***Rajesh R** (S21005) - MS by Research*

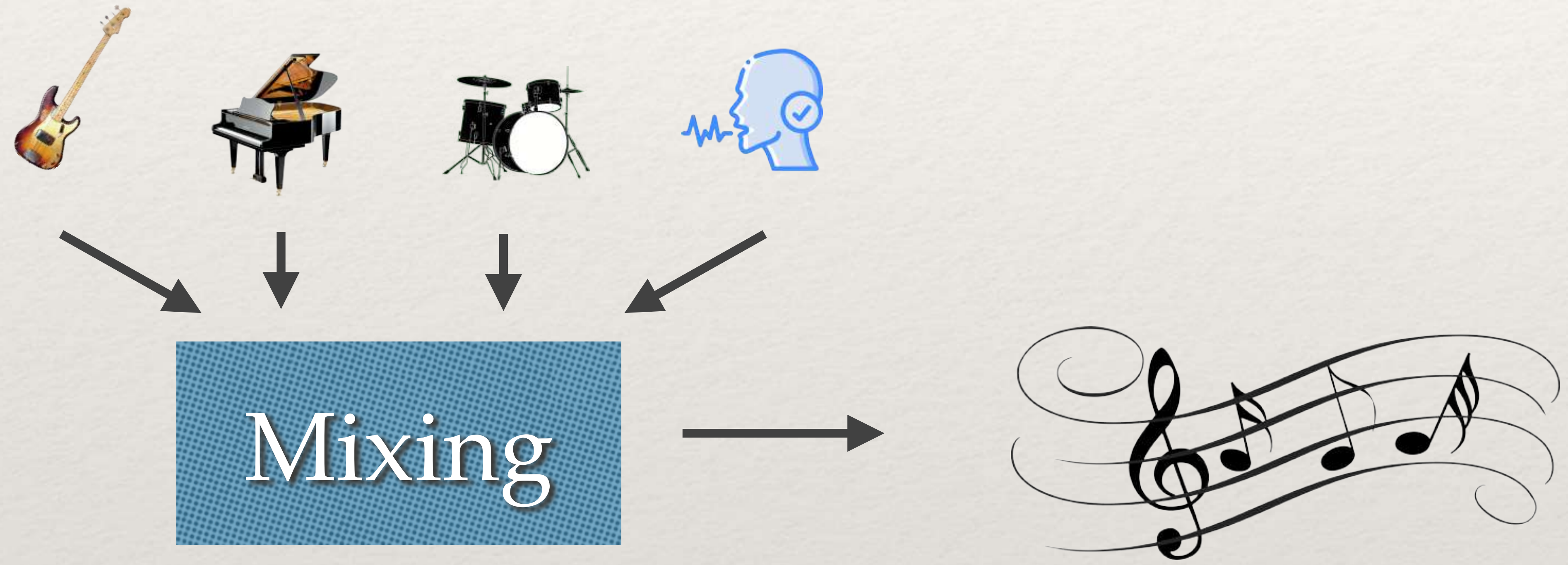
Interference Reduction in Music Source Separation for Live Recordings

Open Seminar | 27 October 2023
Guide: Dr. Padmanabhan Rajan
School of Computing & Electrical Engineering,
Indian Institute of Technology, Mandi

