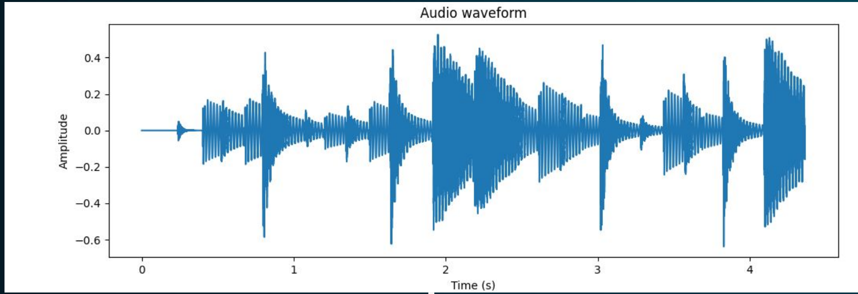
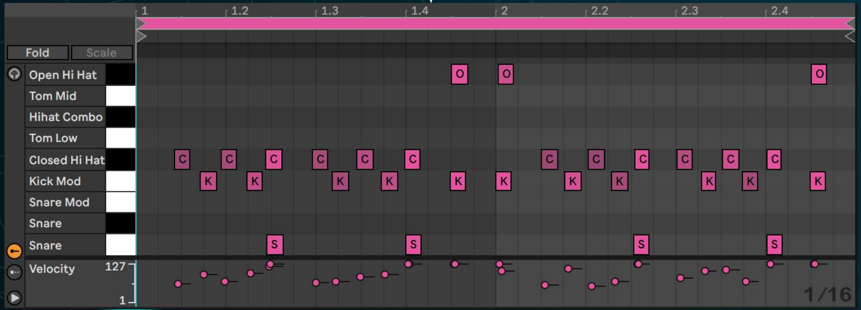
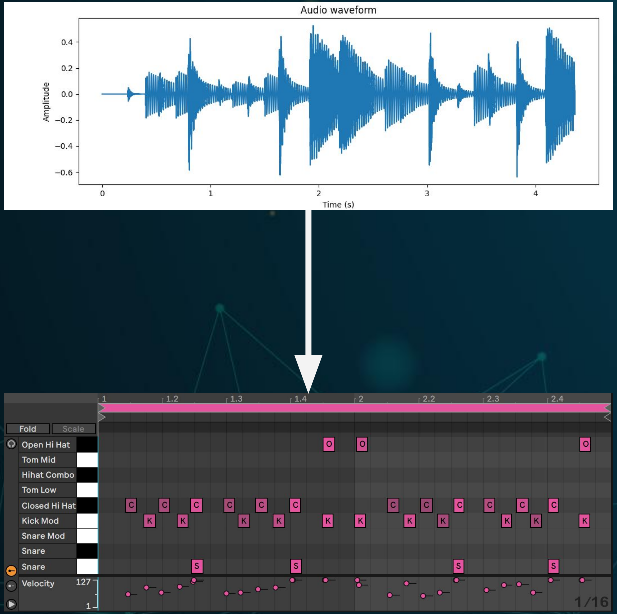
Automatic Drum Transcription Using Template-Initialized Non-Negative Matrix Factorization

Júlia Vághy

Bachelor’s project presentation


Audio waveform
Amplitude
0.4
0.2
0.0
−0.2
−0.4
−0.6
0
1
2
3
4
Time (s)


1
1.2
1.3
1.4
2
2.2
2.3
2.4
Fold
Scale
Open Hi Hat
Tom Mid
Hihat Combo
Tom Low
Closed Hi Hat
Kick Mod
Snare Mod
Snare
Snare
Velocity
127
1
1/16



Hidden Markov Models
F=0.82

Alternate Level Clustering
F=0.71

Recurrent Neural Networks
F=0.95

Non-negative matrix factorization (NMF)
F=0.89

Support Vector Machines
F=0.83

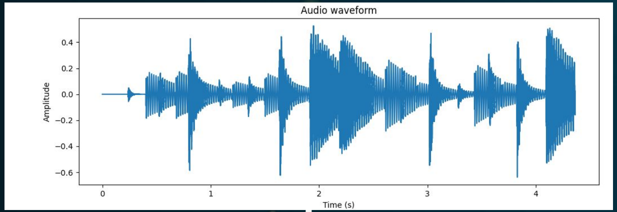
Probabilistic Latent Component Analysis
F=0.7

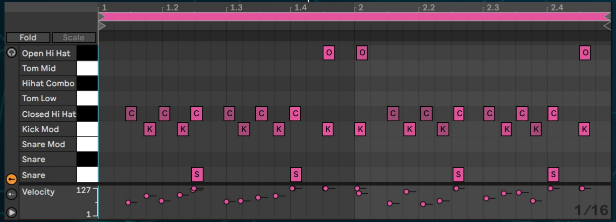
Convolutional Neural Networks
F=0.96

Problem: Datasets

- Lack of complexity
- Poor coverage of instruments and playing style
- Homogeneity of recording conditions

Wu, C. W., Dittmar, C., Southall, C., Vogl, R., Widmer, G., Hockman, J., & Lerch, A. (2018). A review of automatic drum transcription. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(9), 1457-1483.





Hidden Markov Models
F=0.82

Alternate Level Clustering
F=0.71

Recurrent Neural Networks
F=0.95

Non-negative matrix factorization (NMF)
F=0.89

Support Vector Machines
F=0.83

Probabilistic Latent Component Analysis
F=0.7

Convolutional Neural Networks
F=0.96

Research Question

NMF variants in drum transcription

→ with background noise

→ using a wide range of instruments

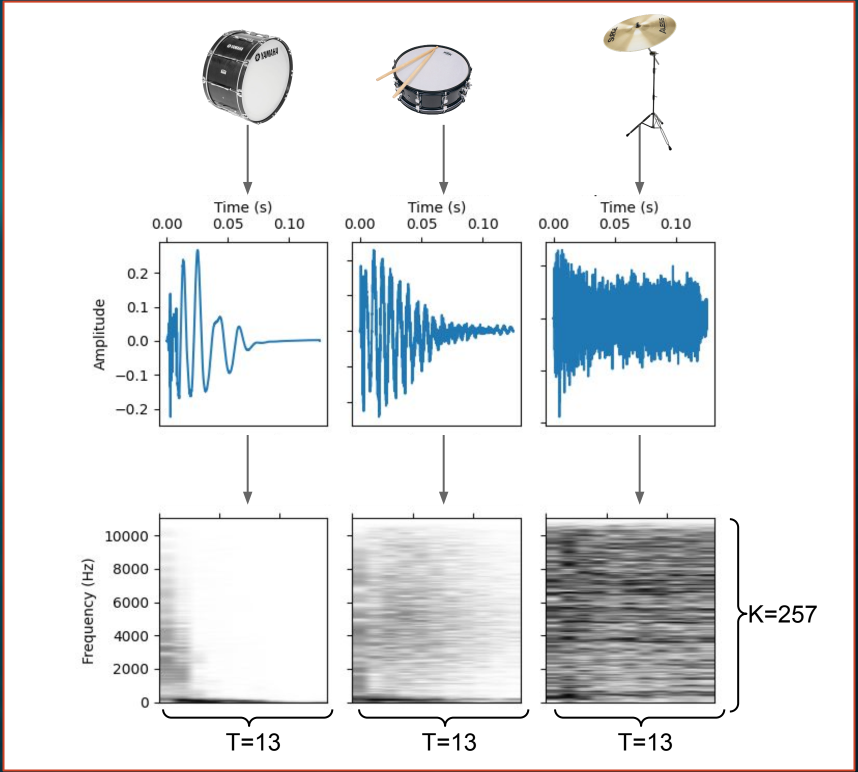
Data synthesis

20 drum loops and corresponding instrument sound samples

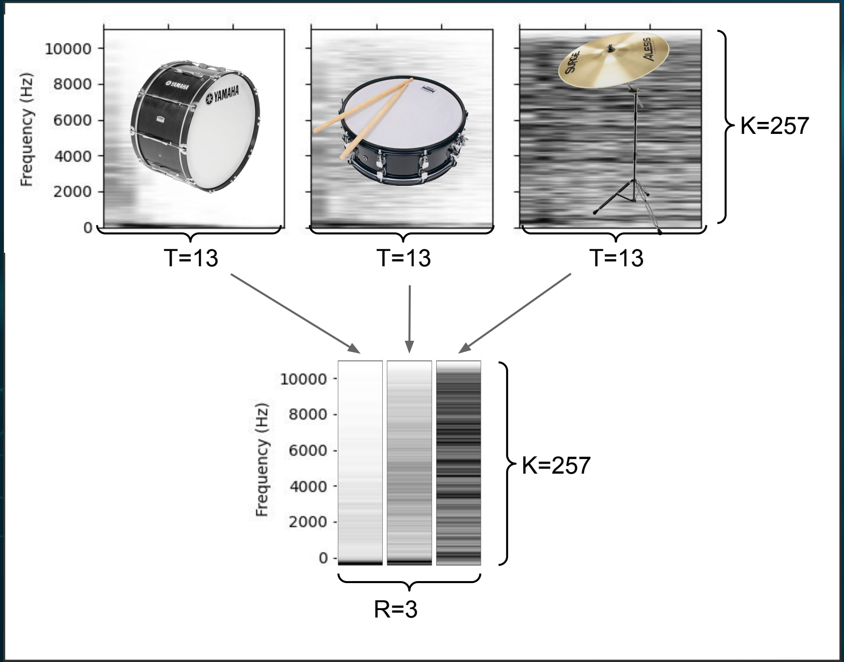
Each on four noise conditions: none, mild, loud, and extreme

