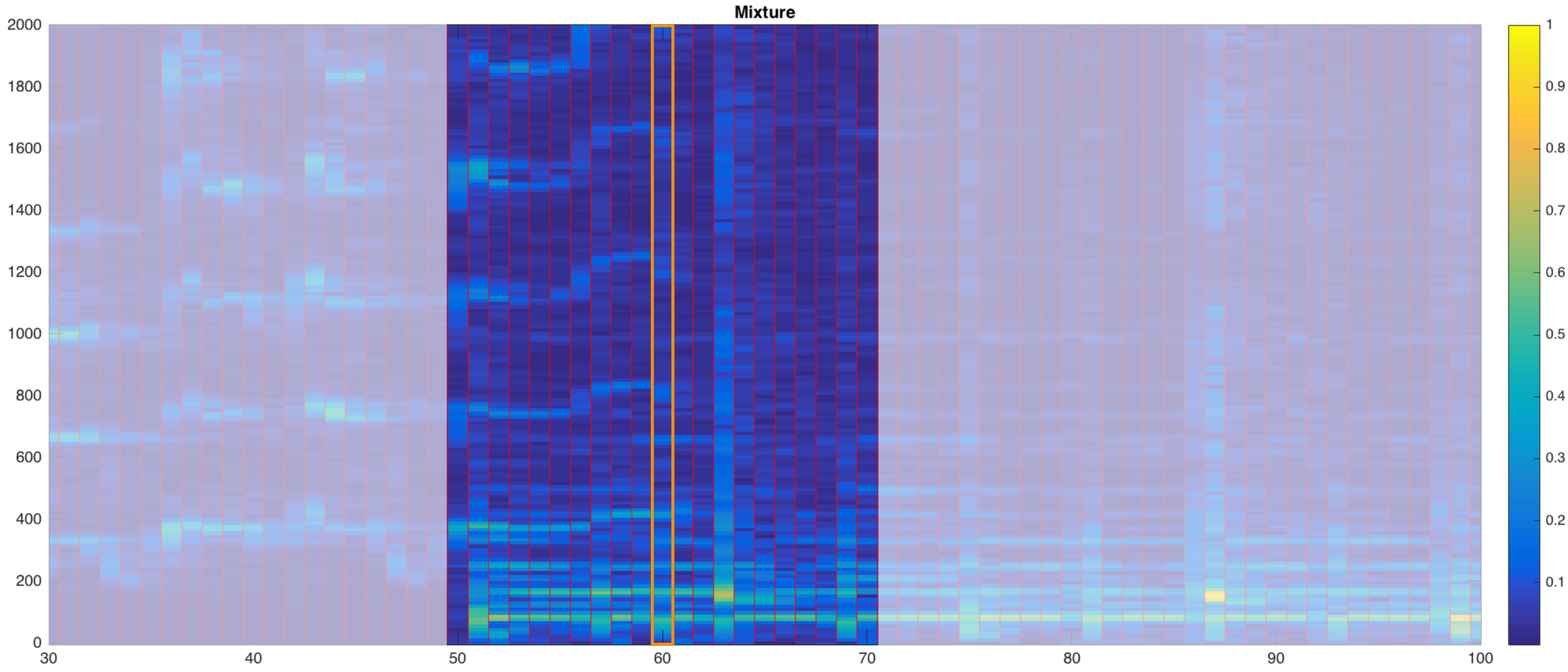
PROPOSED EXTENSION



PROPOSED EXTENSION

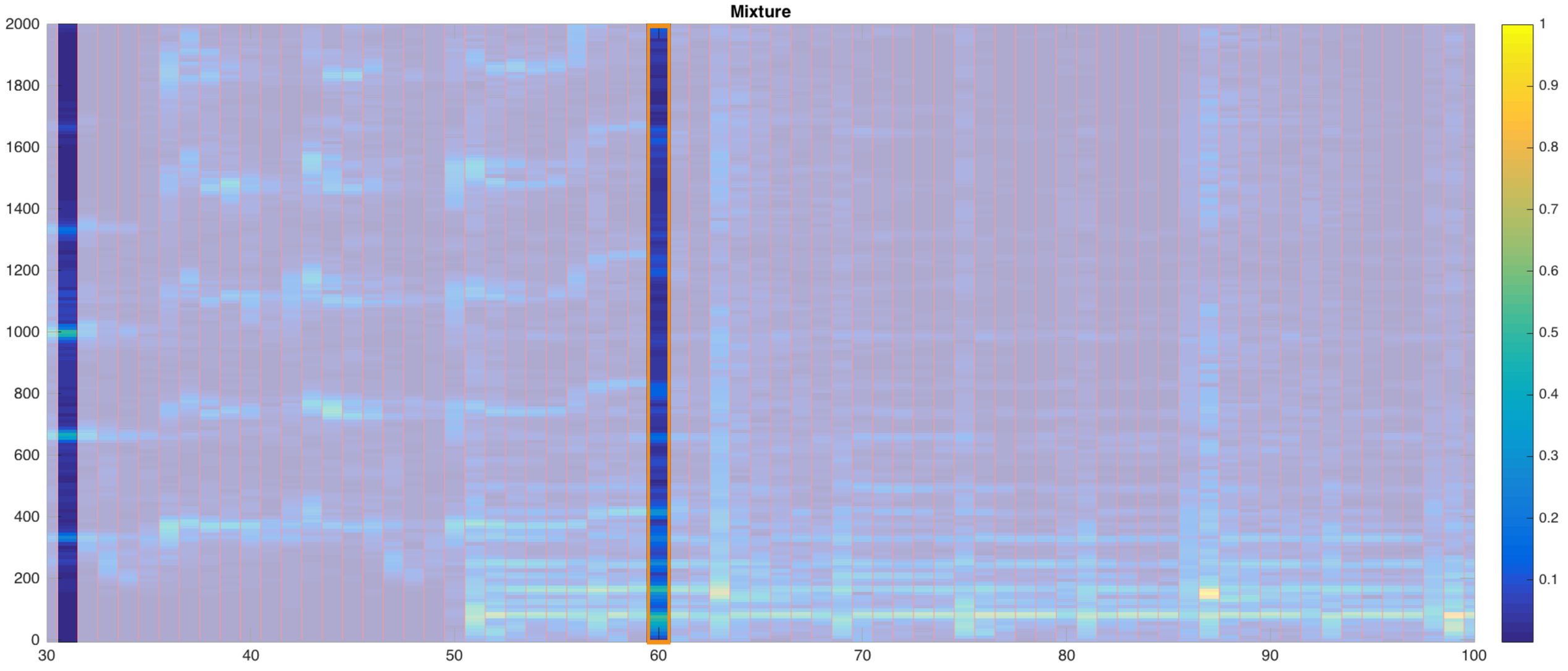


PROPOSED EXTENSION



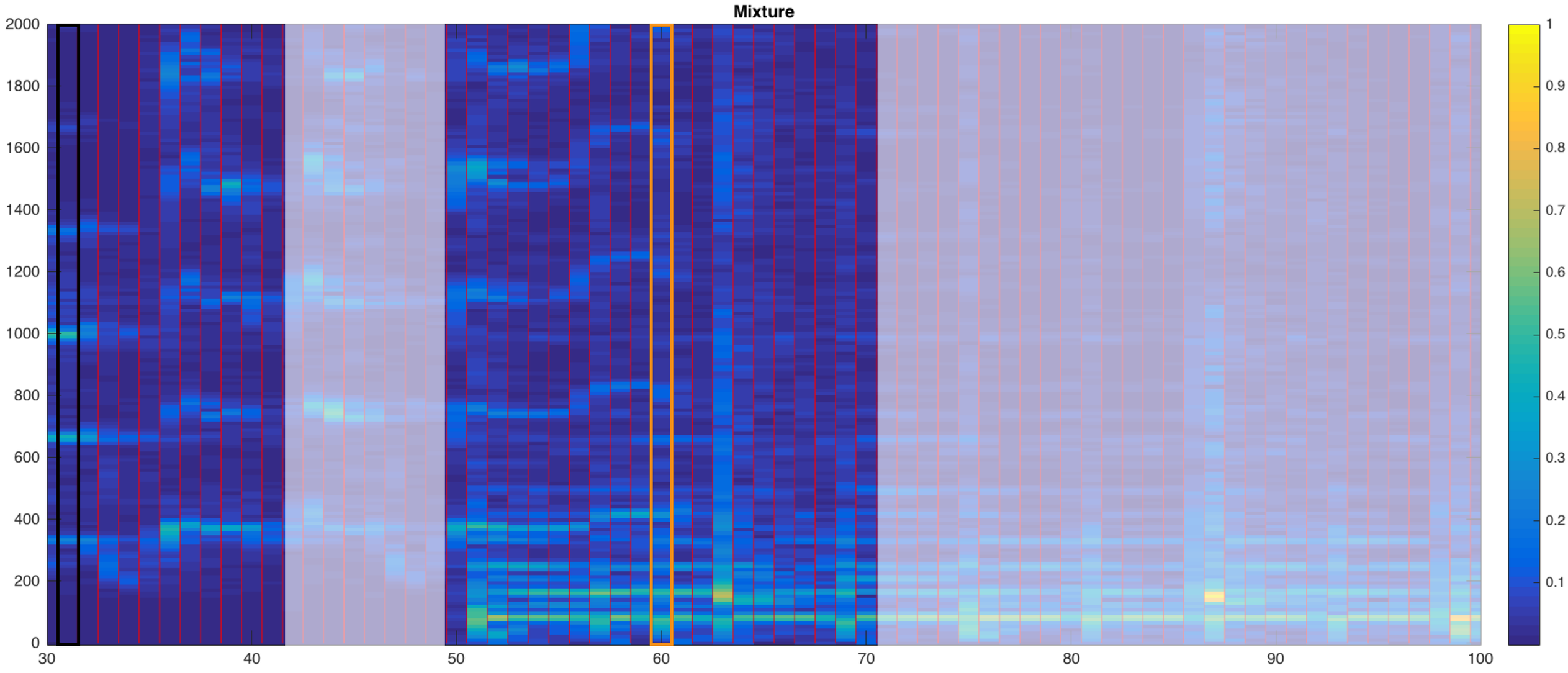
TEMPORAL CONTEXT

2. LOOK FOR SIMILAR TIME FRAMES



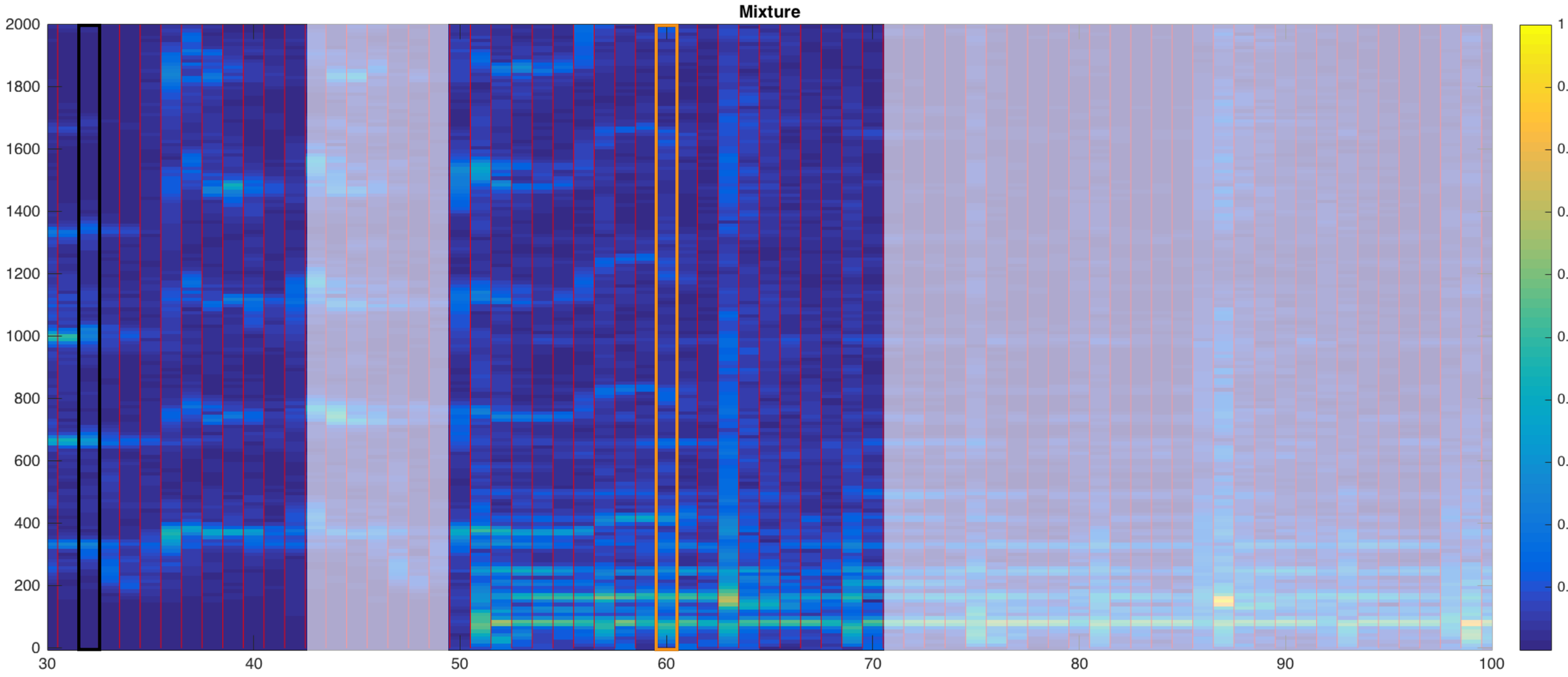