Music Source Separation

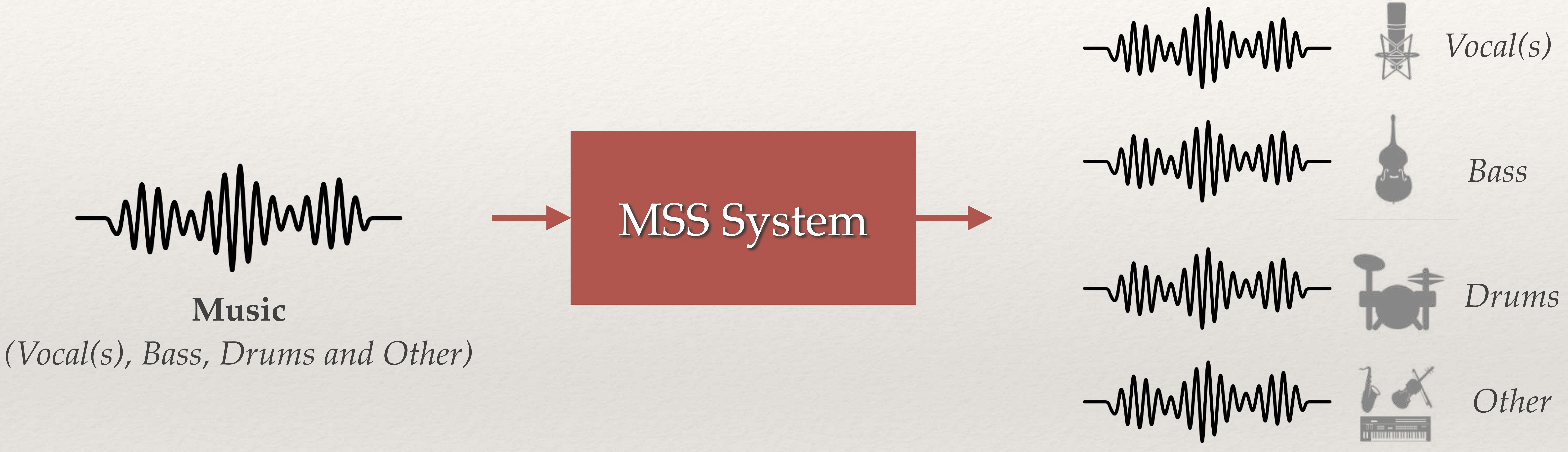
Music

- Music production

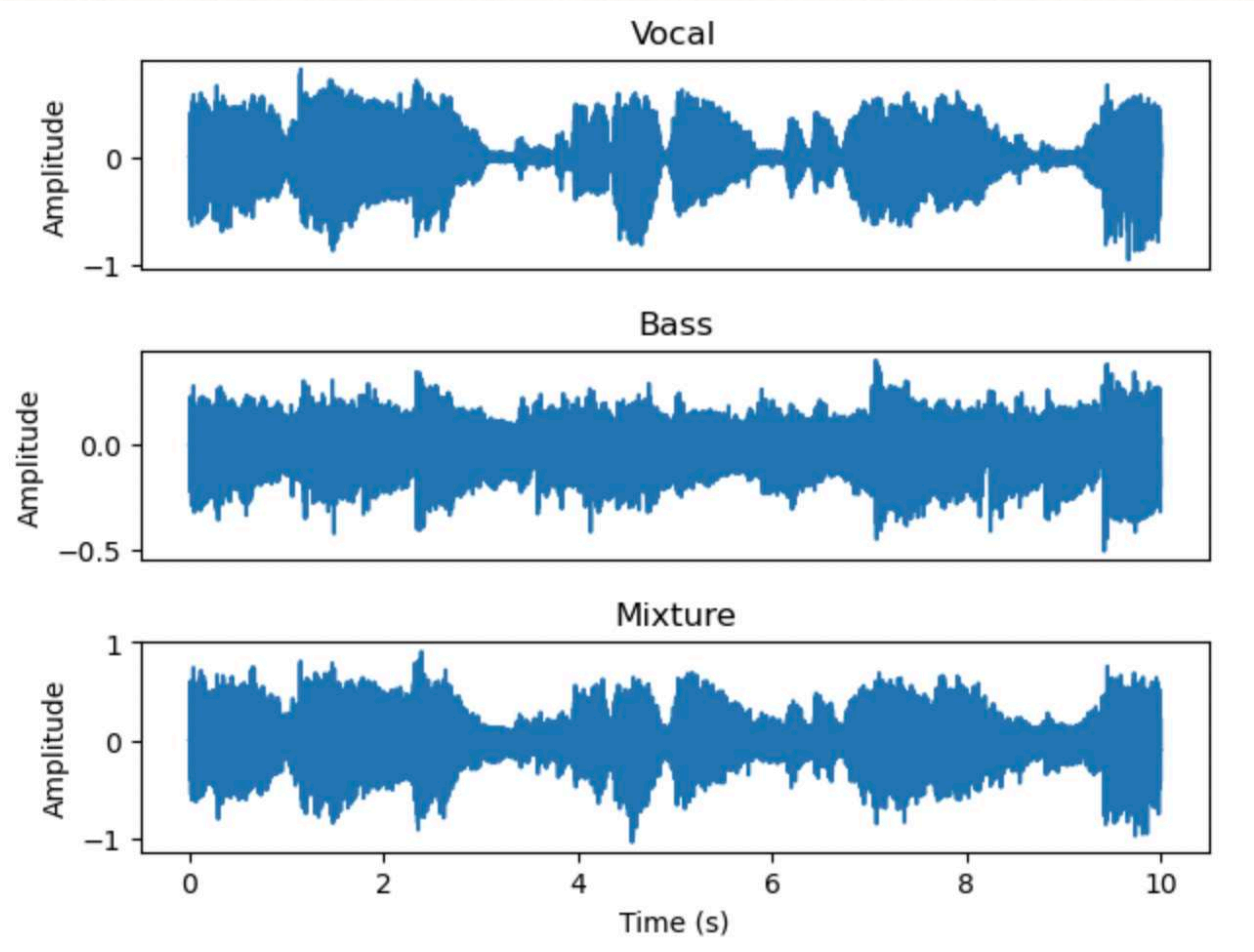


https://www.seekpng.com/ima/u2q8r5y3e6y3r5u2/