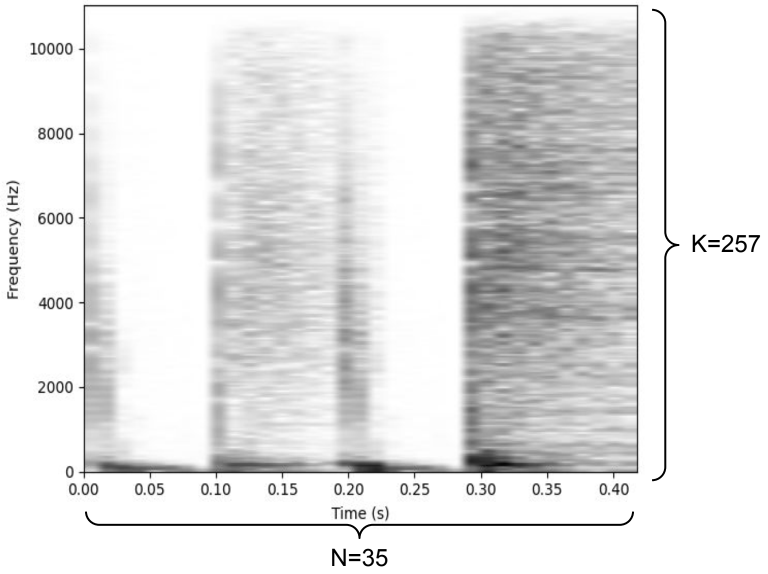
NMF

Template initialization

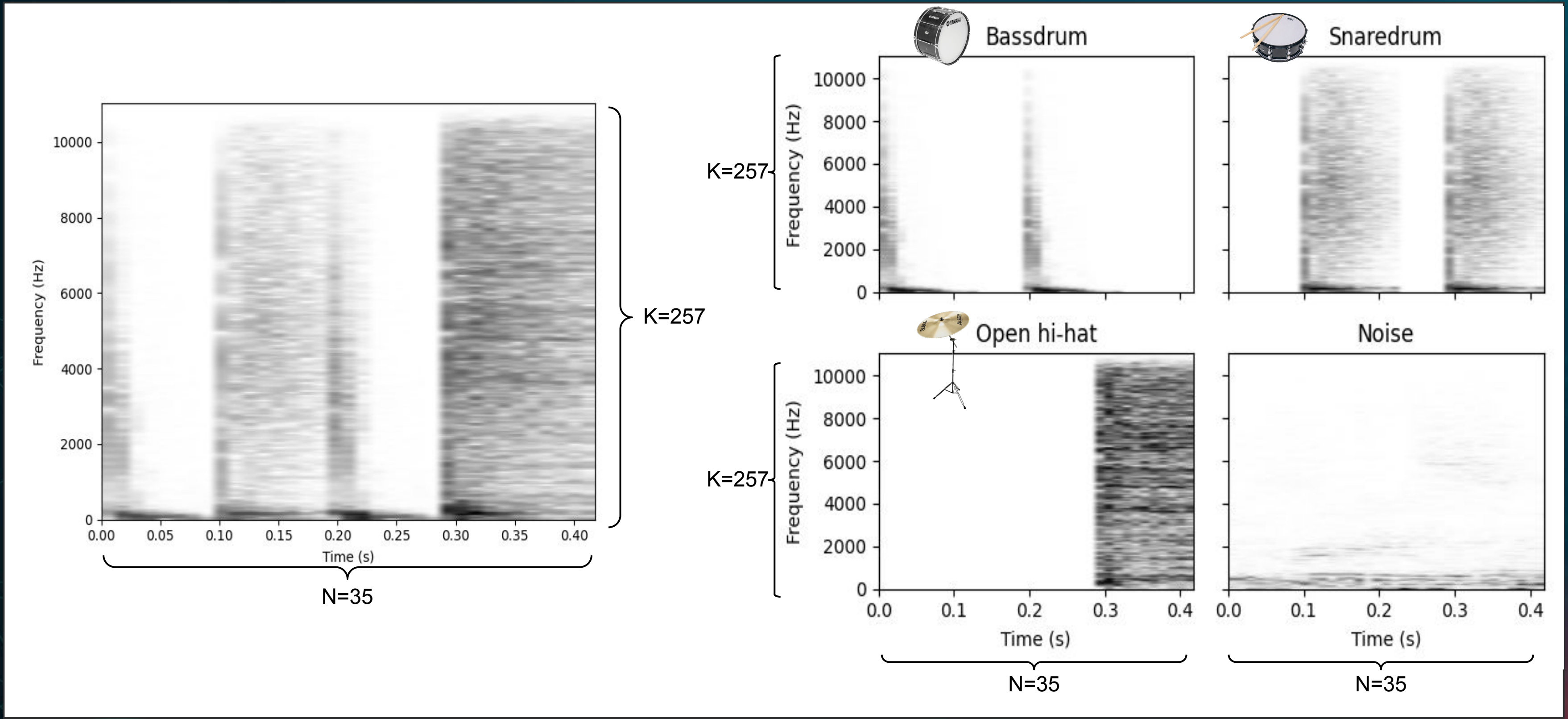


NMF Template Matrix

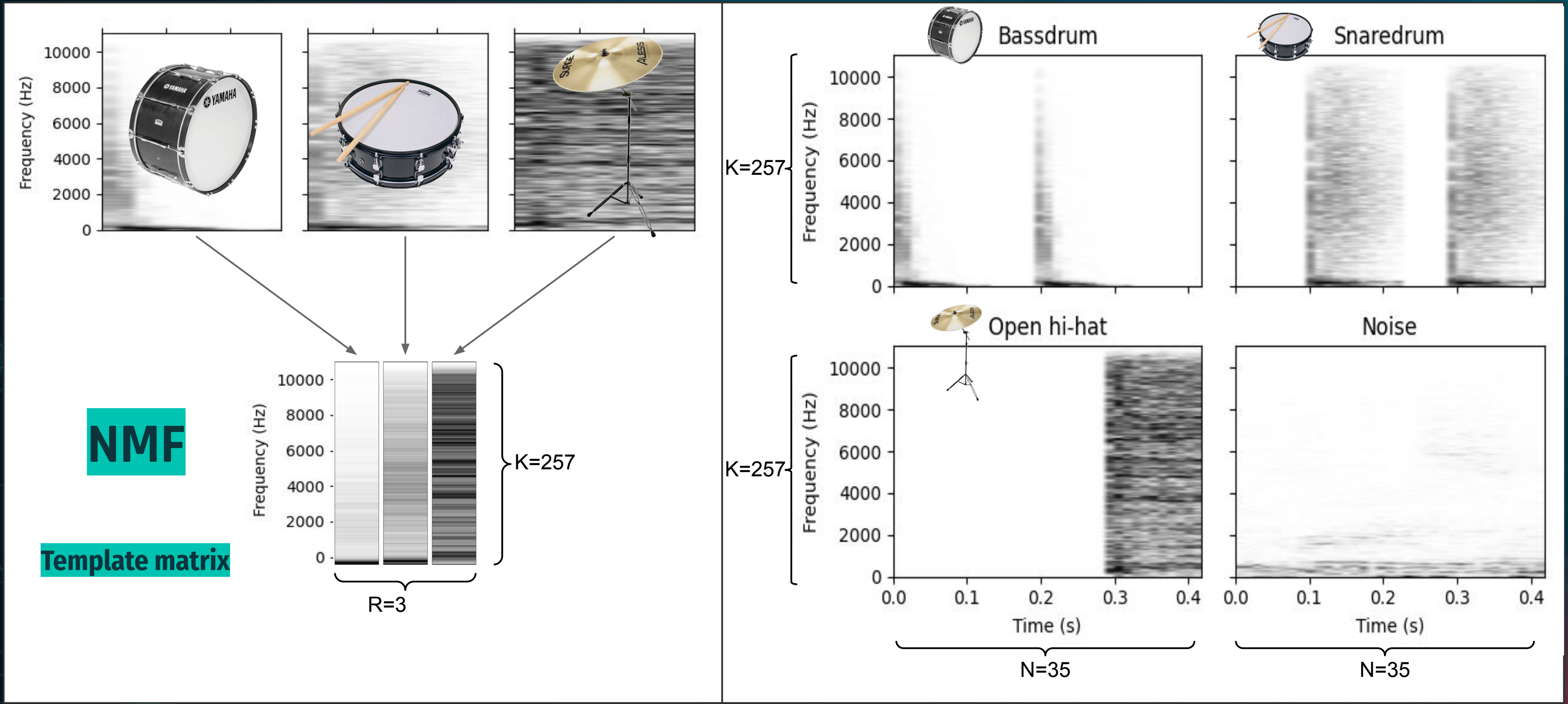




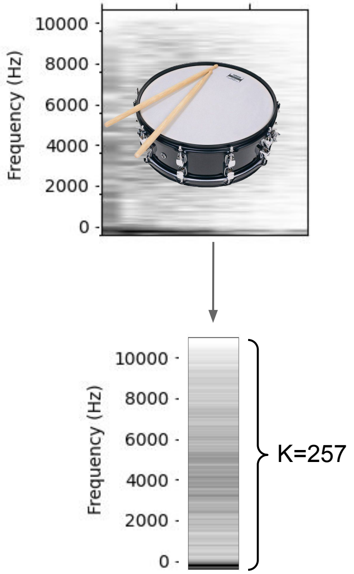
Frequency (Hz)
10000
8000
6000
4000
2000
0
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Time (s)
K=257
N=35

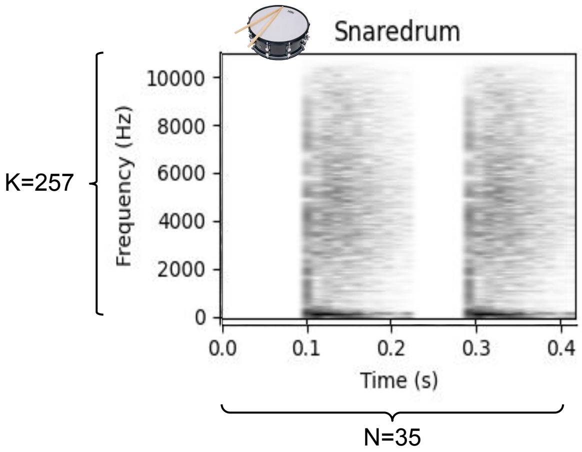


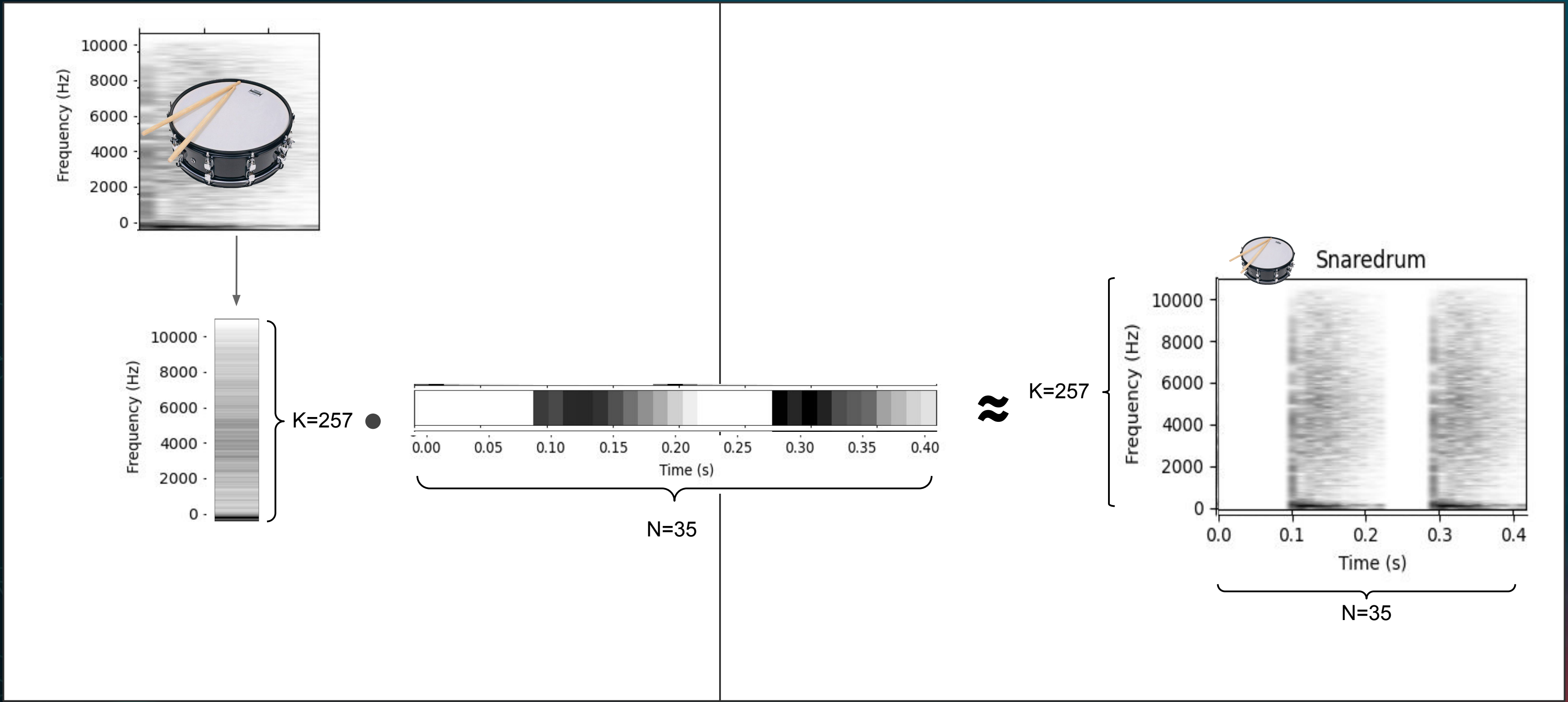
Frequency (Hz)
10000
8000
6000
4000
2000
0
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Time (s)
N=35
K=257
Bassdrum
Snaredrum
Open hi-hat
Noise
K=257
K=257
Frequency (Hz)
Frequency (Hz)
0.0
0.1
0.2
0.3
0.4
Time (s)
Time (s)
N=35
N=35



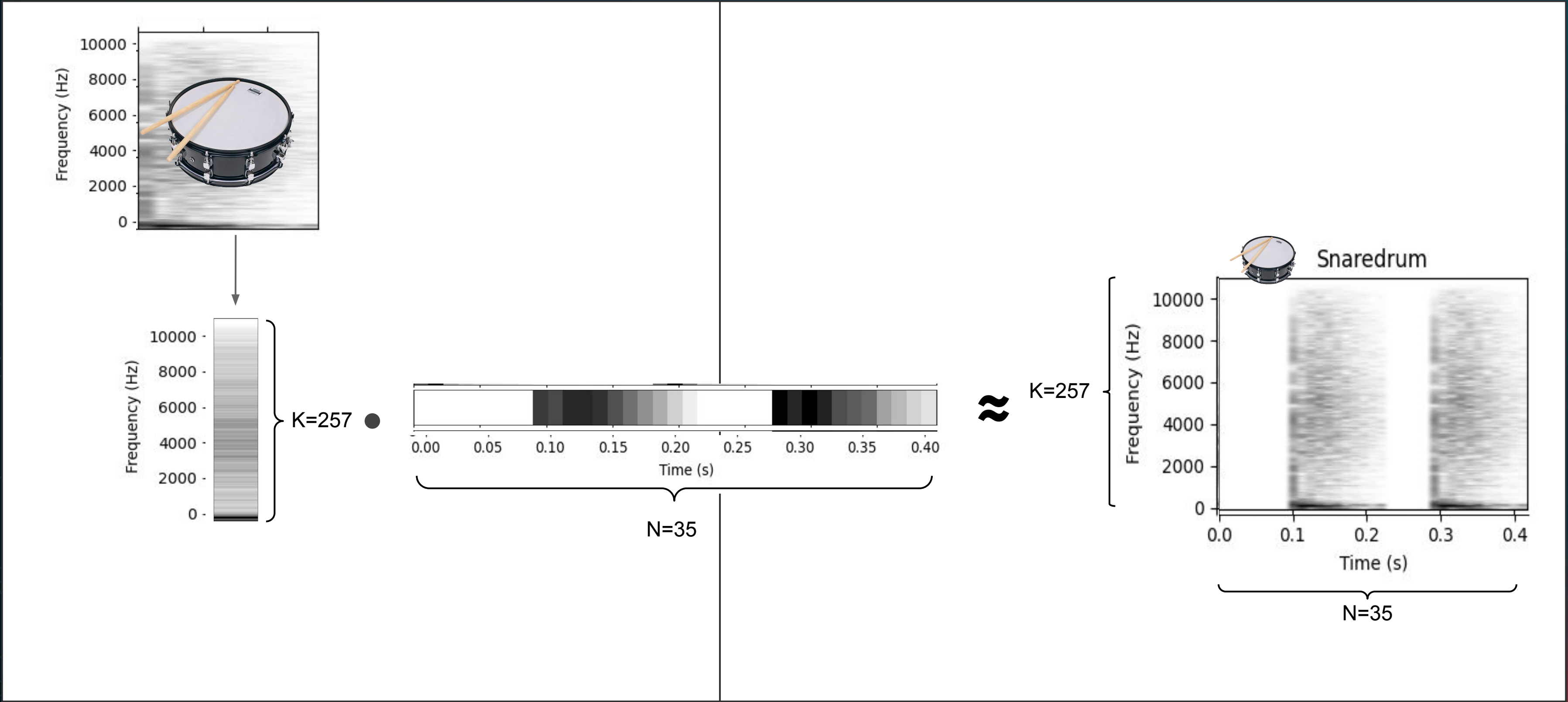
Frequency (Hz)
10000
8000
6000
4000
2000
0
NMF
Template matrix
R=3
K=257
Bassdrum
Snaredrum
Open hi-hat
Noise
Time (s)
0.0
0.1
0.2
0.3
0.4
N=35
N=35