DISTANCE BETWEEN FRAMES

2. LOOK FOR SIMILAR TIME GROUP OF FRAMES



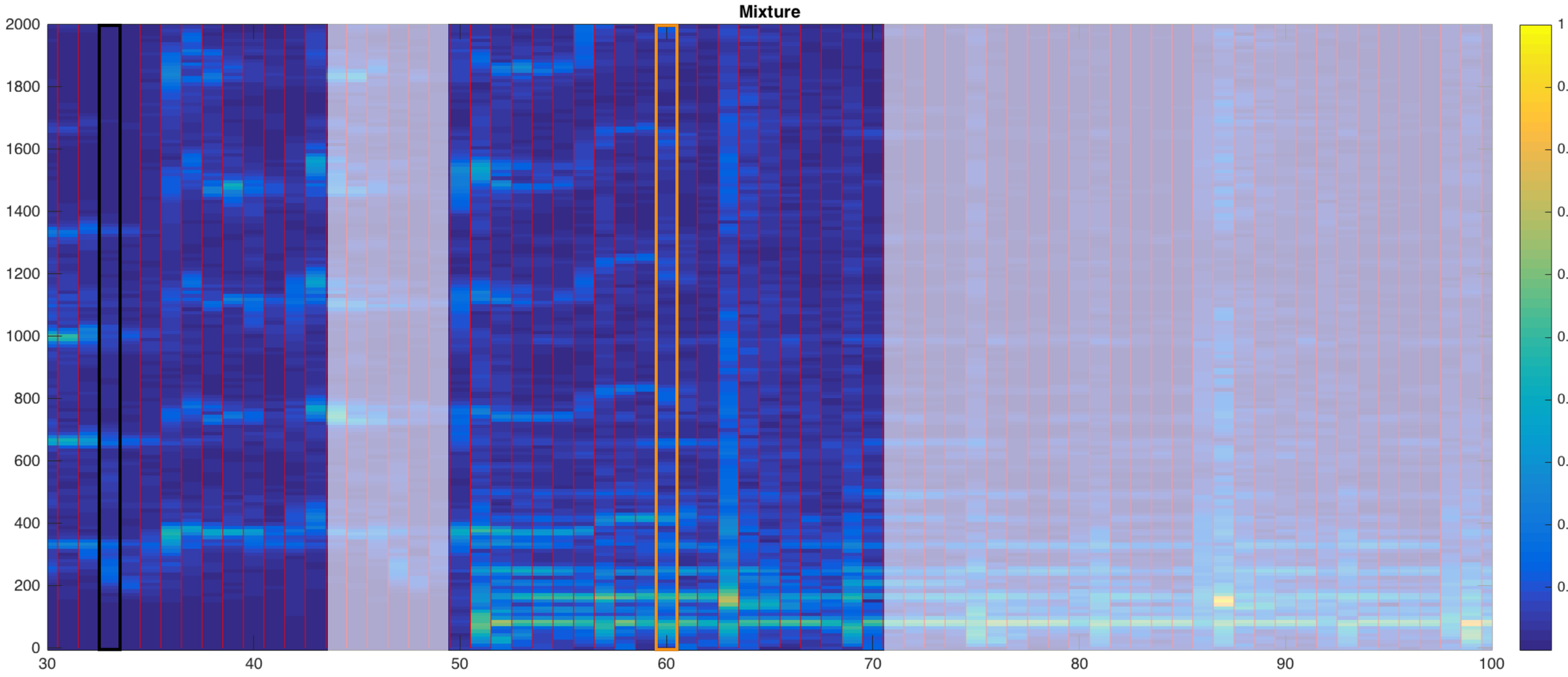
DISTANCE BETWEEN GROUP OF FRAMES

2. LOOK FOR SIMILAR TIME FRAMES



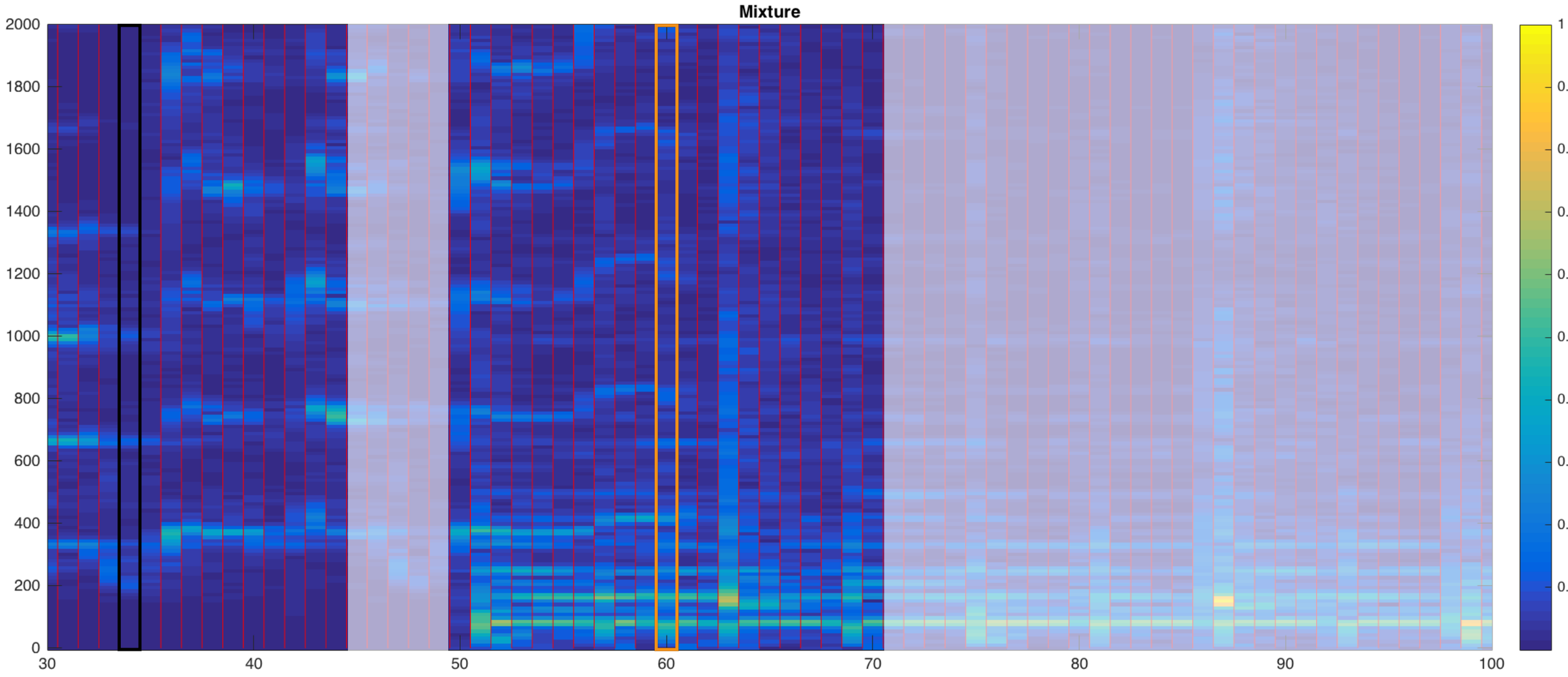
DISTANCE BETWEEN GROUP OF FRAMES

2. LOOK FOR SIMILAR TIME FRAMES