Music Source Separation



Hard Problem?



Music Characteristics

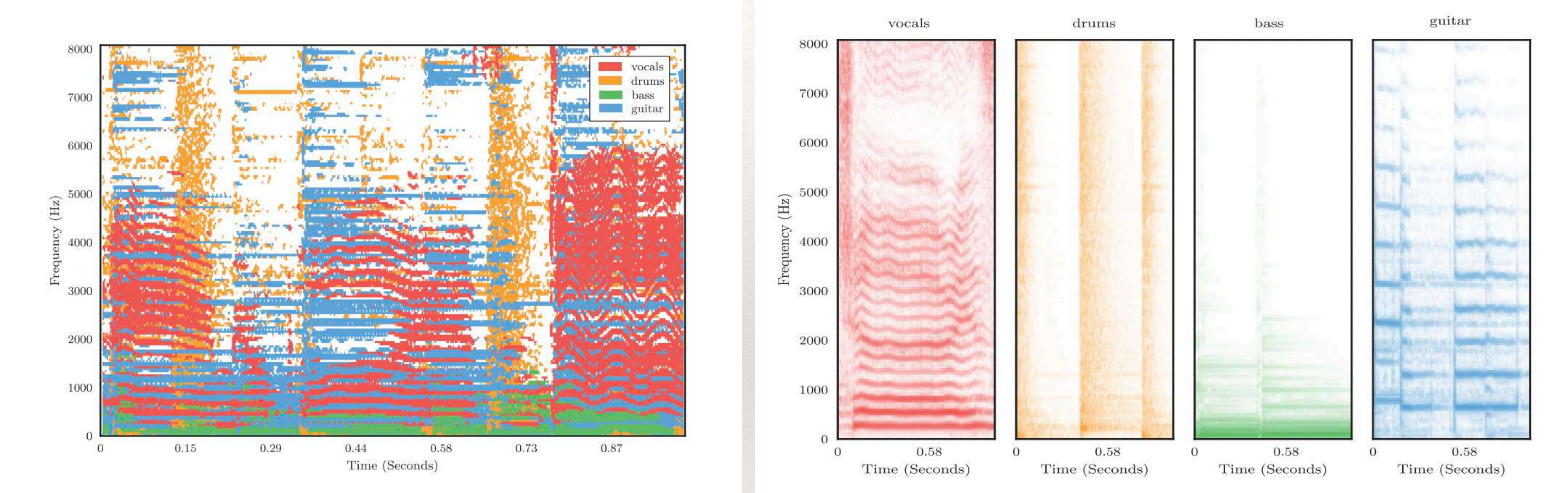
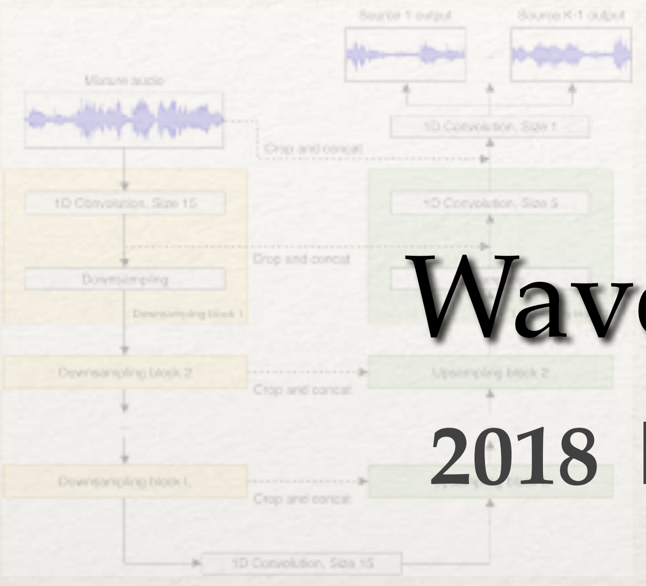