Frequency (Hz)
10000
8000
6000
4000
2000
0
K=257

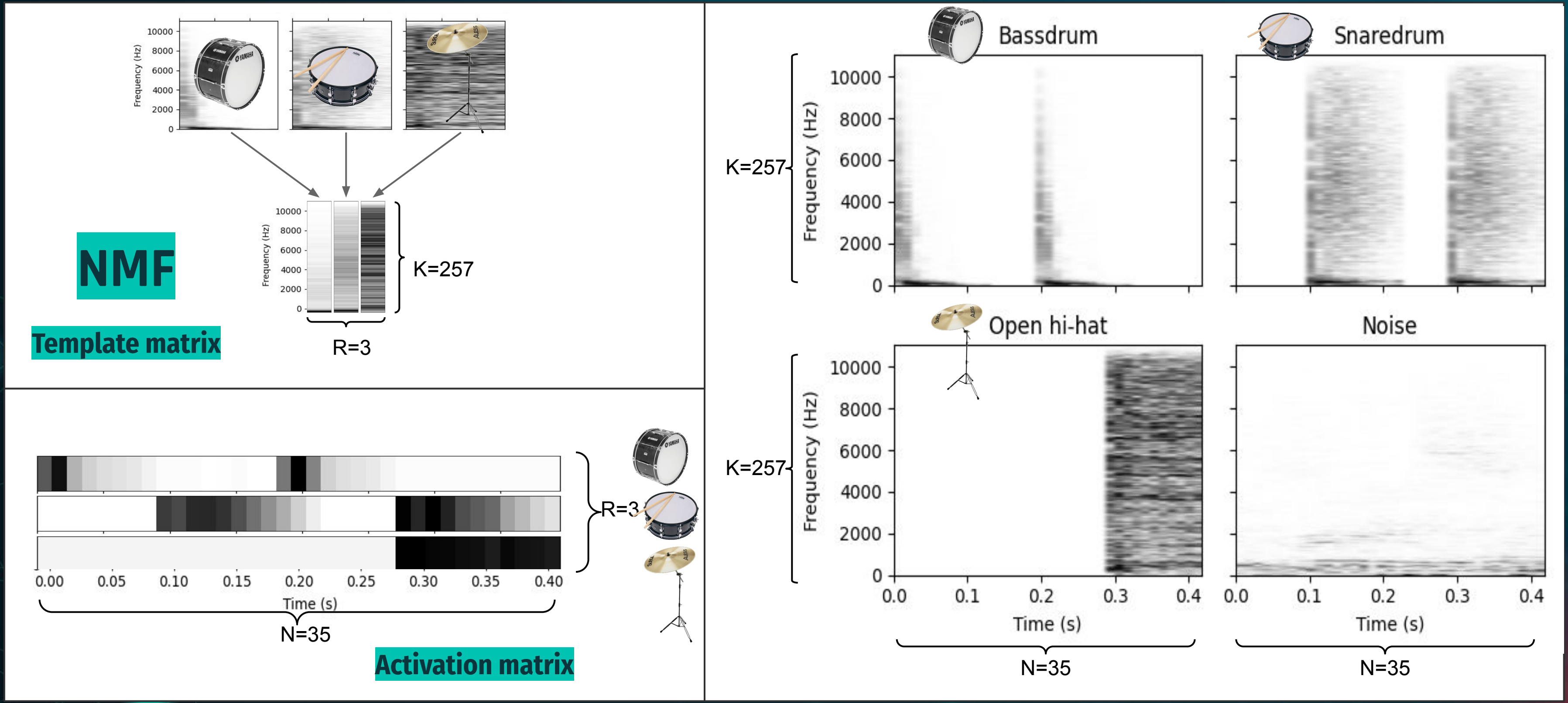

Snaredrum
K=257
Frequency (Hz)
10000
8000
6000
4000
2000
0
0.0
0.1
0.2
0.3
0.4
Time (s)
N=35



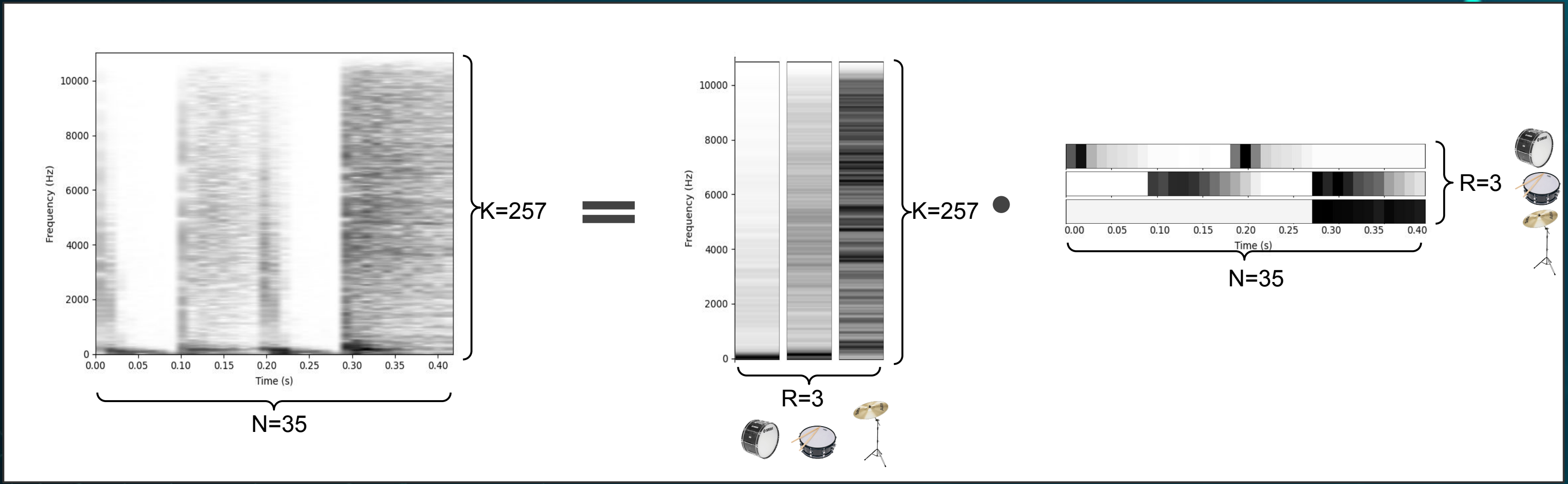
Frequency (Hz)
10000
8000
6000
4000
2000
0
K=257
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Time (s)
N=35
≈
K=257
Snaredrum
Frequency (Hz)
0.0
0.1
0.2
0.3
0.4
Time (s)
N=35



10000
8000
6000
4000
2000
0
Frequency (Hz)
K=257
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40
Time (s)
N=35
≈
K=257
Snaredrum
Frequency (Hz)
0.0 0.1 0.2 0.3 0.4
Time (s)
N=35



NMF
Template matrix
Frequency (Hz)
K=257
R=3
Activation matrix
Time (s)
N=35
Bassdrum
Snaredrum
Open hi-hat
Noise



Recording V

Template matrix W

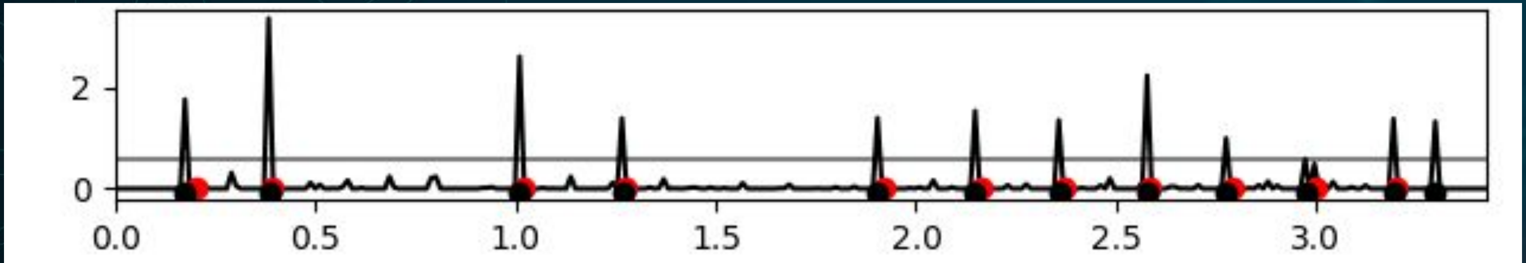
Activation matrix H

$$V = W \cdot H$$

Onset detection

H_r

Novelty function

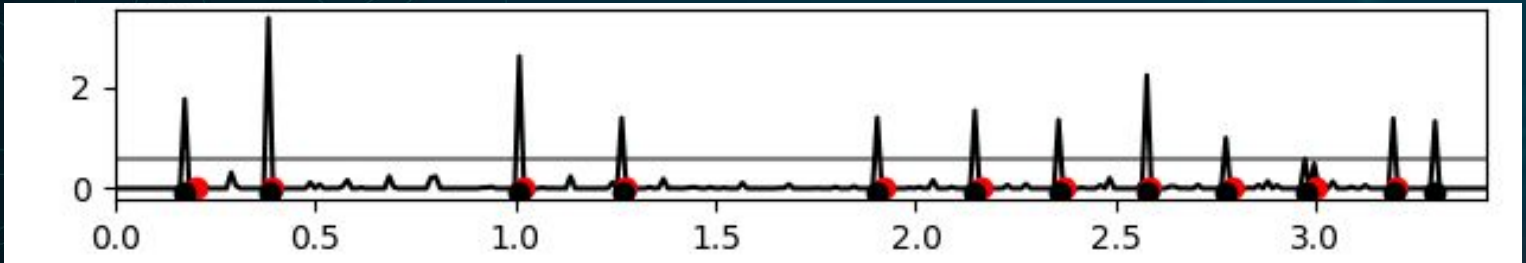


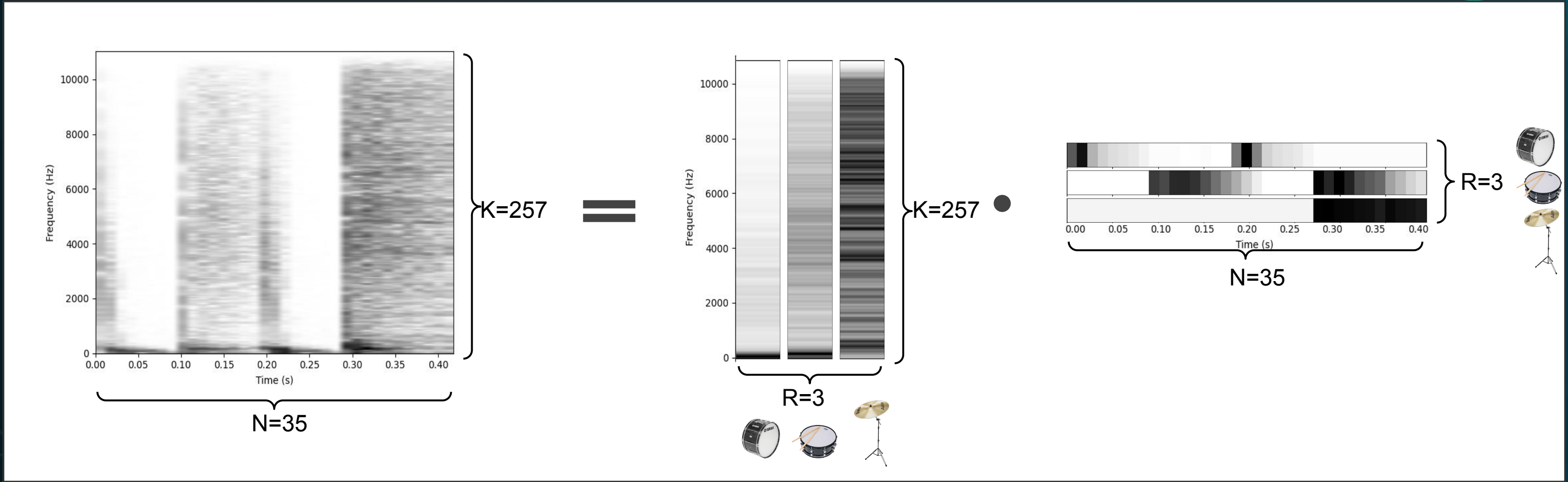
Müller Meinard. (2021). Tempo and Beat Tracking. In *Fundamentals of Music Processing: Using Python and Jupyter Notebooks* (2nd ed., pp. 311–315)., Springer.

Evaluation

$$F = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \qquad P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN}$$

Novelty function





Recording V

Template matrix W

Activation matrix H

$$V = W \cdot H$$

Gradient descent

$$\mathcal{L}^{(i)}(\mathbf{V}|\hat{\mathbf{V}}^{(i)}) = \sum \left(\mathbf{V} \odot \log\left(\frac{\mathbf{V}}{\hat{\mathbf{V}}^{(i)}}\right) - \mathbf{V} + \hat{\mathbf{V}}^{(i)} \right)$$

$$\mathbf{V} \approx \hat{\mathbf{V}} = \mathbf{W} \cdot \mathbf{H}$$

Lee, D., & Seung, H. S. (2000). Algorithms for non-negative matrix factorization. *Advances in neural information processing systems, 13*.

NMF

update rules

$$\mathbf{V} \approx \hat{\mathbf{V}} = \mathbf{W} \cdot \mathbf{H}$$

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\frac{\mathbf{V}}{\hat{\mathbf{V}}}\mathbf{H}^\top}{\mathbf{J}\mathbf{H}^\top}$$

Template matrix W

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^\top \frac{\mathbf{V}}{\hat{\mathbf{V}}}}{\mathbf{W}^\top \mathbf{J}},$$

Activation matrix H

Lee, D., & Seung, H. S. (2000). Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, *13*.

NMF Variants

Adaptive Semi-adaptive Fixed

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\frac{\mathbf{V}}{\hat{\mathbf{V}}}\mathbf{H}^T}{\mathbf{J}\mathbf{H}^T}$$

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^T\frac{\mathbf{V}}{\hat{\mathbf{V}}}}{\mathbf{W}^T\mathbf{J}},$$

Yang, Z., Zhang, H., Yuan, Z., & Oja, E. (2011, June). Kullback-Leibler divergence for nonnegative matrix factorization. In *International Conference on Artificial Neural Networks* (pp. 250-257). Springer, Berlin, Heidelberg.

Adaptive

Semi-adaptive

Fixed

$$\mathbf{W} \leftarrow (1-\alpha) \cdot \mathbf{W}^{(0)} + \alpha \cdot \mathbf{W} \odot \frac{\frac{\mathbf{V}}{\hat{\mathbf{V}}}\mathbf{H}^{\mathsf{T}}}{\mathbf{J}\mathbf{H}^{\mathsf{T}}} \qquad \alpha = \left(\frac{k}{K}\right)^{\beta}$$

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^{\mathsf{T}}\frac{\mathbf{V}}{\hat{\mathbf{V}}}}{\mathbf{W}^{\mathsf{T}}\mathbf{J}},$$

Dittmar, C., & Gärtner, D. (2014, September). Real-Time Transcription and Separation of Drum Recordings Based on NMF Decomposition. In *DAFx* (pp. 187-194).