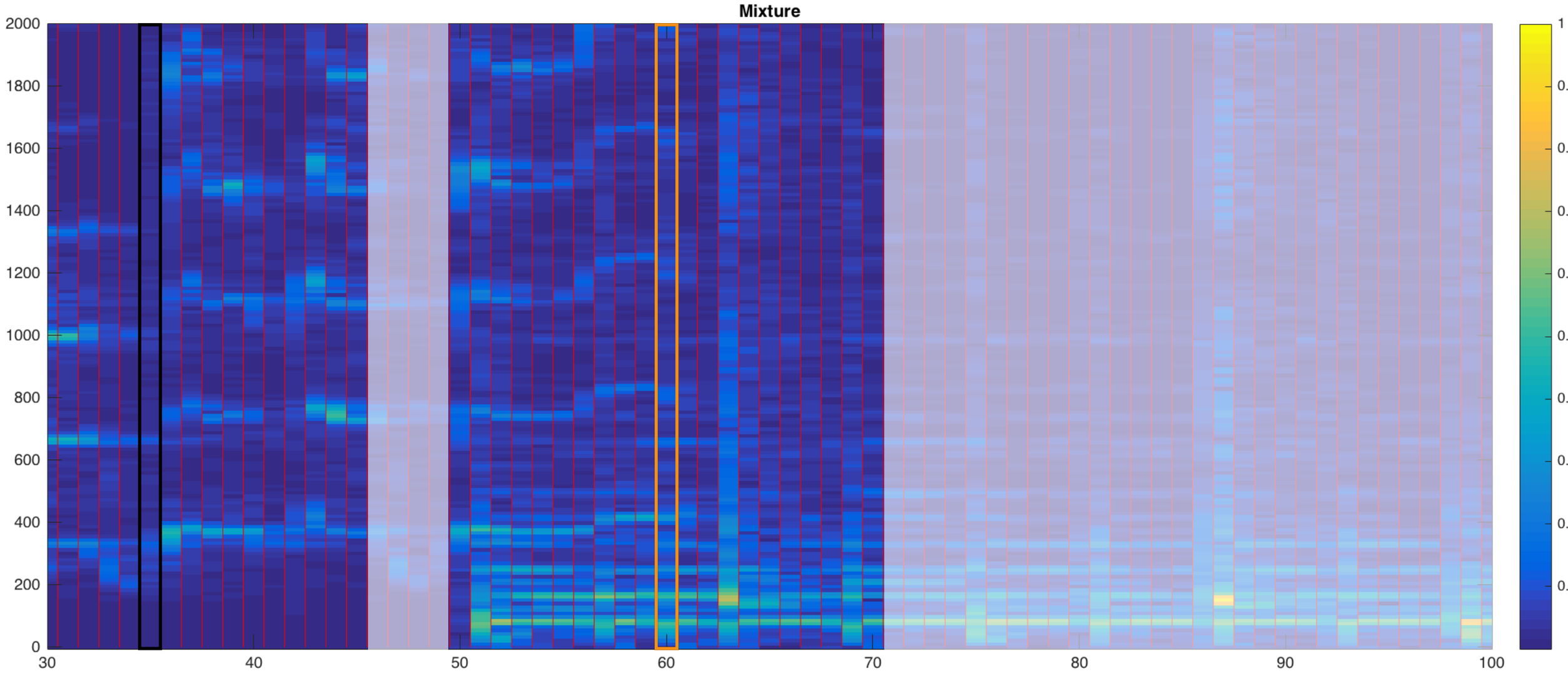
DISTANCE BETWEEN GROUP OF FRAMES

2. LOOK FOR SIMILAR TIME FRAMES



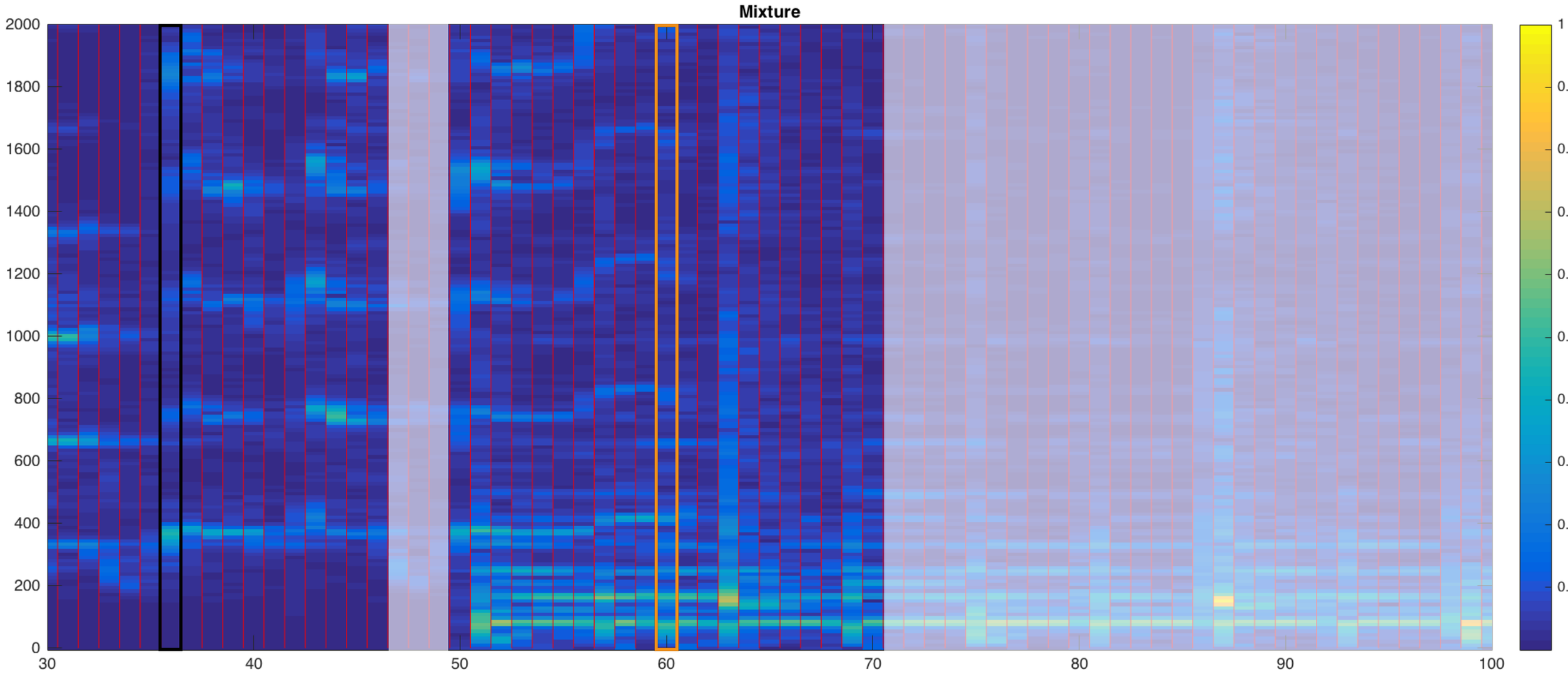
DISTANCE BETWEEN GROUP OF FRAMES

2. LOOK FOR SIMILAR TIME FRAMES

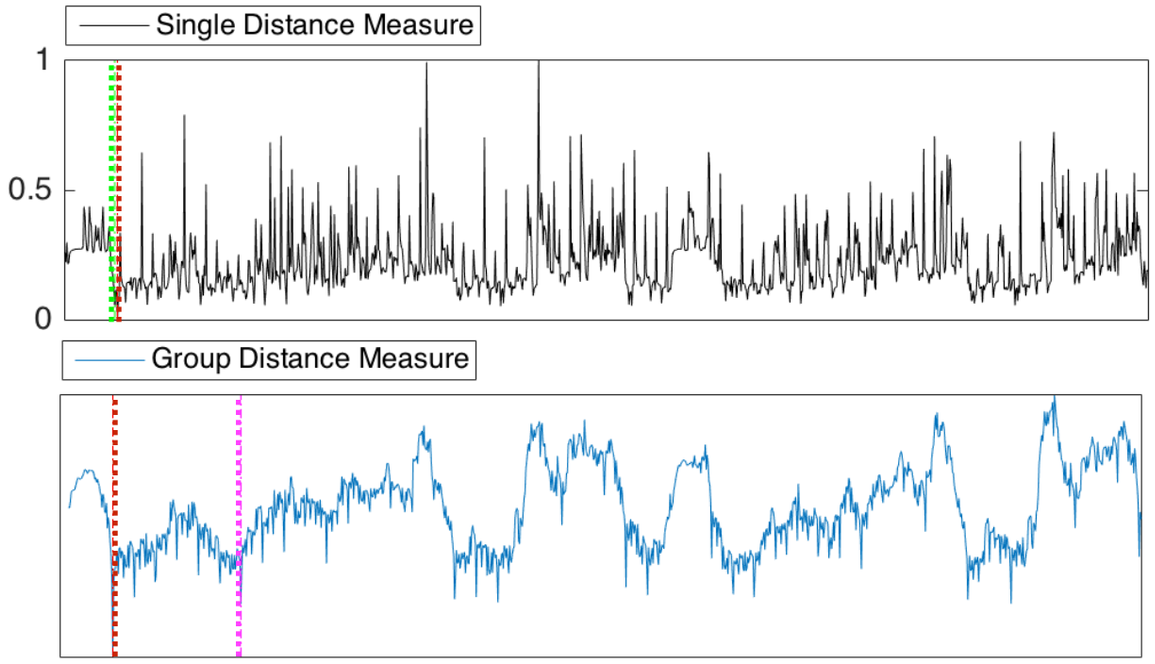