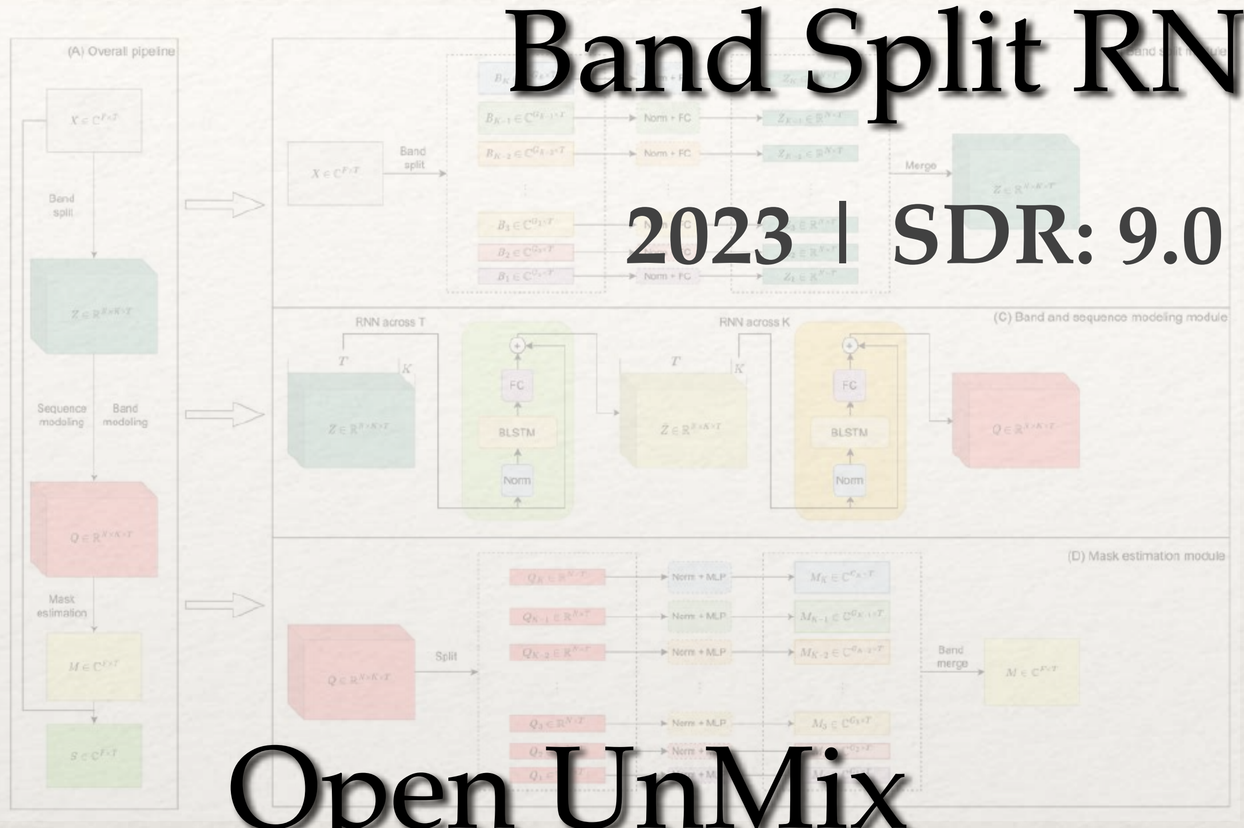
Figure taken from: *Estefania Cano, Derry Fitzgerald, Antoine Liutkus, Mark Plumbley, Fabian-Robert Stöter. Musical Source Separation: An Introduction. IEEE Signal Processing Magazine, Institute of Electrical and Electronics Engineers, 2019, 36 (1), pp.31-40.*