Adaptive

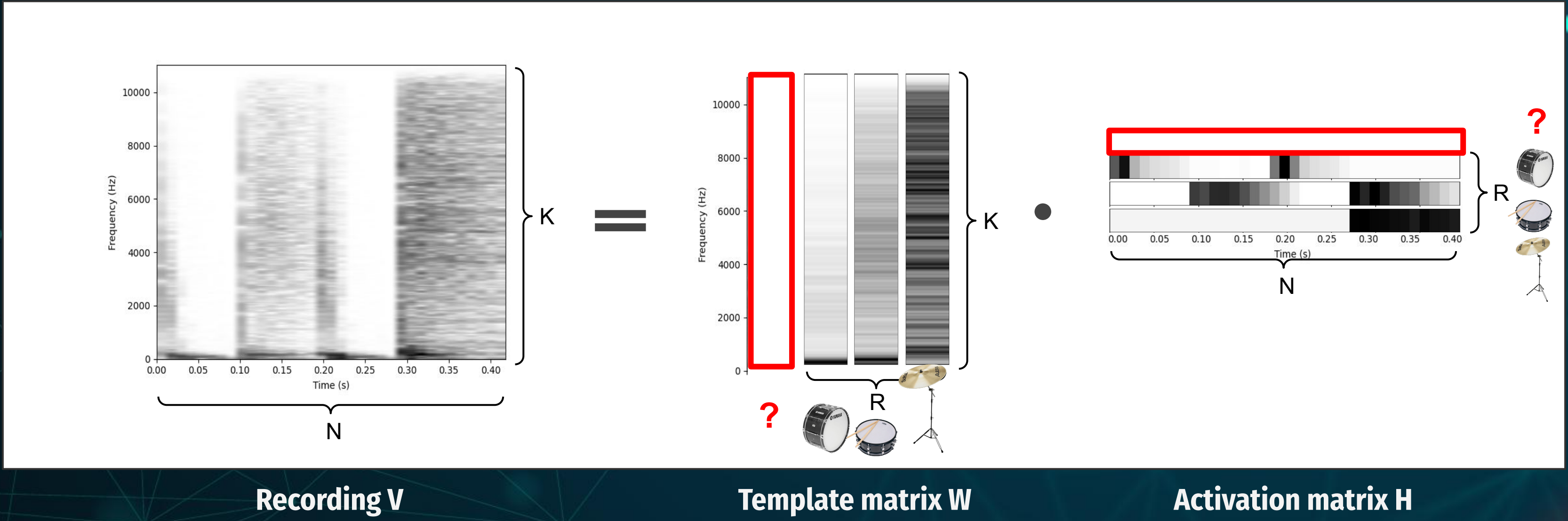
Semi-adaptive

Fixed

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^{\mathsf{T}} \frac{\mathbf{V}}{\hat{\mathbf{V}}}}{\mathbf{W}^{\mathsf{T}} \mathbf{J}},$$

Wu, C. W., & Lerch, A. (2015, August). Drum transcription using partially fixed non-negative matrix factorization. In *2015 23rd European Signal Processing Conference (EUSIPCO)* (pp. 1281-1285). IEEE.

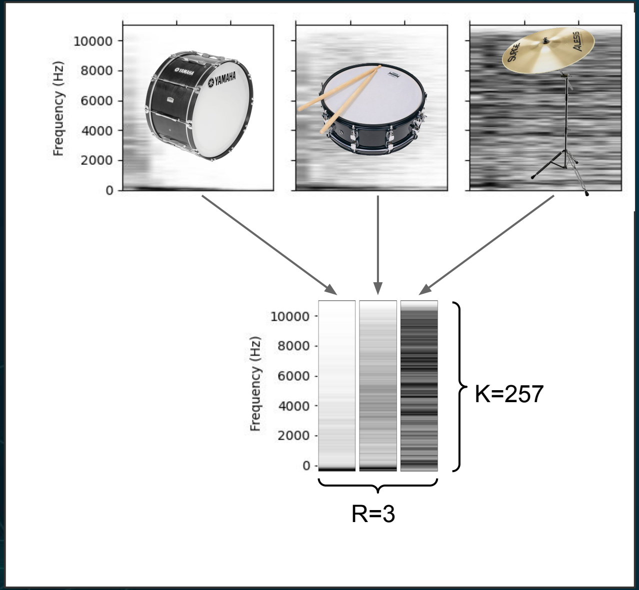
Added component(s)



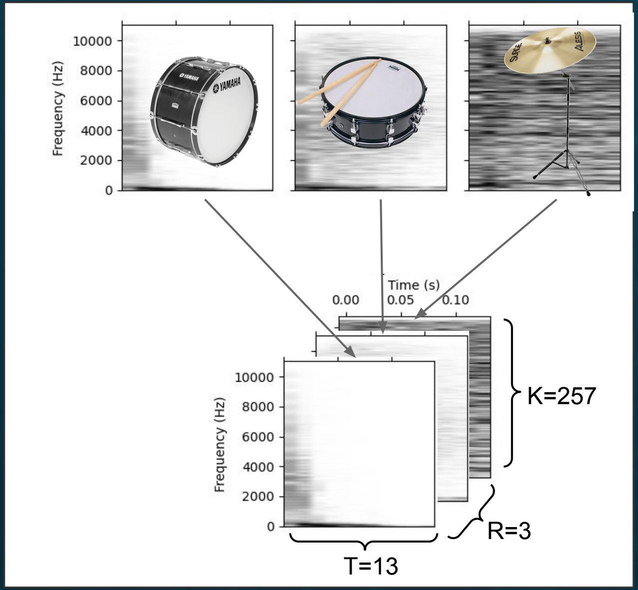
Wu, C. W., & Lerch, A. (2015, August). Drum transcription using partially fixed non-negative matrix factorization. In *2015 23rd European Signal Processing Conference (EUSIPCO)* (pp. 1281-1285). IEEE.

NMFD

Template matrix

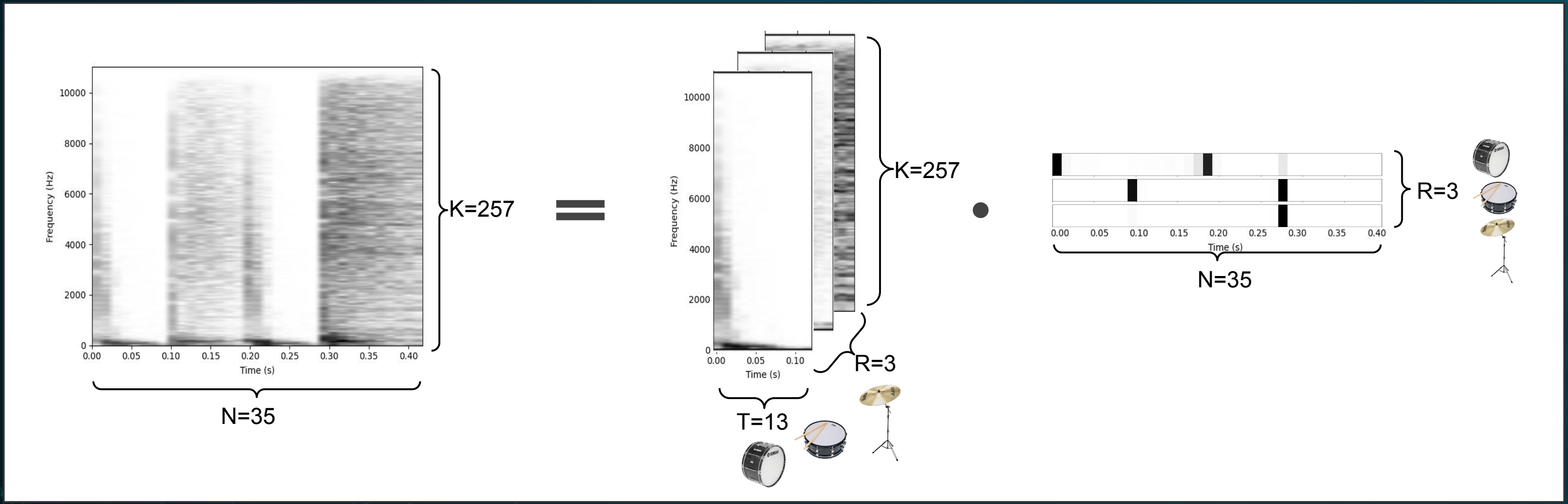


NMF



NMFD

Smaragdis, P. (2004, September). Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In *International Conference on Independent Component Analysis and Signal Separation* (pp. 494-499). Springer, Berlin, Heidelberg.



Recording V **Pattern tensor P** **Activation matrix H**

$$\mathbf{V} = \sum_{t=0}^{T-1} \mathbf{P}(,,t) \cdot \overset{t\rightarrow}{\mathbf{H}}, \quad \text{for } \forall t \in [1:T]$$

Smaragdis, P. (2004, September). Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In *International Conference on Independent Component Analysis and Signal Separation* (pp. 494-499). Springer, Berlin, Heidelberg.

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix}, \overset{0\rightarrow}{\mathbf{A}} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix}, \overset{1\rightarrow}{\mathbf{A}} = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 0 & 5 & 6 & 7 \end{bmatrix}, \overset{2\rightarrow}{\mathbf{A}} = \begin{bmatrix} 0 & 0 & 1 & 2 \\ 0 & 0 & 5 & 6 \end{bmatrix}$$

$$\mathbf{V} = \sum_{t=0}^{T-1} \mathbf{P}(,,t) \cdot \overset{t\rightarrow}{\mathbf{H}}, \quad \text{for } \forall t \in [1:T]$$

NMFD

update rules

$$\mathbf{V} \approx \hat{\mathbf{V}} = \sum_{t=0}^{T-1} \mathbf{P}(,,t) \cdot \overset{t\rightarrow}{\mathbf{H}}, \quad \text{for } \forall t \in [1:T]$$

$$\mathbf{P}(,,t) \leftarrow \mathbf{P}(,,t) \odot \frac{\frac{\mathbf{V}}{\hat{\mathbf{V}}} \cdot \left(\overset{t\rightarrow}{\mathbf{H}}\right)^{\mathsf{T}}}{\mathbf{J} \cdot \left(\overset{t\rightarrow}{\mathbf{H}}\right)^{\mathsf{T}}}$$

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{P}(,,t)^{\mathsf{T}} \cdot \overset{\leftarrow t}{\left[\frac{\mathbf{V}}{\hat{\mathbf{V}}}\right]}}{\mathbf{P}(,,t)^{\mathsf{T}} \cdot \mathbf{J}}$$

$$\text{for } \forall t \in [1:T]$$

Smaragdis, P. (2004, September). Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In *International Conference on Independent Component Analysis and Signal Separation* (pp. 494-499). Springer, Berlin, Heidelberg.