DISTANCE BETWEEN GROUP OF FRAMES

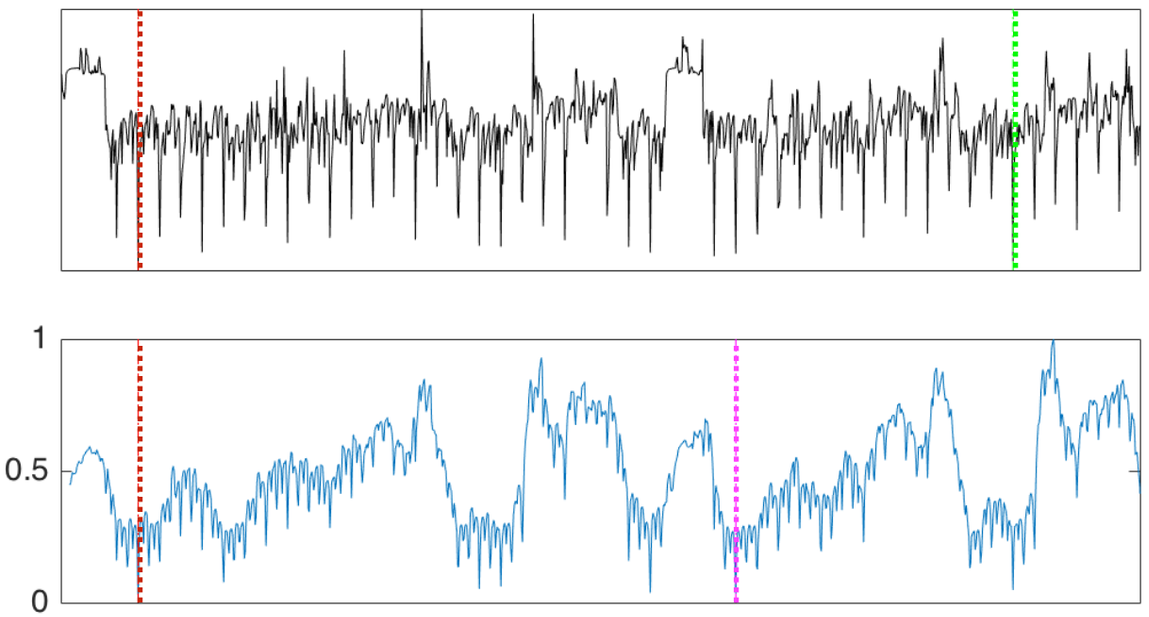
2. LOOK FOR SIMILAR TIME FRAMES



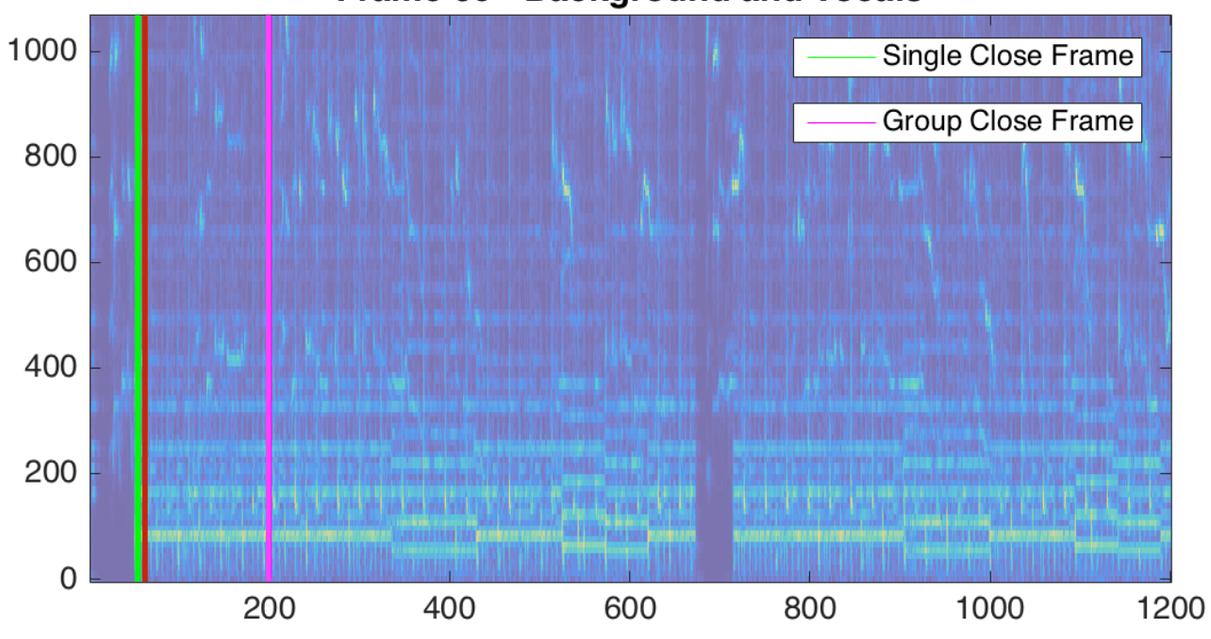
DISTANCE BETWEEN GROUP OF FRAMES

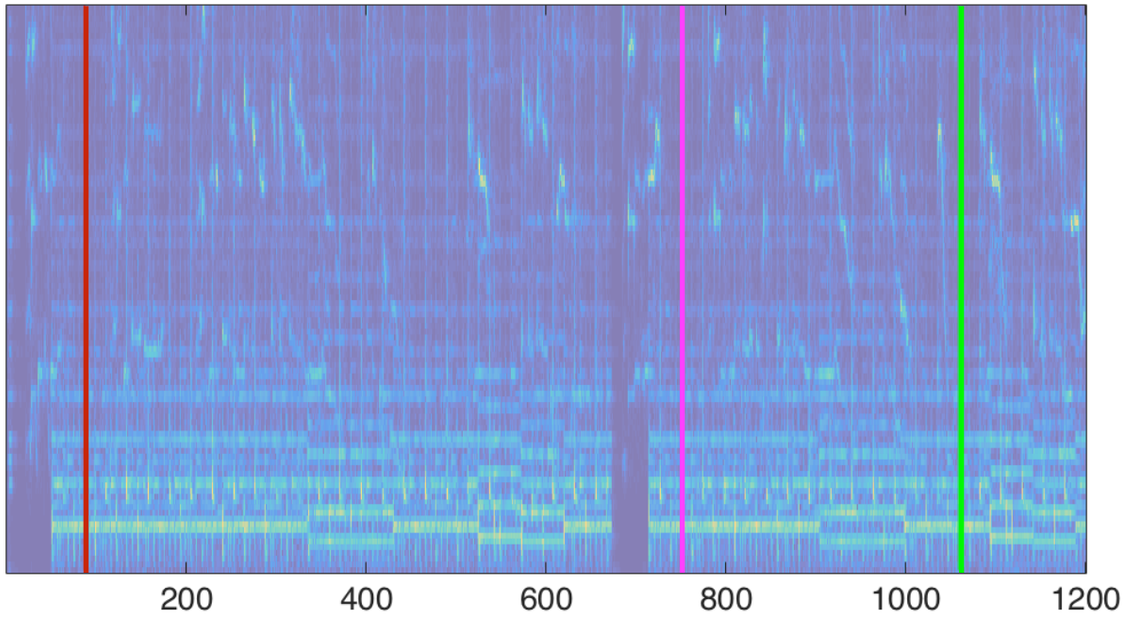


Frame 60 - Background and vocals

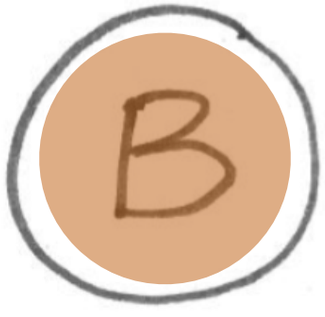


Frame 87 - Only background music





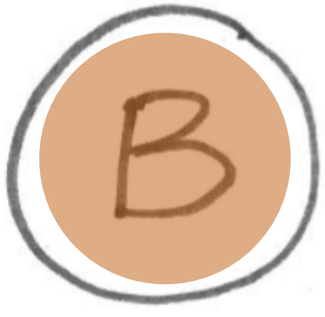
A


B

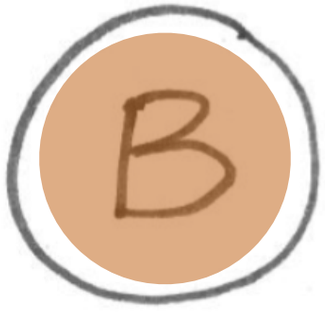

A


B

A


B

SINGLE FRAME
SIMILARITY



GROUPS OF FRAMES
SIMILARITY

EXPERIMENTS

EXPERIMENTS

Demixing Secrets Dataset 100 (DSD100)