The State-of-the-art



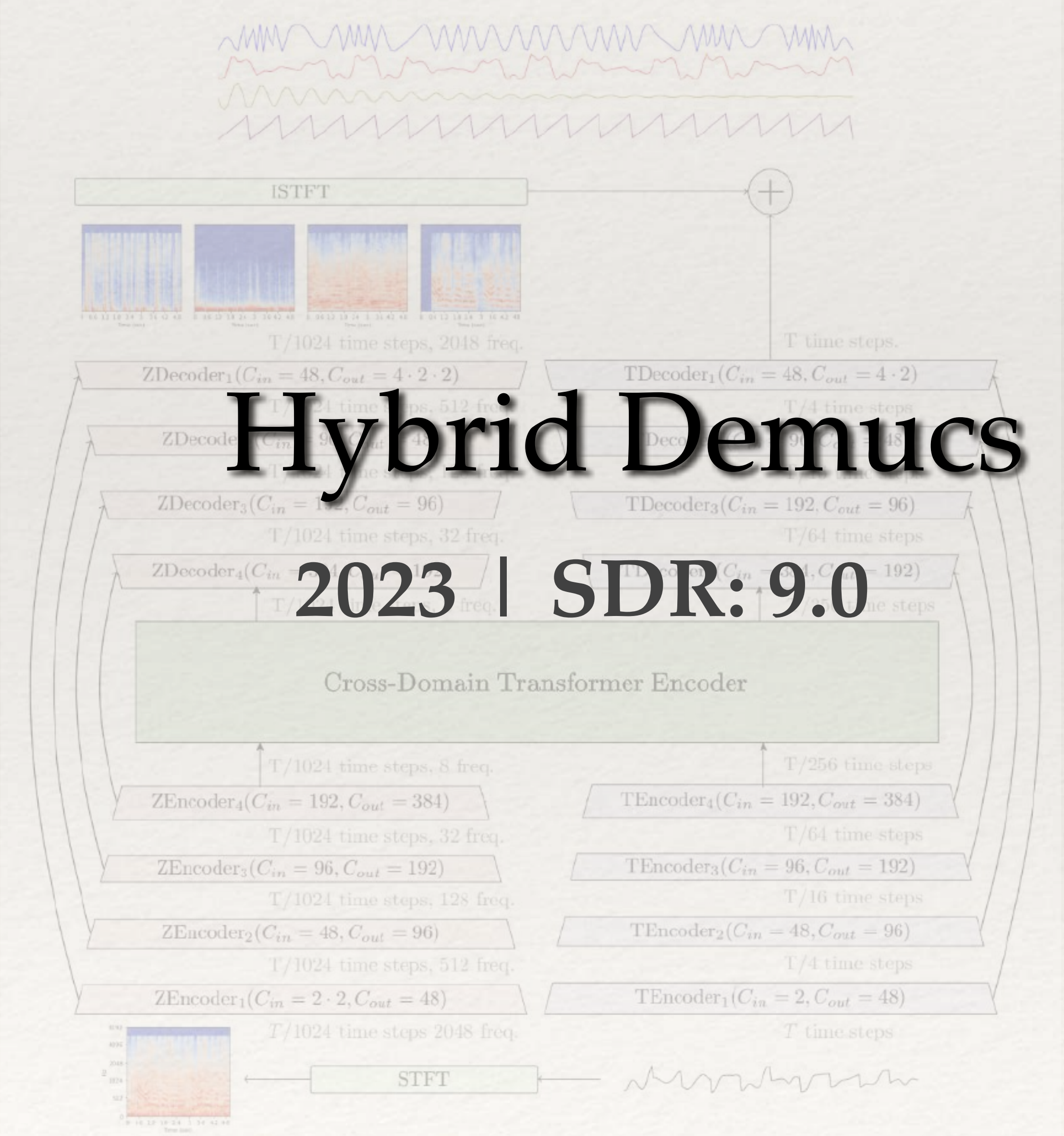
Wave-U-Net

2018 | SDR: 3.2



Band Split RNN

2023 | SDR: 9.0