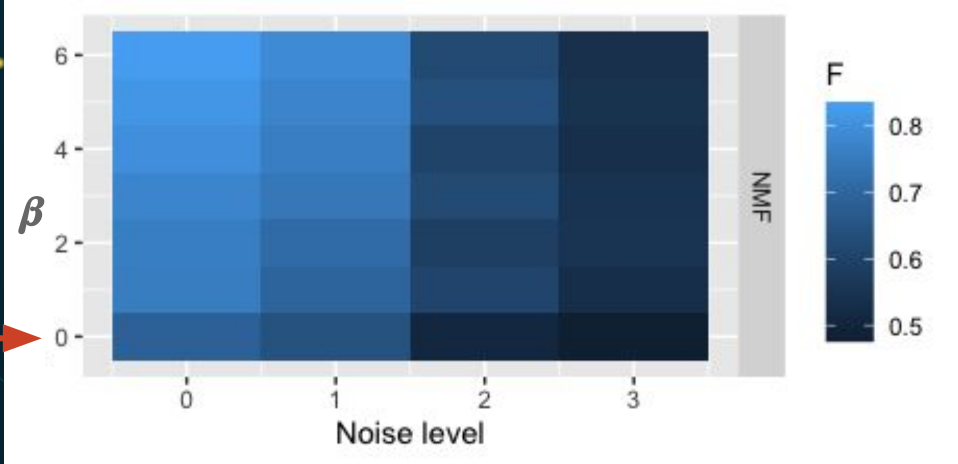
NMFD Variants

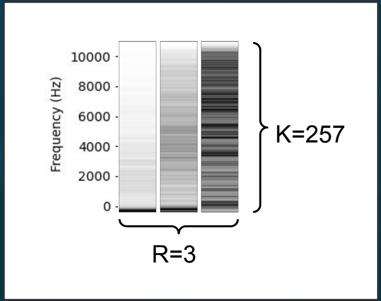
Results

(Semi-)adaptive templates - NMF

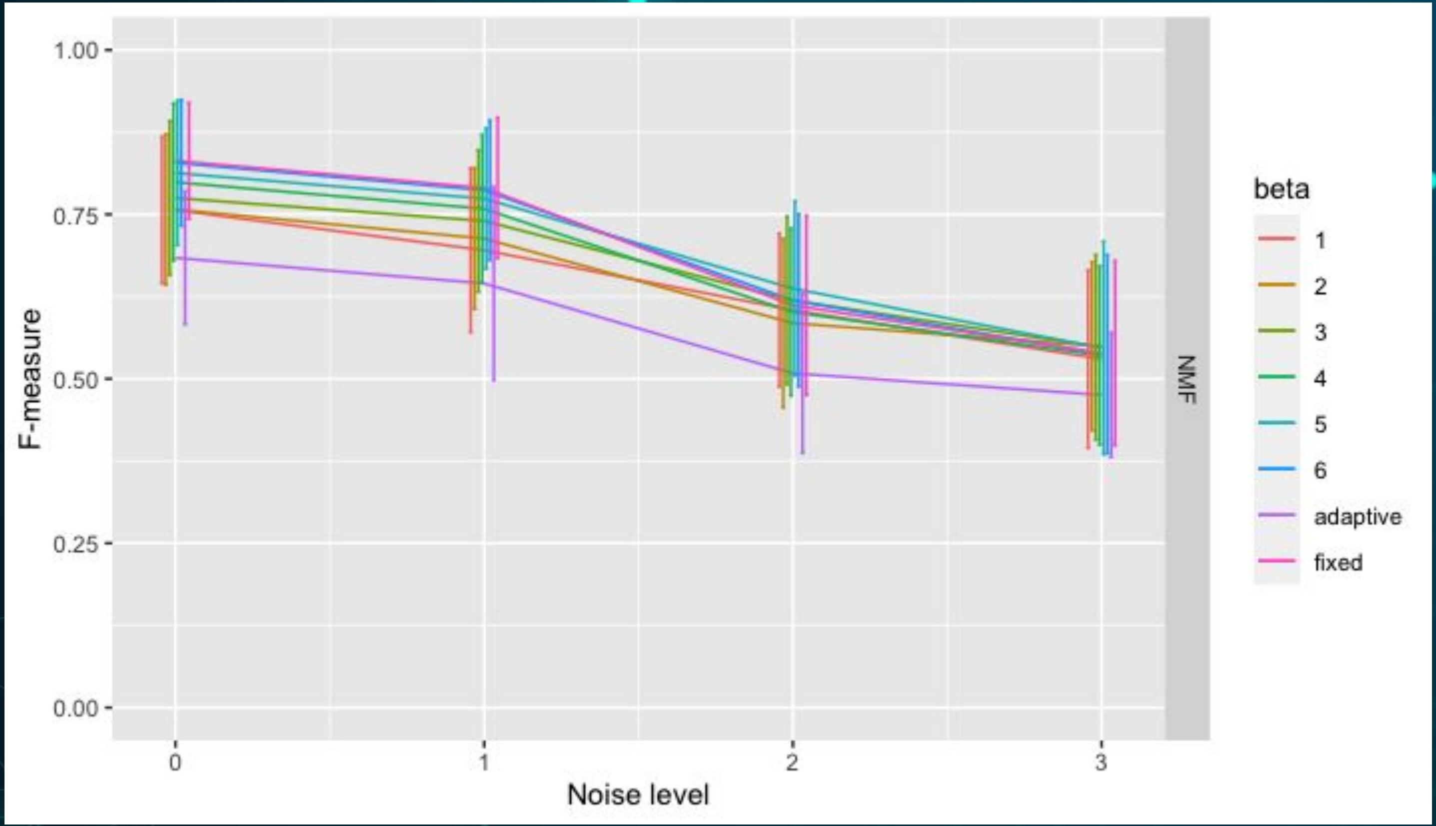
$\beta \to \infty$ - fixed

$\beta = 0$ - adaptive

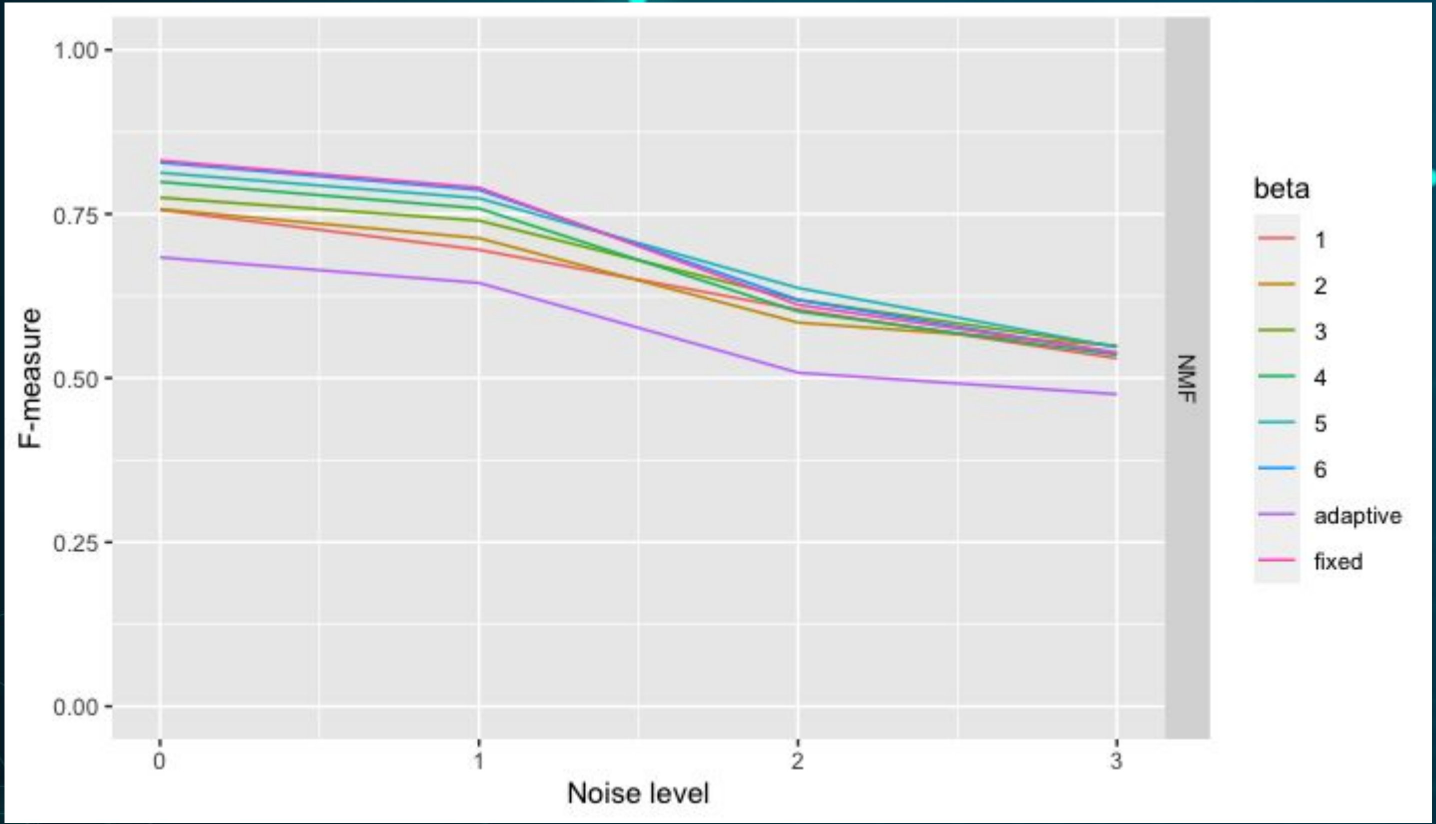




$$\mathbf{W} \leftarrow (1-\alpha) \cdot \mathbf{W}^{(0)} + \alpha \cdot \mathbf{W} \odot \frac{\frac{\mathbf{V}}{\hat{\mathbf{V}}}\mathbf{H}^{\mathsf{T}}}{\mathbf{J}\mathbf{H}^{\mathsf{T}}} \qquad \alpha = \left(\frac{k}{K}\right)^{\beta}$$



$$\mathbf{W} \leftarrow (1-\alpha)\cdot \mathbf{W}^{(0)} + \alpha \cdot \mathbf{W} \odot \frac{\frac{\mathbf{V}}{\hat{\mathbf{V}}}\mathbf{H}^{\mathsf{T}}}{\mathbf{J}\mathbf{H}^{\mathsf{T}}} \qquad \alpha = \left(\frac{k}{K}\right)^{\beta}$$

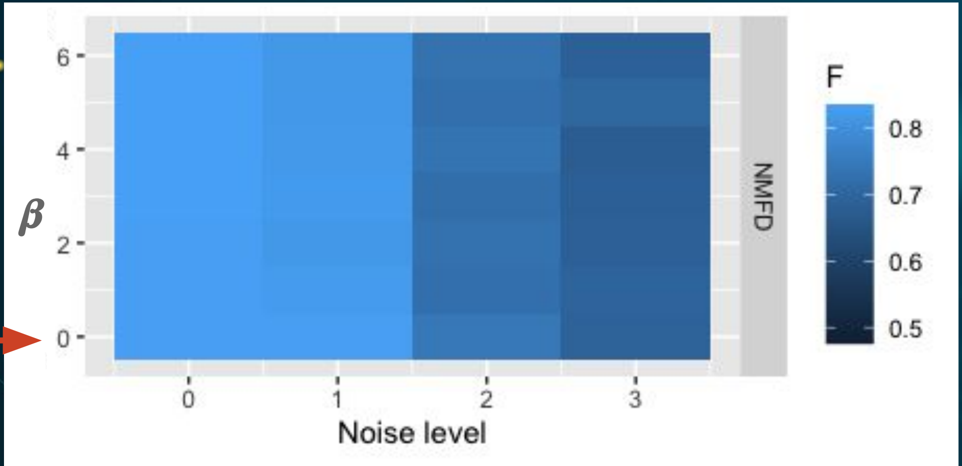


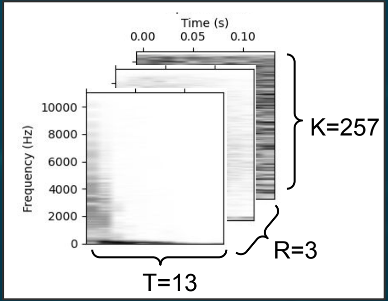
$$\mathbf{W} \leftarrow (1-\alpha) \cdot \mathbf{W}^{(0)} + \alpha \cdot \mathbf{W} \odot \frac{\frac{\mathbf{V}}{\hat{\mathbf{V}}}\mathbf{H}^{\mathsf{T}}}{\mathbf{J}\mathbf{H}^{\mathsf{T}}} \qquad \alpha = \left(\frac{k}{K}\right)^{\beta}$$

(Semi-)adaptive templates - NMFD

$\beta \rightarrow \infty$ - fixed

β=0 - adaptive



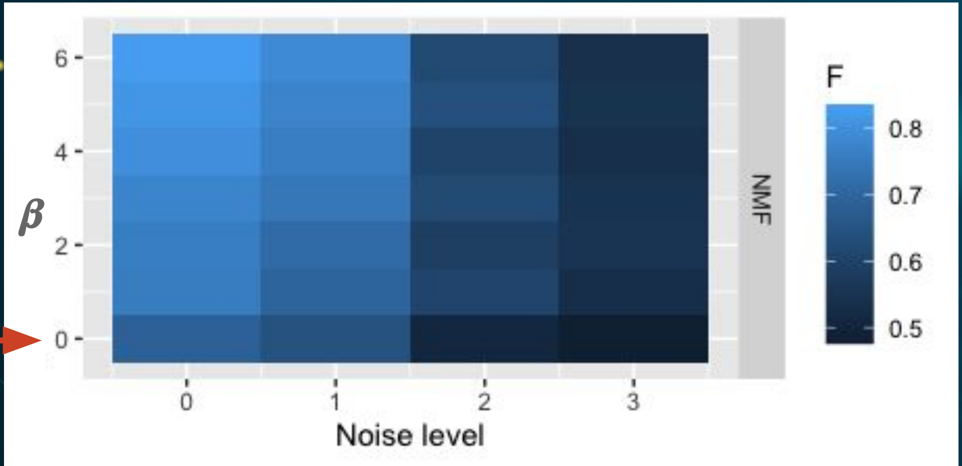


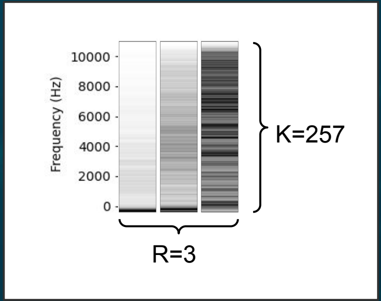
$$\mathbf{W} \leftarrow (1-\alpha) \cdot \mathbf{W}^{(0)} + \alpha \cdot \mathbf{W} \odot \frac{\frac{\mathbf{V}}{\hat{\mathbf{V}}}\mathbf{H}^{\mathsf{T}}}{\mathbf{J}\mathbf{H}^{\mathsf{T}}} \qquad \alpha = \left(\frac{k}{K}\right)^{\beta}$$