2016 Signal Separation Evaluation Campaign (SiSEC)

EXPERIMENTS

Demixing Secrets Dataset 100 (DSD100)

2016 Signal Separation Evalvation Campaign (SiSEC)

100 SONGS

POLYPHONIC

Fs = 44.1 KHz

30s LONG

EXPERIMENTS

Demixing Secrets Dataset 100 (DSD100)

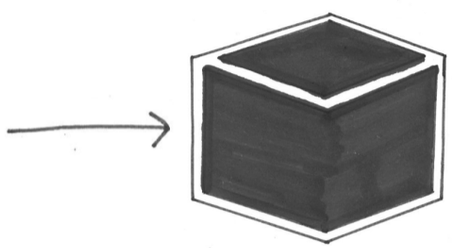
2016 Signal Separation Evalvation Campaign (SiSEC)

100 SONGS

POLYPHONIC

Fs = 44.1 KHz

30s LONG



BSS Eval toolbox 3.0

Signal to Distortion Ratio (SDR)

Source Image to Spatial Distortion Ratio (ISR)

Source to Interference Ratio (SIR)

Source to Artifacts Ratio (SAR)

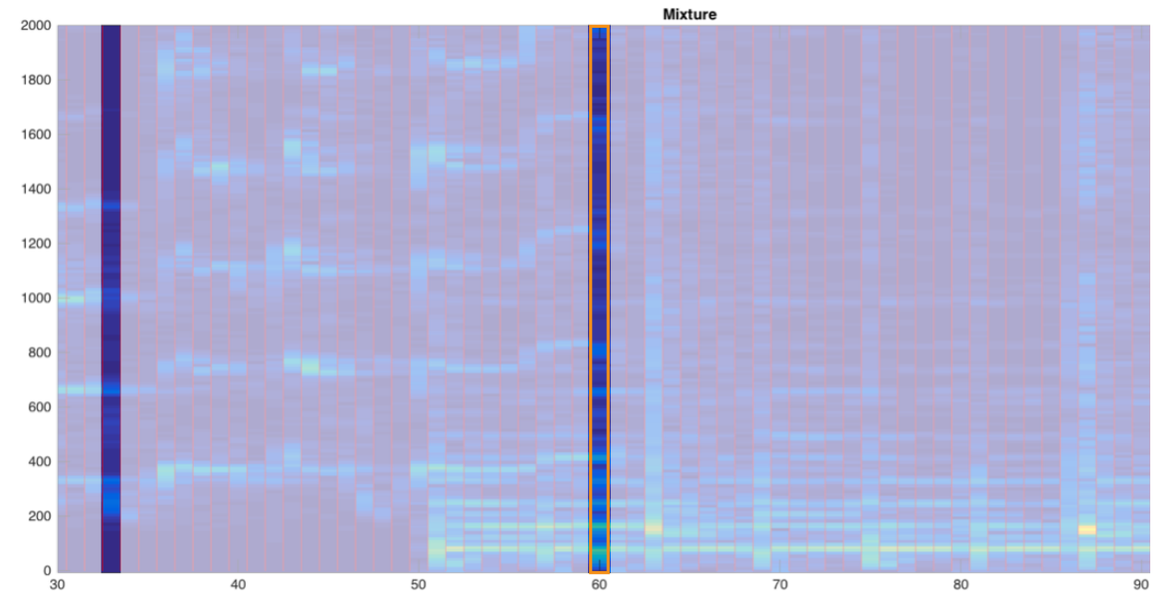
FFT size = 4096

Hopsize = 2048

BASELINE :

Instance of K.A.M

for vocal separation



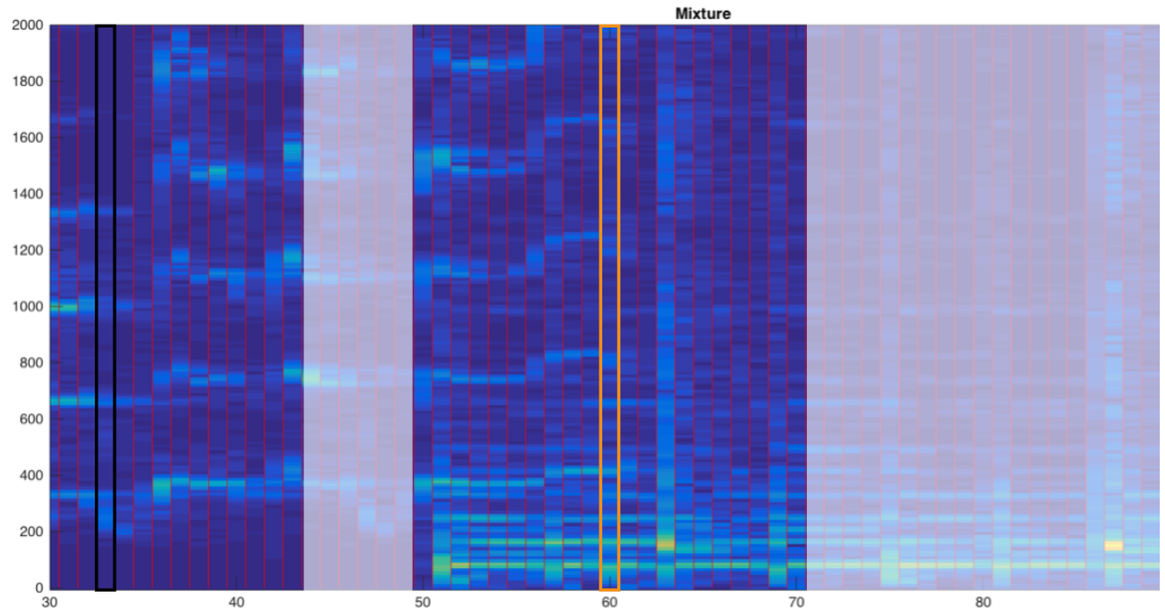
DISTANCE BETWEEN FRAMES

PROPOSED EXTENSION :