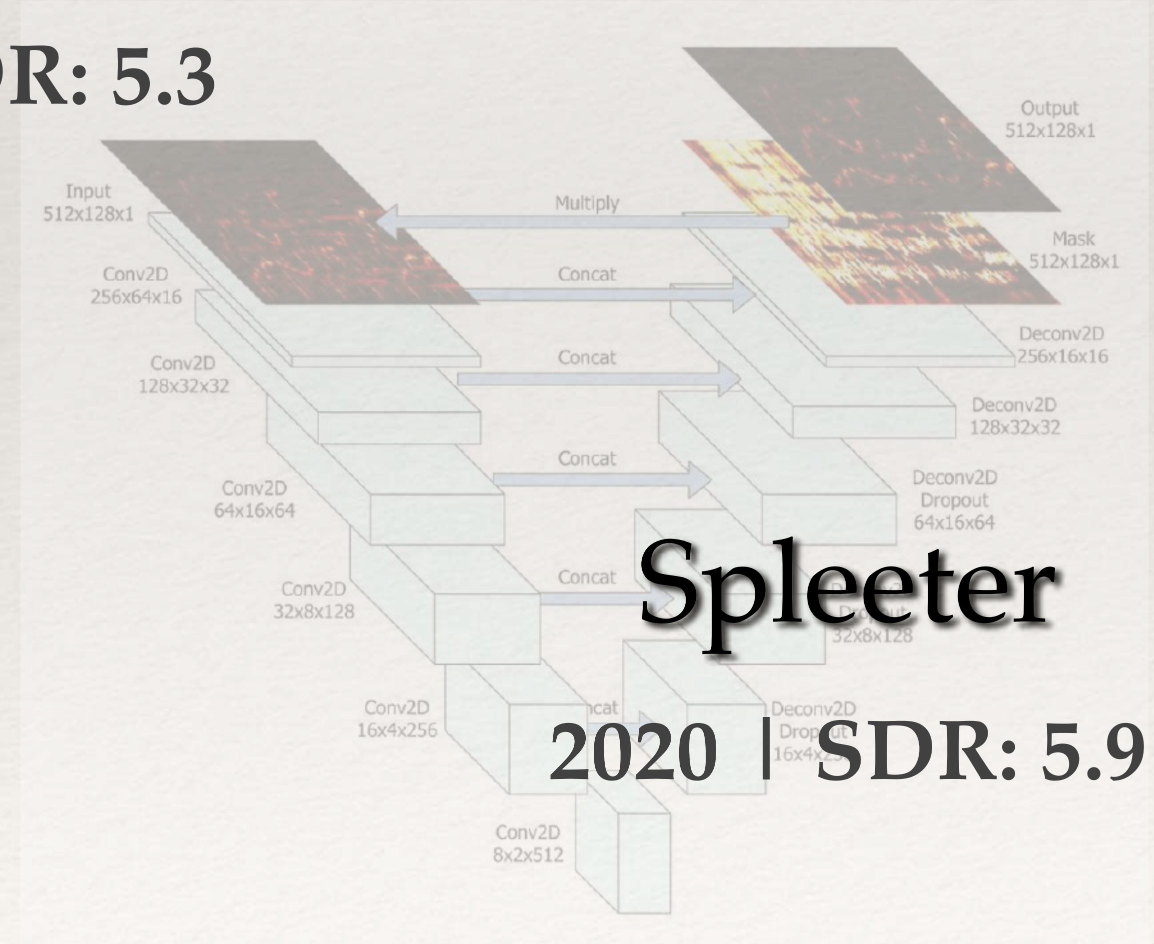


Hybrid Demucs

2023 | SDR: 9.0

Open UnMix

2019 | SDR: 5.3



Spleeter

2020 | SDR: 5.9