(Semi-)adaptive templates - NMF

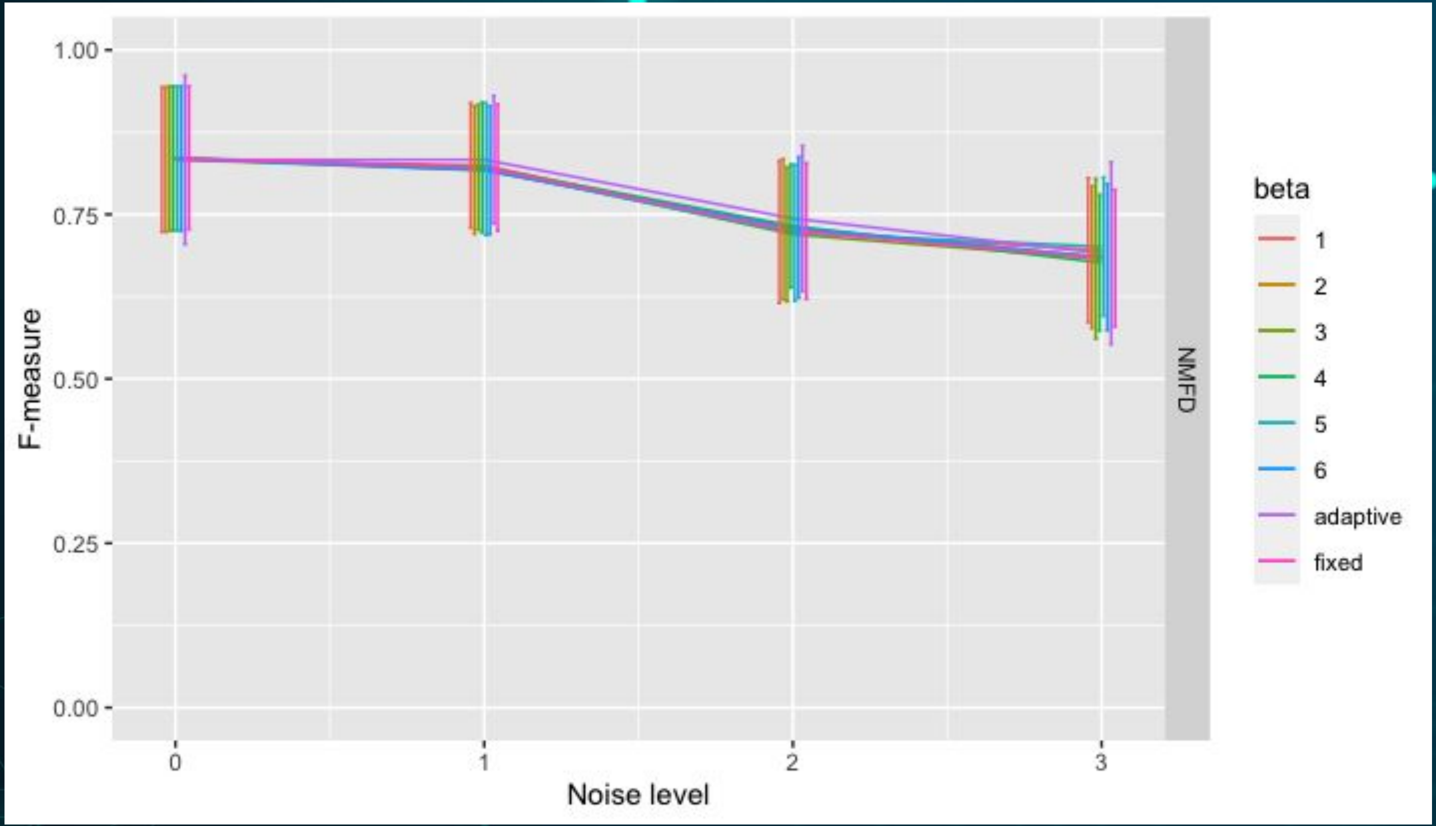
$\beta \rightarrow \infty$ - fixed

$\beta = 0$ - adaptive

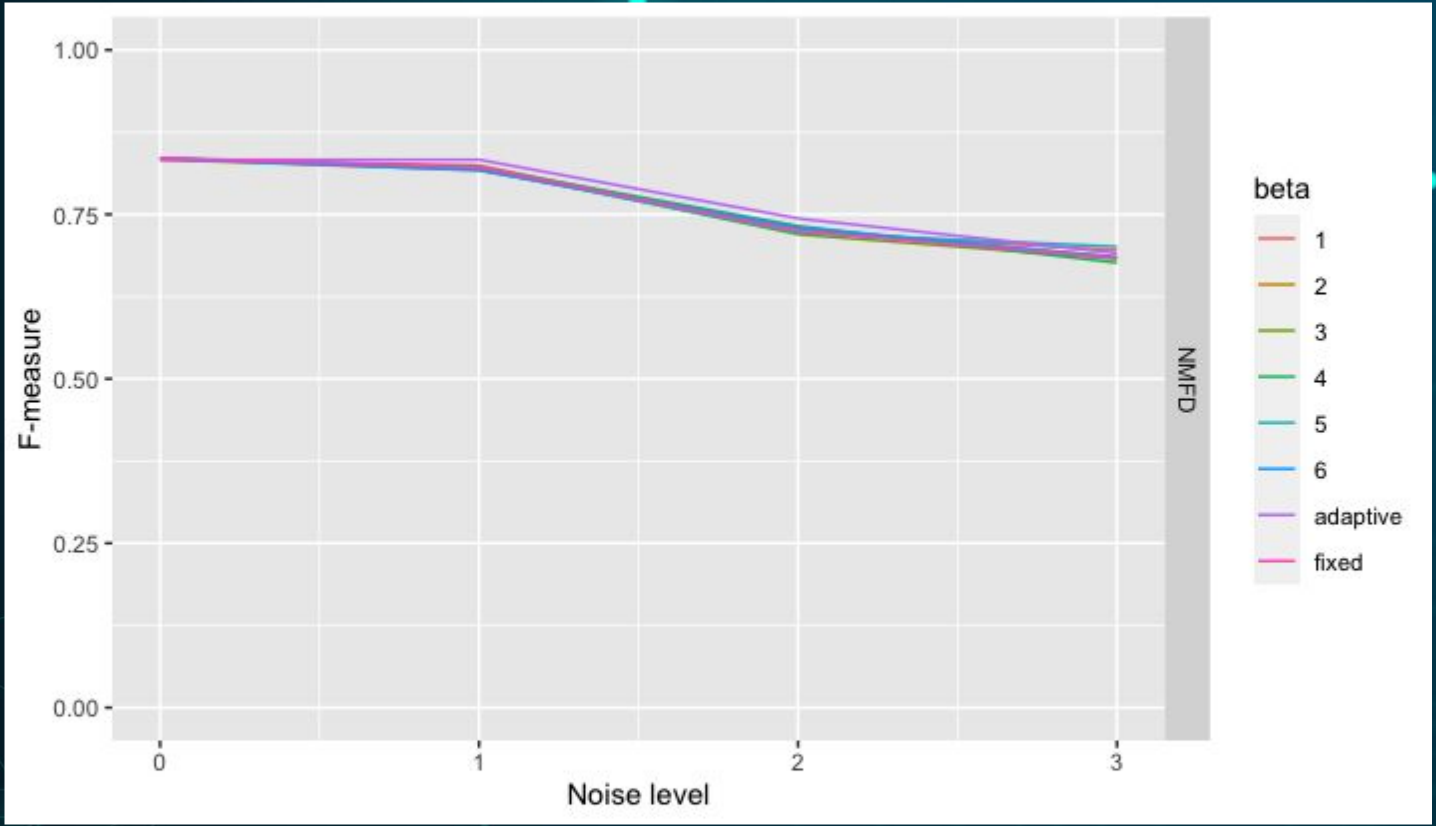




$$\mathbf{W} \leftarrow (1-\alpha) \cdot \mathbf{W}^{(0)} + \alpha \cdot \mathbf{W} \odot \frac{\frac{\mathbf{V}}{\hat{\mathbf{V}}}\mathbf{H}^{\mathsf{T}}}{\mathbf{J}\mathbf{H}^{\mathsf{T}}} \qquad \alpha = \left(\frac{k}{K}\right)^{\beta}$$

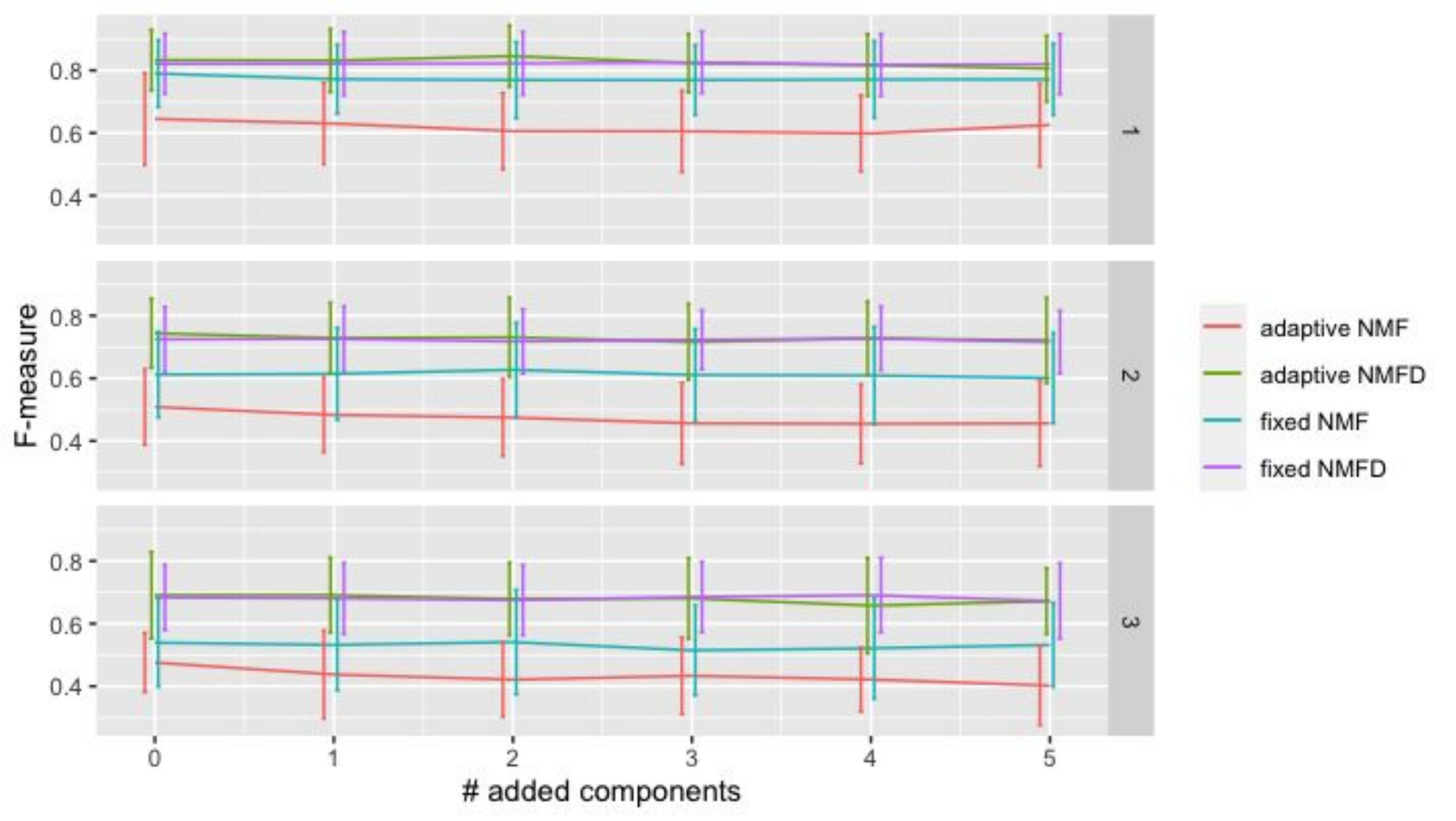


$$\mathbf{W} \leftarrow (1-\alpha) \cdot \mathbf{W}^{(0)} + \alpha \cdot \mathbf{W} \odot \frac{\frac{\mathbf{V}}{\hat{\mathbf{V}}}\mathbf{H}^{\mathsf{T}}}{\mathbf{J}\mathbf{H}^{\mathsf{T}}} \qquad \alpha = \left(\frac{k}{K}\right)^{\beta}$$

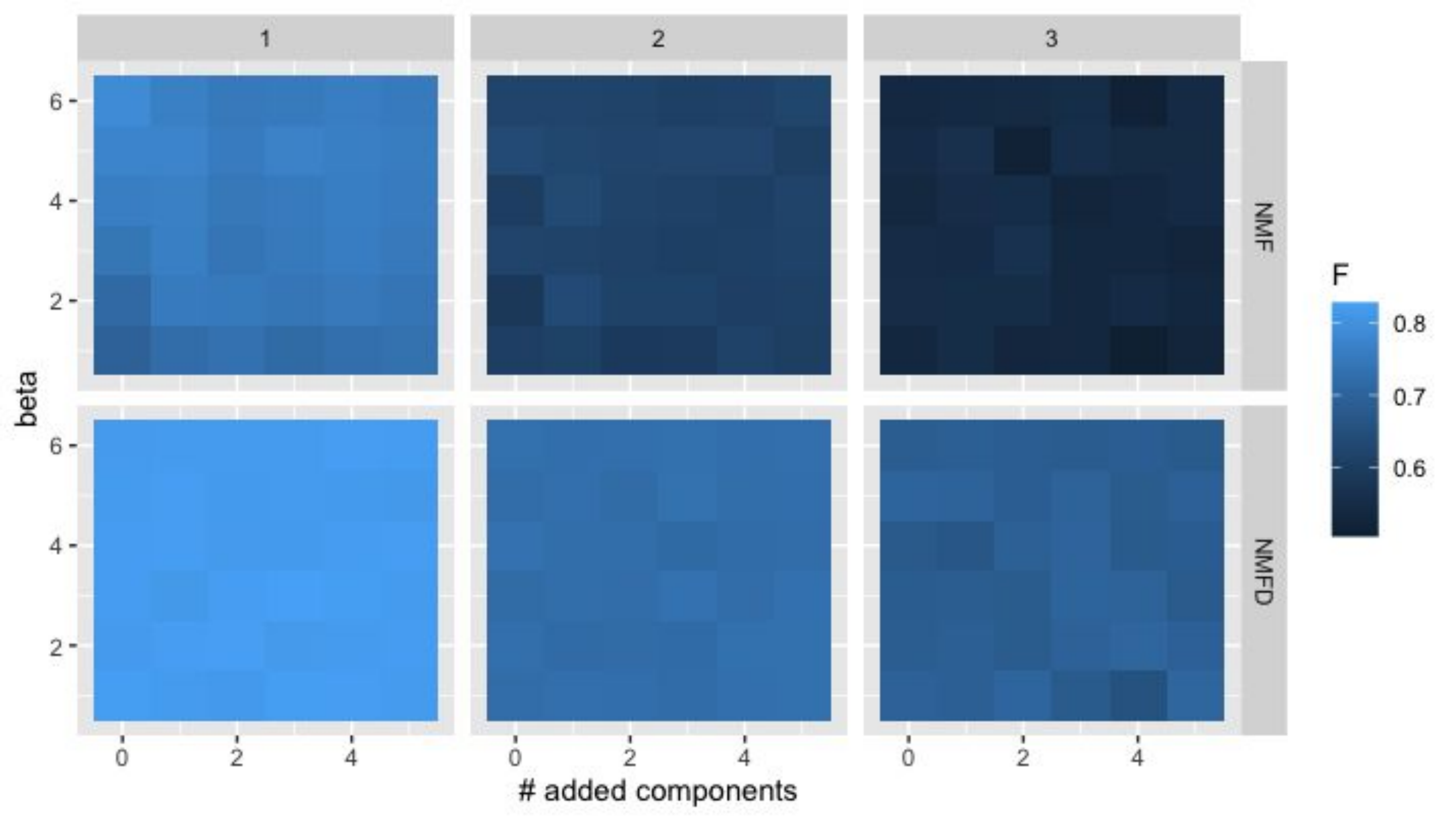


$$\mathbf{W} \leftarrow (1-\alpha) \cdot \mathbf{W}^{(0)} + \alpha \cdot \mathbf{W} \odot \frac{\frac{\mathbf{V}}{\hat{\mathbf{V}}}\mathbf{H}^{\mathsf{T}}}{\mathbf{J}\mathbf{H}^{\mathsf{T}}} \qquad \alpha = \left(\frac{k}{K}\right)^{\beta}$$

Added noise components



Added noise components



Best: Adaptive NMFD

No noise	Mild	Loud	Extreme
0.83 ± 0.13	0.83 ± 0.1	0.74 ± 0.11	0.69 ± 0.14

Questions

Removed slides

Multiplicative update rules

NMF

update rules

$$\mathbf{V} \approx \hat{\mathbf{V}} = \mathbf{W} \cdot \mathbf{H}$$

$$\mathbf{H} \leftarrow \mathbf{H} - \gamma_{\mathbf{H}} \nabla \mathcal{L}(\mathbf{H})$$
$$\mathbf{W} \leftarrow \mathbf{W} - \gamma_{\mathbf{W}} \nabla \mathcal{L}(\mathbf{W})$$

Lee, D., & Seung, H. S. (2000). Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, *13*.

NMF
update rules

$$\mathbf{V} \approx \hat{\mathbf{V}} = \mathbf{W} \cdot \mathbf{H}$$

$$\mathbf{H} \leftarrow \mathbf{H} - \gamma_{\mathbf{H}} \nabla \mathcal{L}(\mathbf{H})$$
$$\mathbf{W} \leftarrow \mathbf{W} - \gamma_{\mathbf{W}} \nabla \mathcal{L}(\mathbf{W})$$

$$\gamma_{\mathbf{H}} := \frac{\mathbf{H}}{\mathbf{W}^{\mathsf{T}}\mathbf{W}\mathbf{H}}$$

$$\gamma_{\mathbf{W}} := \frac{\mathbf{W}}{\mathbf{W}\mathbf{H}\mathbf{H}^{\mathsf{T}}}$$

Lee, D., & Seung, H. S. (2000). Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, *13*.