Baseline +

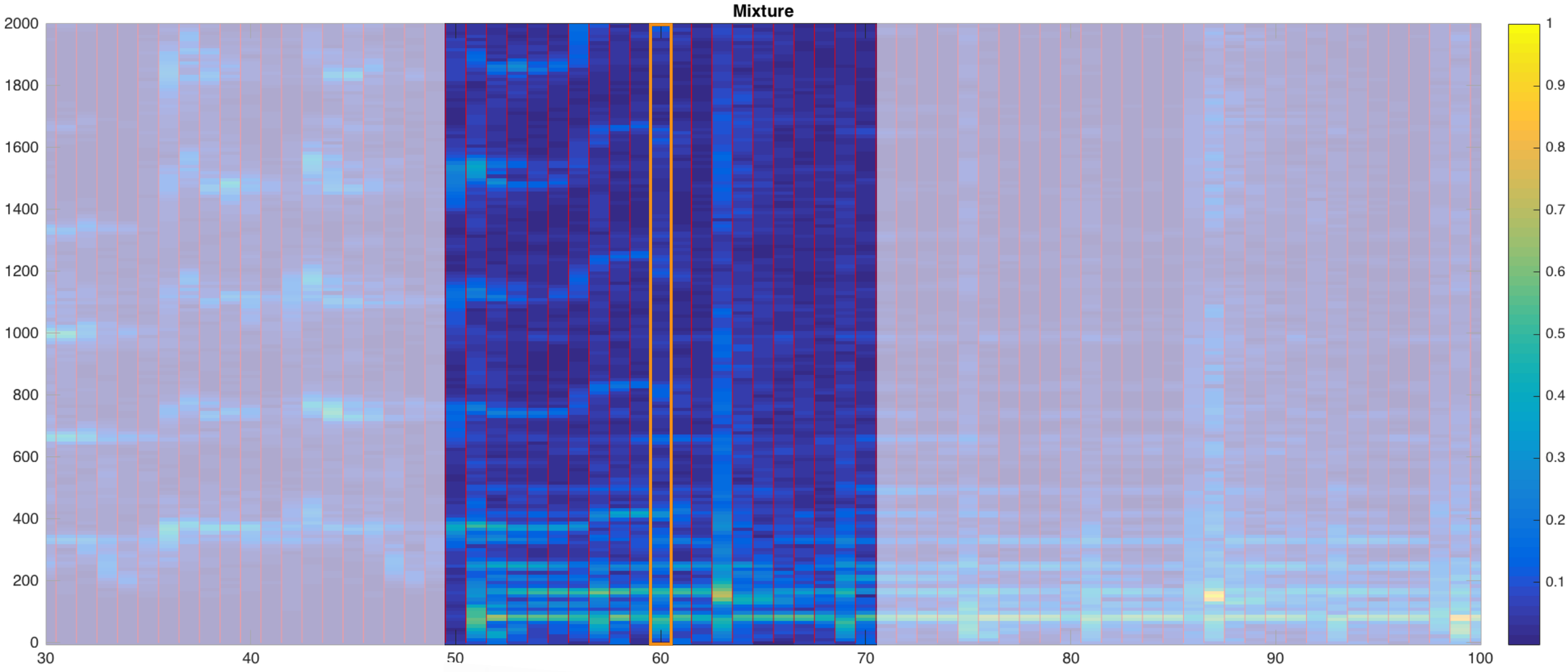
temporal context

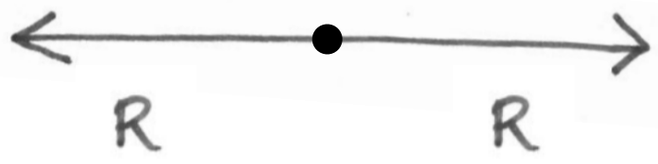
in proximity kernel



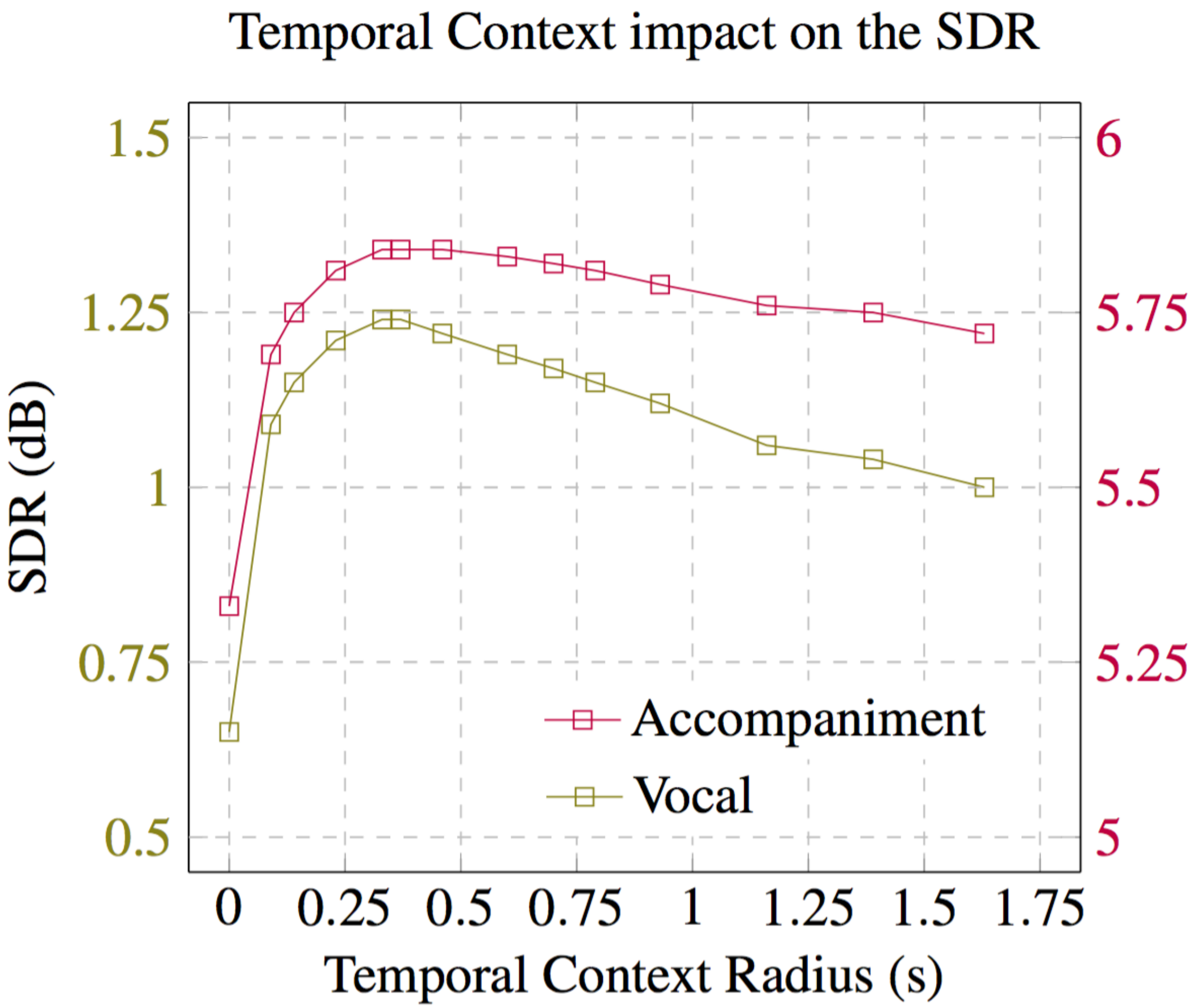
DISTANCE BETWEEN GROUP OF FRAMES

PROPOSED EXTENSION

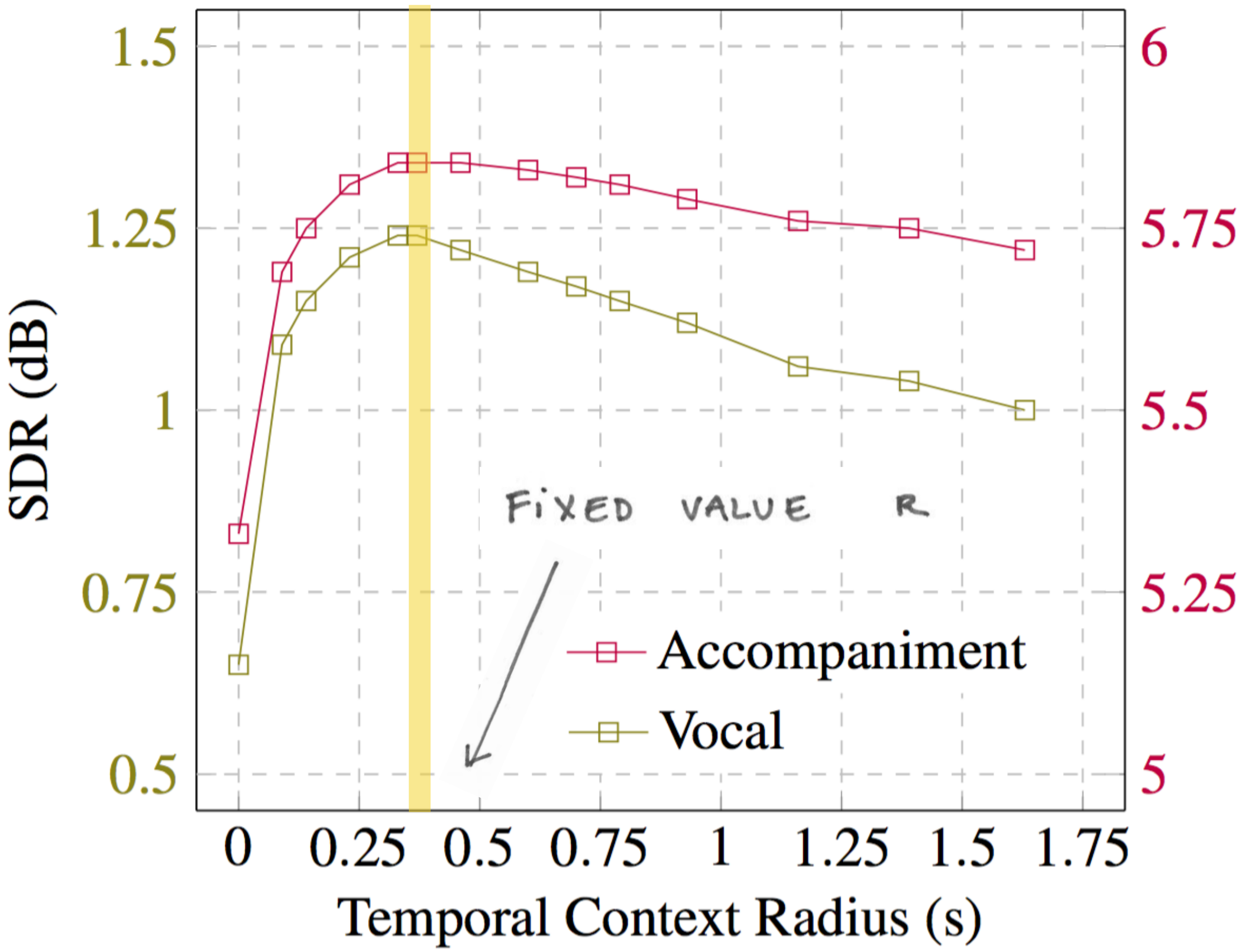


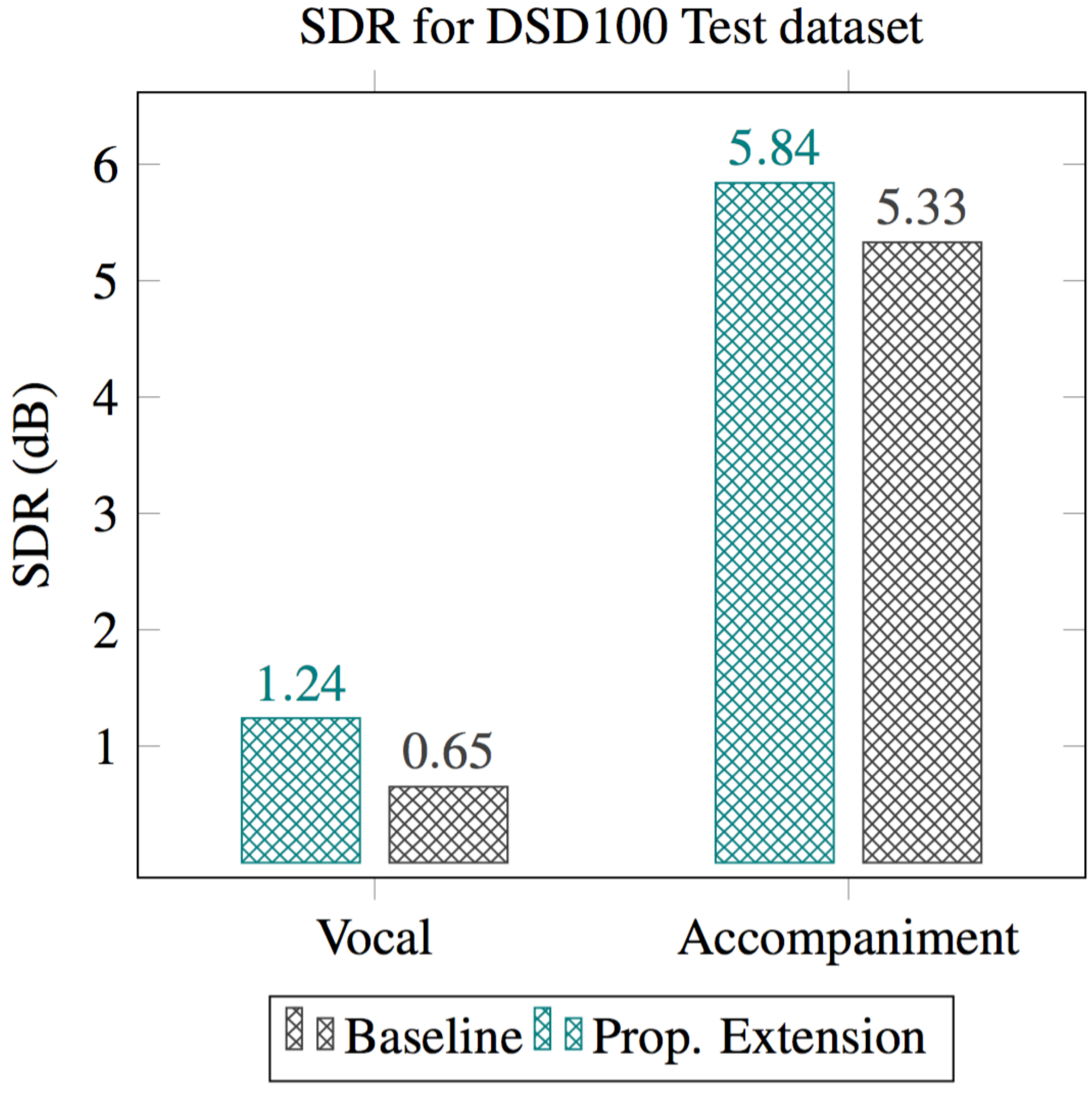


TEMPORAL CONTEXT


Temporal Context impact on the SDR
SDR (dB)
1.5
1.25
1
0.75
0.5
6
5.75
5.5
5.25
5
Accompaniment
Vocal
0
0.25
0.5
0.75
1
1.25
1.5
1.75
Temporal Context Radius (s)

Temporal Context impact on the SDR




SDR for DSD100 Test dataset
SDR (dB)
6
5
4
3
2
1
5.84
5.33
1.24
0.65
Vocal
Accompaniment
Baseline
Prop. Extension

WE PROPOSE A SIMPLE EXTENSION:

WE PROPOSE A SIMPLE EXTENSION:

INTRODUCE A TEMPORAL CONTEXT IN PROXIMITY KERNELS

WE PROPOSE A SIMPLE EXTENSION:

INTRODUCE A TEMPORAL CONTEXT IN PROXIMITY KERNELS

⊕ STABILISES THE SOURCE ESTIMATES

WE PROPOSE A SIMPLE EXTENSION:

INTRODUCE A TEMPORAL CONTEXT IN PROXIMITY KERNELS

⊕ STABILISES THE SOURCE ESTIMATES

⊕ IMPROVES (A BIT) THE SEPARATION PERFORMANCE

ANY QUESTIONS?

Delia Fano Yela Sebastian Ewert

Derry FitzGerald Mark Sandler

d.fanoyela@qmul.ac.uk