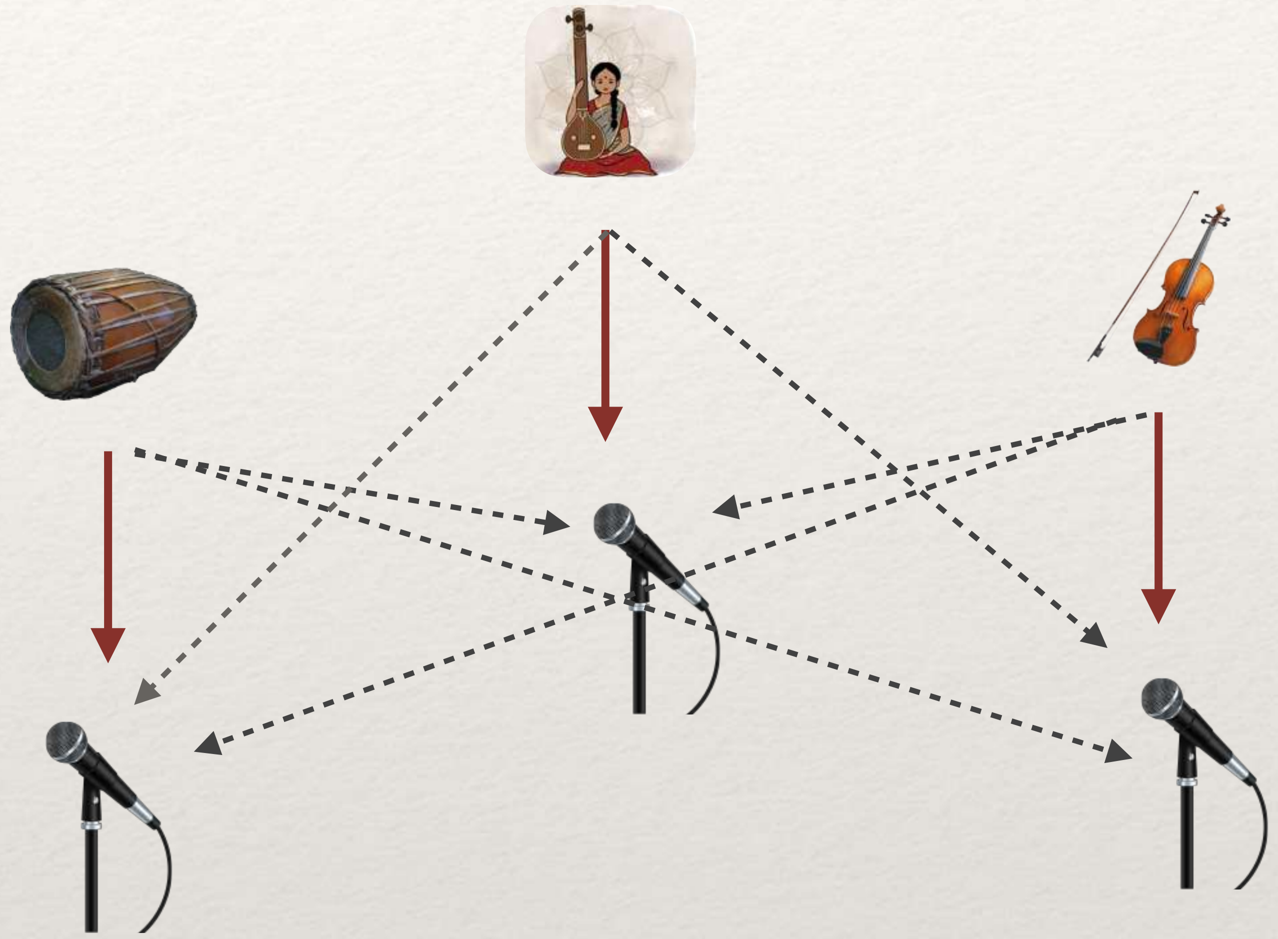
Western Pop Music
MUSDB18

SOURCE TO DISTORTION RATIO

Spotify
Google