NMF

update rules

$$\mathbf{V} \approx \hat{\mathbf{V}} = \mathbf{W} \cdot \mathbf{H}$$

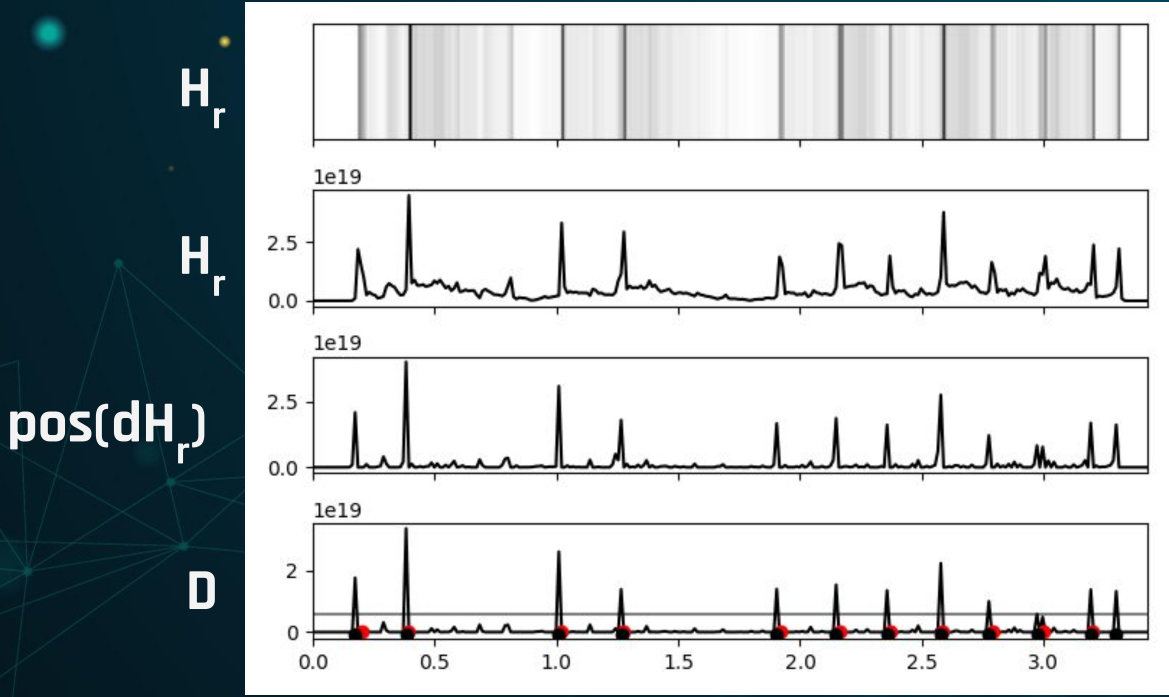
$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\frac{\mathbf{V}}{\hat{\mathbf{V}}}\mathbf{H}^\top}{\mathbf{J}\mathbf{H}^\top}$$

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^\top \frac{\mathbf{V}}{\hat{\mathbf{V}}}}{\mathbf{W}^\top \mathbf{J}},$$

Lee, D., & Seung, H. S. (2000). Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, *13*.

Onset detection

Onset detection



Müller Meinard. (2021). Tempo and Beat Tracking. In *Fundamentals of Music Processing: Using Python and Jupyter Notebooks* (2nd ed., pp. 311–315)., Springer.

Semi-adaptive

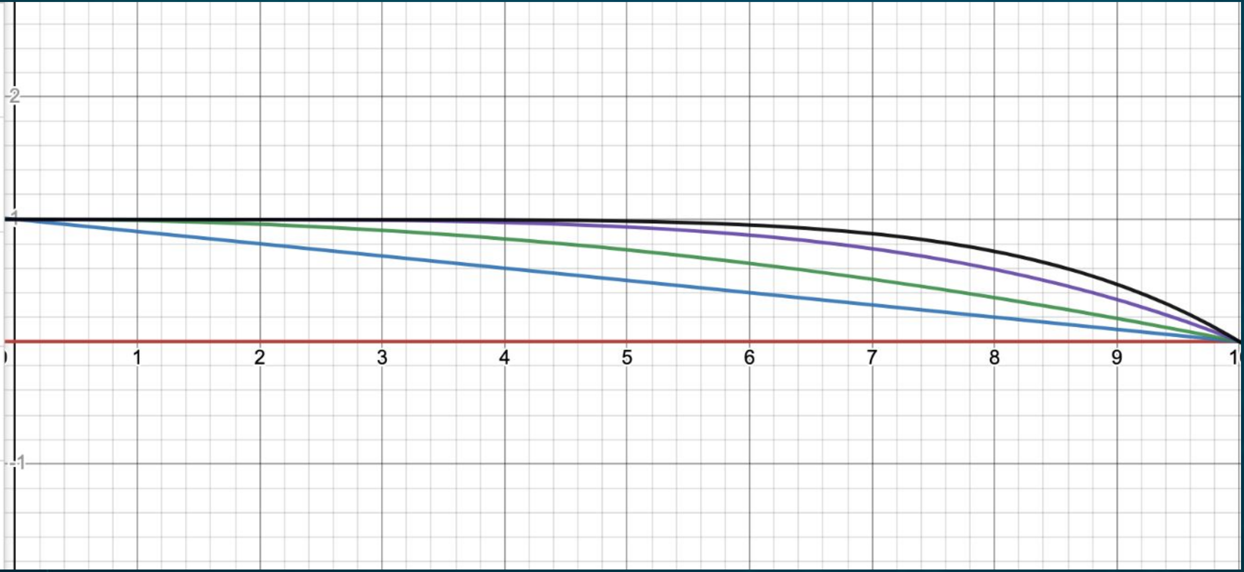
1 $1-\left(\frac{x}{10}\right)^{0}$

2 $1-\left(\frac{x}{10}\right)^{1}$

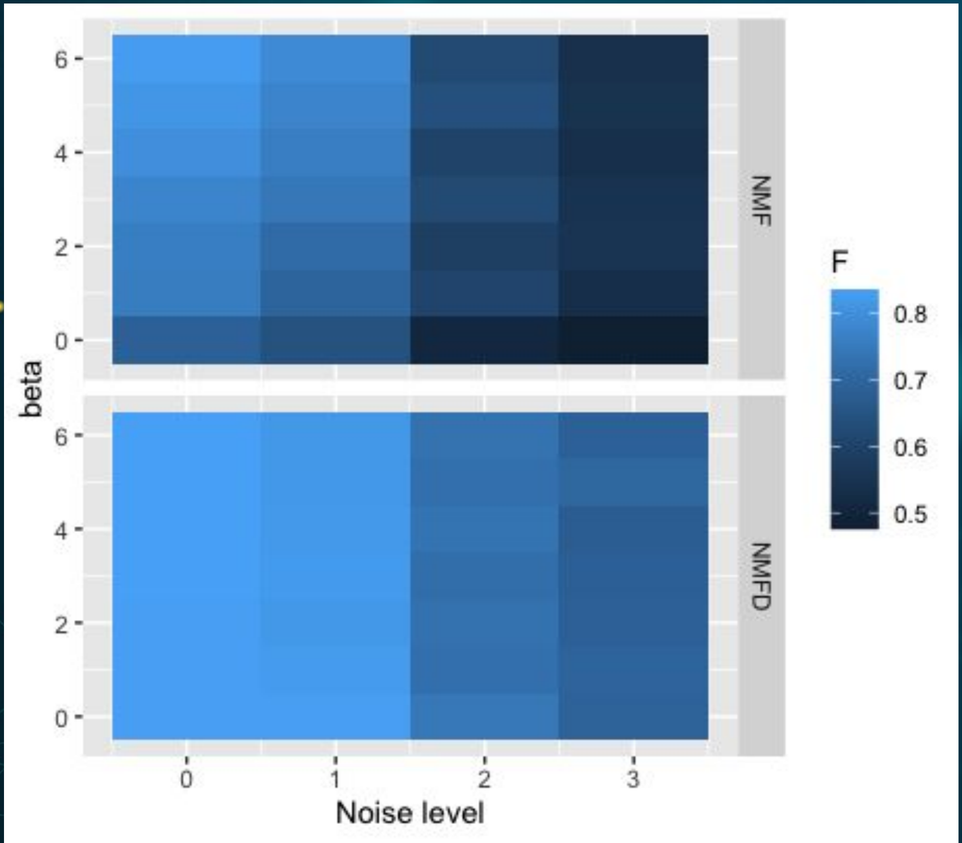
3 $1-\left(\frac{x}{10}\right)^{2}$

4 $1-\left(\frac{x}{10}\right)^{4}$

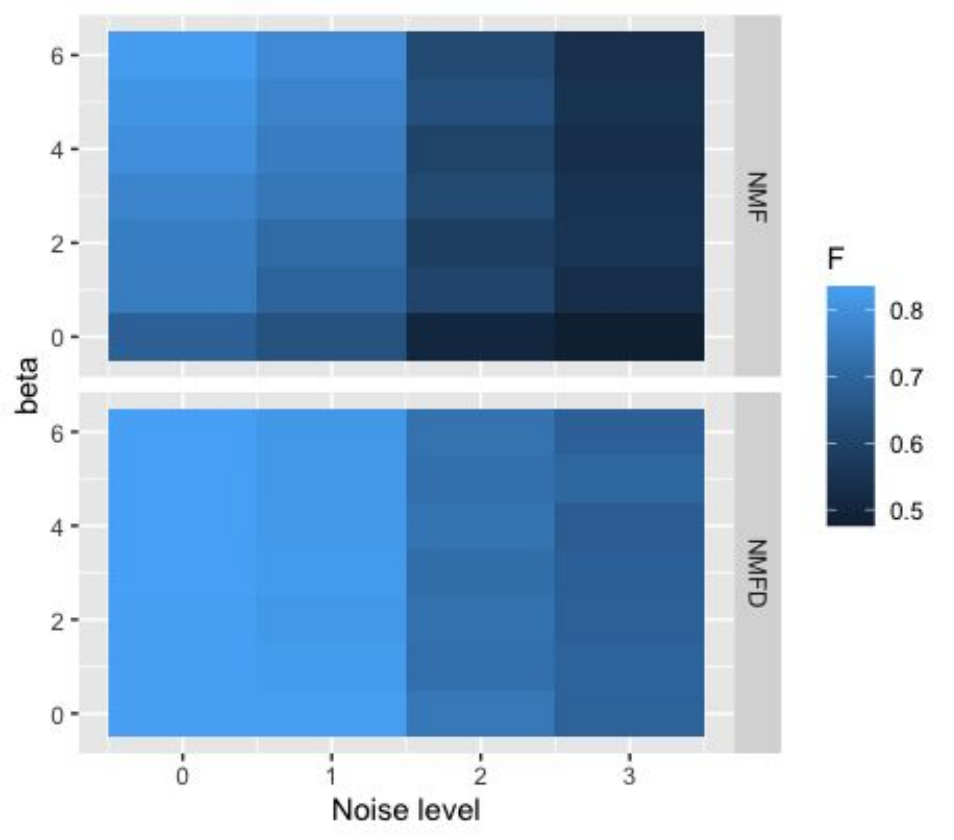
5 $1-\left(\frac{x}{10}\right)^{6}$



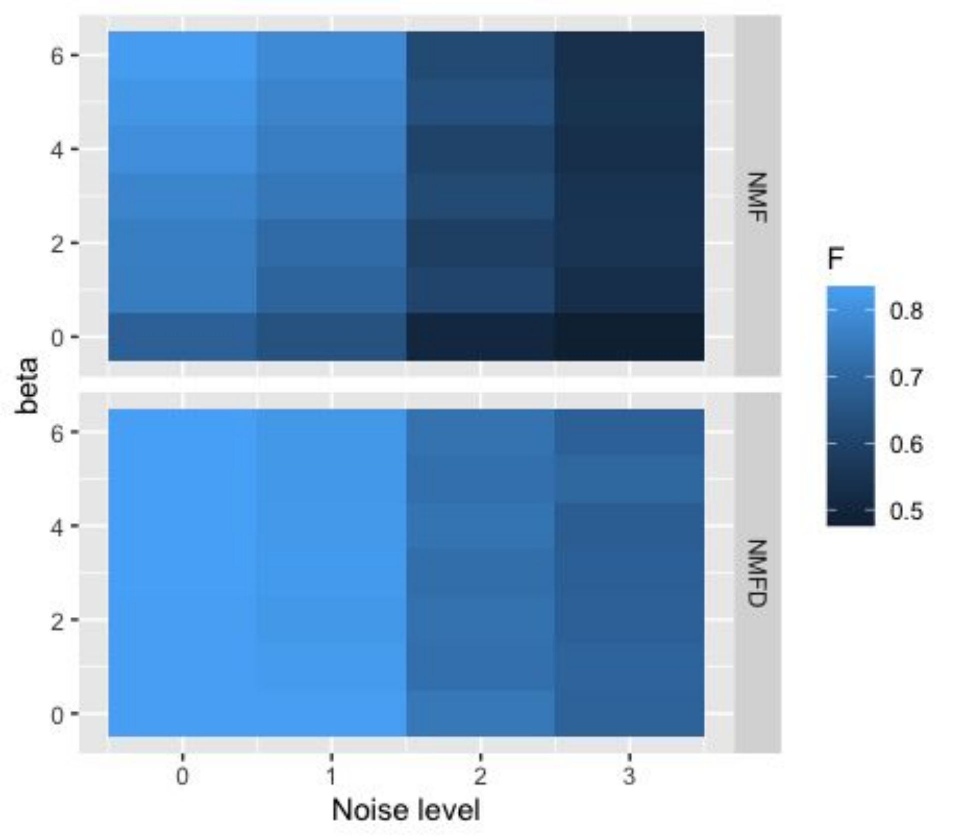
$$\mathbf{W} \leftarrow (1-\alpha)\cdot\mathbf{W}^{(0)} + \alpha\cdot\mathbf{W}\odot\frac{\frac{\mathbf{V}}{\hat{\mathbf{V}}}\mathbf{H}^{\mathsf{T}}}{\mathbf{J}\mathbf{H}^{\mathsf{T}}} \qquad \alpha = \left(\frac{k}{K}\right)^{\beta}$$



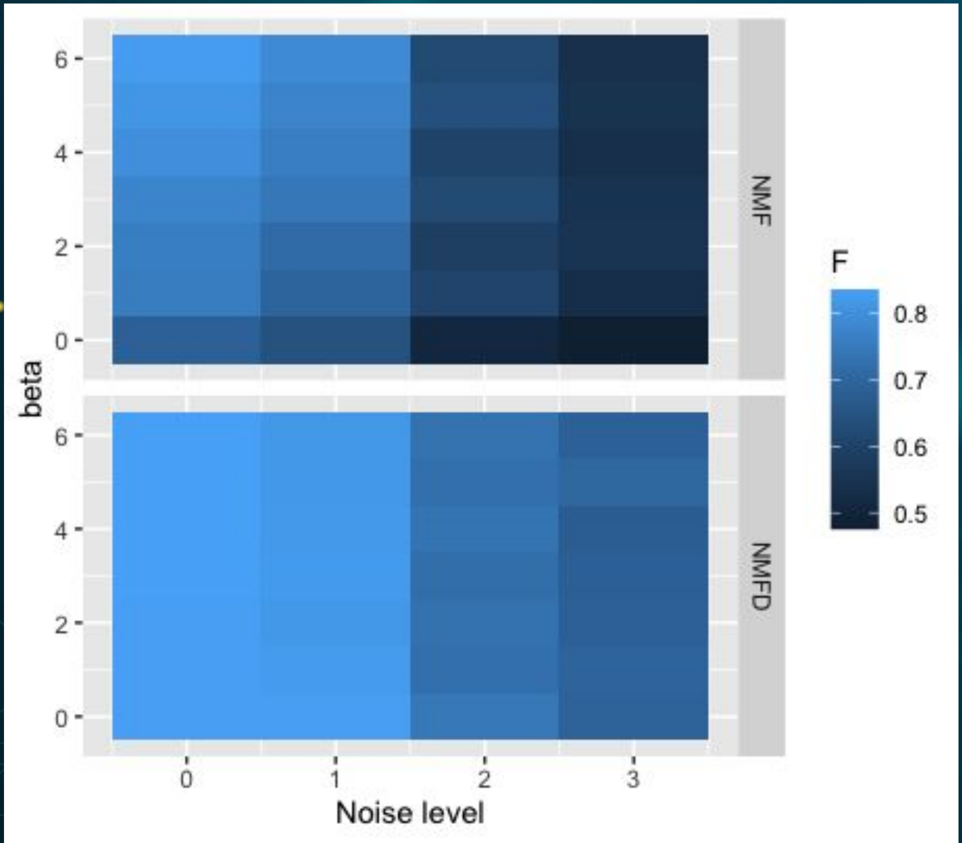
$$\mathbf{W} \leftarrow (1-\alpha) \cdot \mathbf{W}^{(0)} + \alpha \cdot \mathbf{W} \odot \frac{\frac{\mathbf{V}}{\hat{\mathbf{V}}}\mathbf{H}^{\mathsf{T}}}{\mathbf{J}\mathbf{H}^{\mathsf{T}}} \qquad \alpha = \left(\frac{k}{K}\right)^{\beta}$$



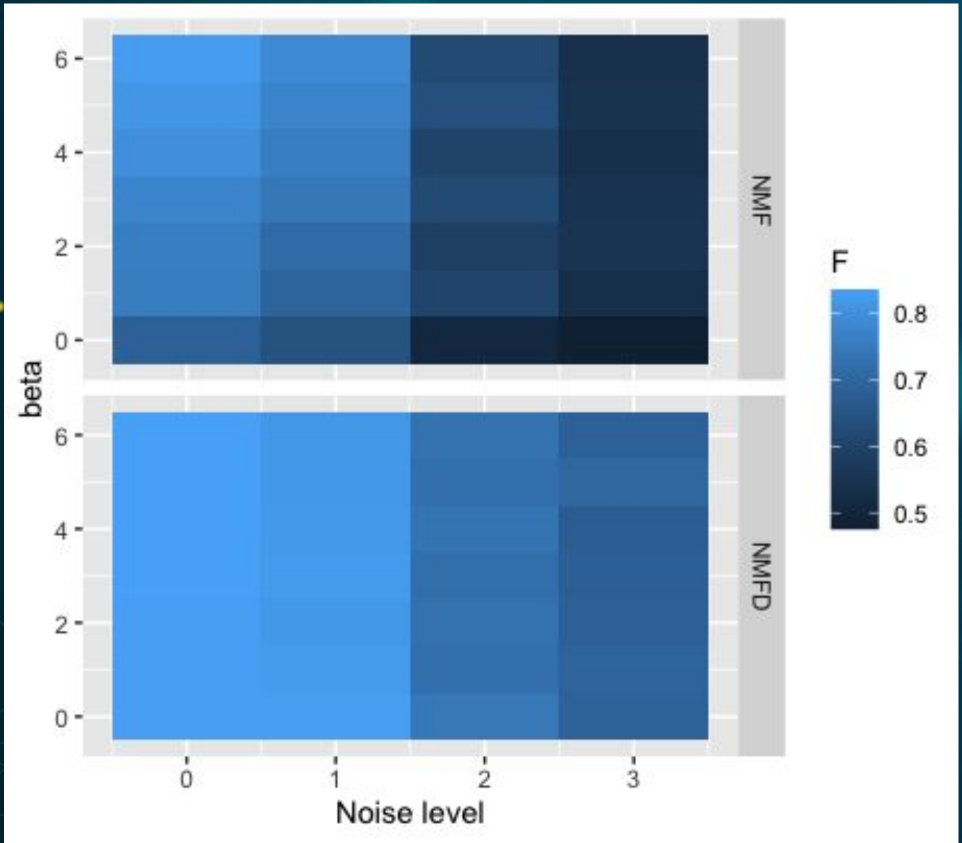
$$\mathbf{W} \leftarrow (1-\alpha) \cdot \mathbf{W}^{(0)} + \alpha \cdot \mathbf{W} \odot \frac{\frac{\mathbf{V}}{\hat{\mathbf{V}}}\mathbf{H}^{\mathsf{T}}}{\mathbf{J}\mathbf{H}^{\mathsf{T}}} \qquad \alpha = \left(\frac{k}{K}\right)^{0} = 1$$



$$\mathbf{W} \leftarrow \cancel{0 \cdot \mathbf{W}^{(0)} +} 1 \cdot \mathbf{W} \odot \frac{\frac{\mathbf{V}}{\hat{\mathbf{V}}}\mathbf{H}^\mathsf{T}}{\mathbf{J}\mathbf{H}^\mathsf{T}} \qquad \alpha = \left(\frac{k}{K}\right)^0 = 1$$

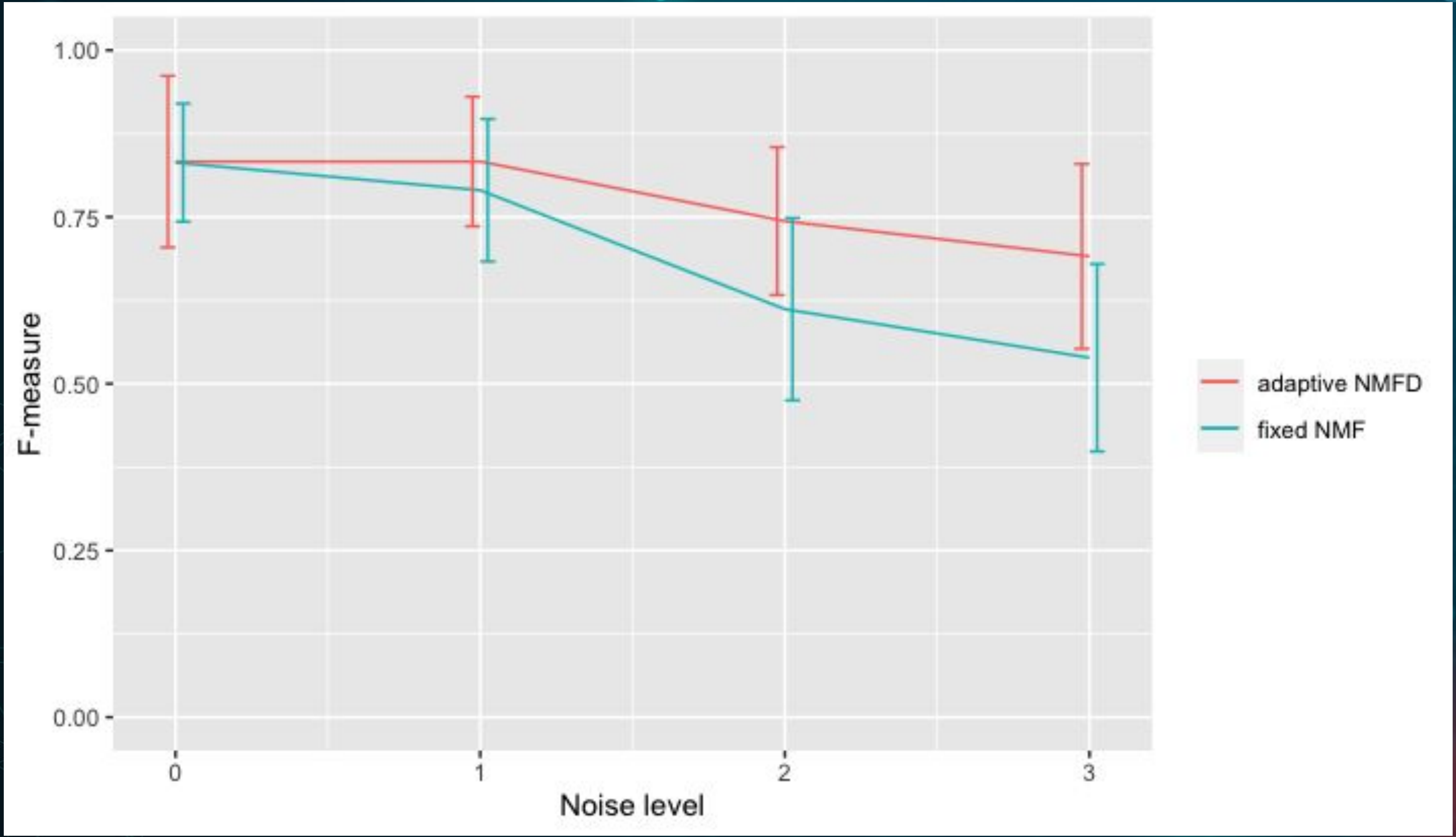


$$\mathbf{W} \leftarrow (1-\alpha)\cdot\mathbf{W}^{(0)} + \alpha\cdot\mathbf{W}\odot\frac{\frac{\mathbf{V}}{\hat{\mathbf{V}}}\mathbf{H}^{\mathsf{T}}}{\mathbf{J}\mathbf{H}^{\mathsf{T}}} \qquad \alpha = \left(\frac{k}{K}\right)^{\infty} = 0$$



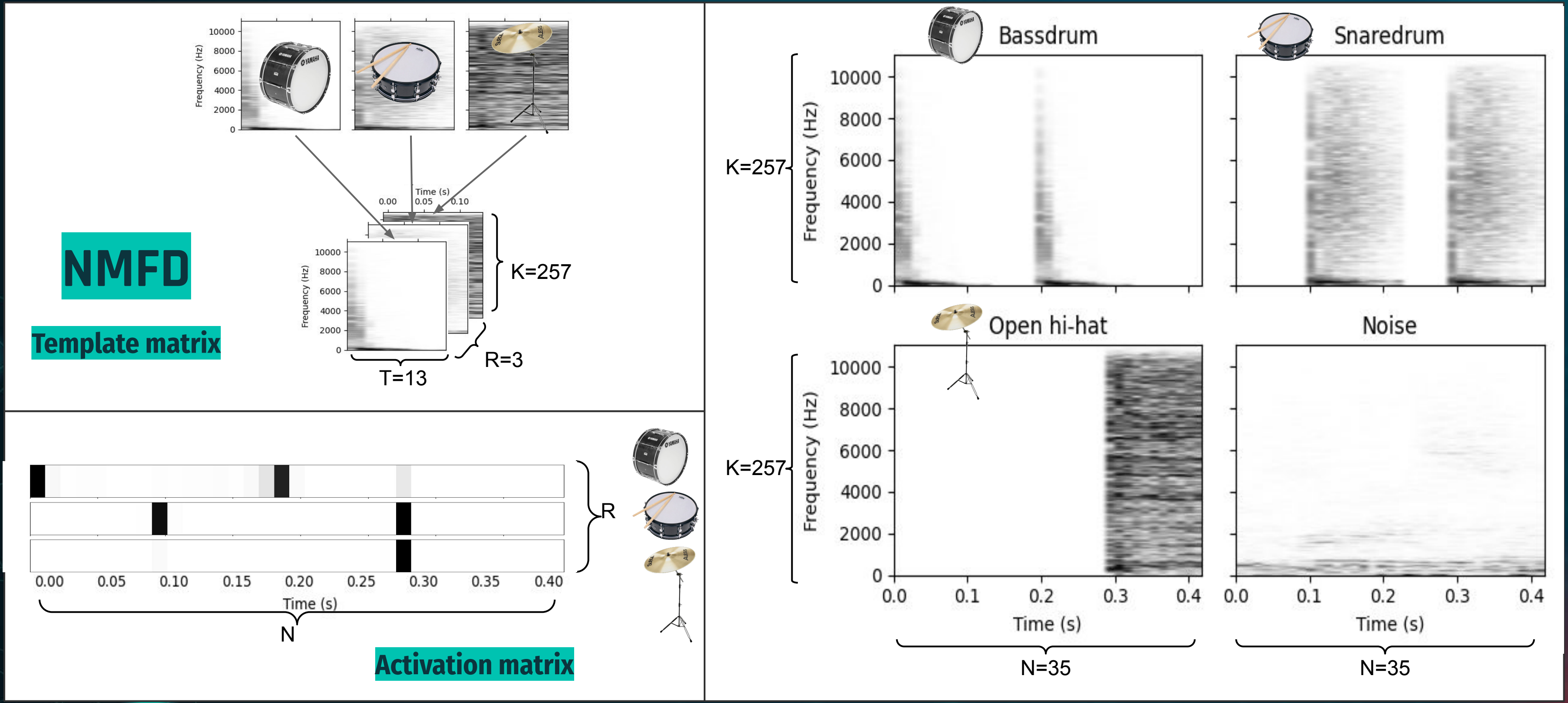
$$\mathbf{W} \leftarrow \quad 1 \quad \cdot \mathbf{W}^{(0)} + \; 0 \; \cdot \cancel{\mathbf{W} \odot \frac{\frac{\mathbf{V}}{\hat{\mathbf{V}}}\mathbf{H}^{\mathsf{T}}}{\mathbf{J}\mathbf{H}^{\mathsf{T}}}} \qquad \alpha = \left(\frac{k}{K}\right)^{\infty} = \; 0$$

Results



1.00
0.75
0.50
0.25
0.00
F-measure
0
1
2
3
Noise level
adaptive NMFD
fixed NMF

NMFD



NMFD
Template matrix
Frequency (Hz)
10000
8000
6000
4000
2000
0
Time (s)
0.00
0.05
0.10
K=257
R=3
T=13
R
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Time (s)
N
Activation matrix
Bassdrum
Snaredrum
Open hi-hat
Noise
K=257
Frequency (Hz)
0.0
0.1
0.2
0.3
0.4
Time (s)
N=35
N=35