facebook research

Recordings from Live Concerts

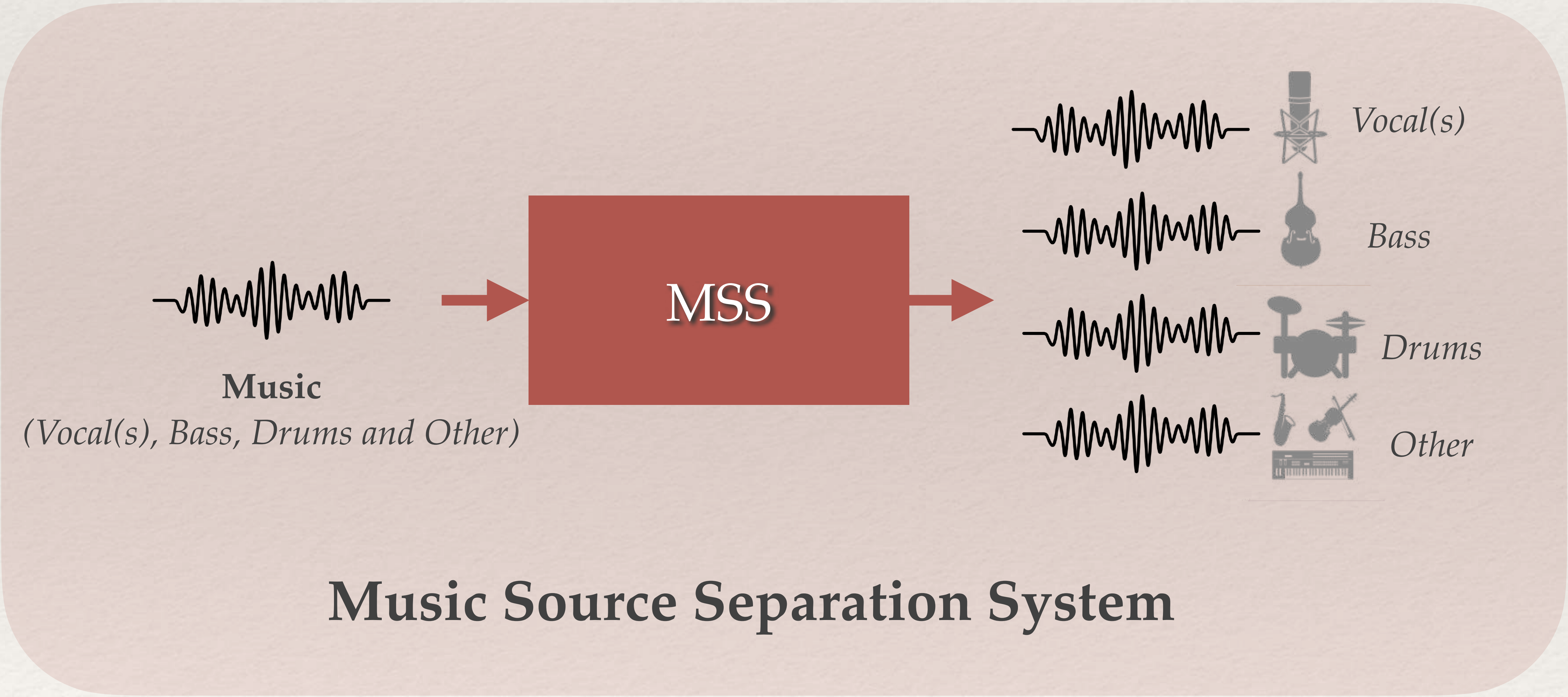
https://images.app.goo.gl/g9MPV2bNE5faJz4M7

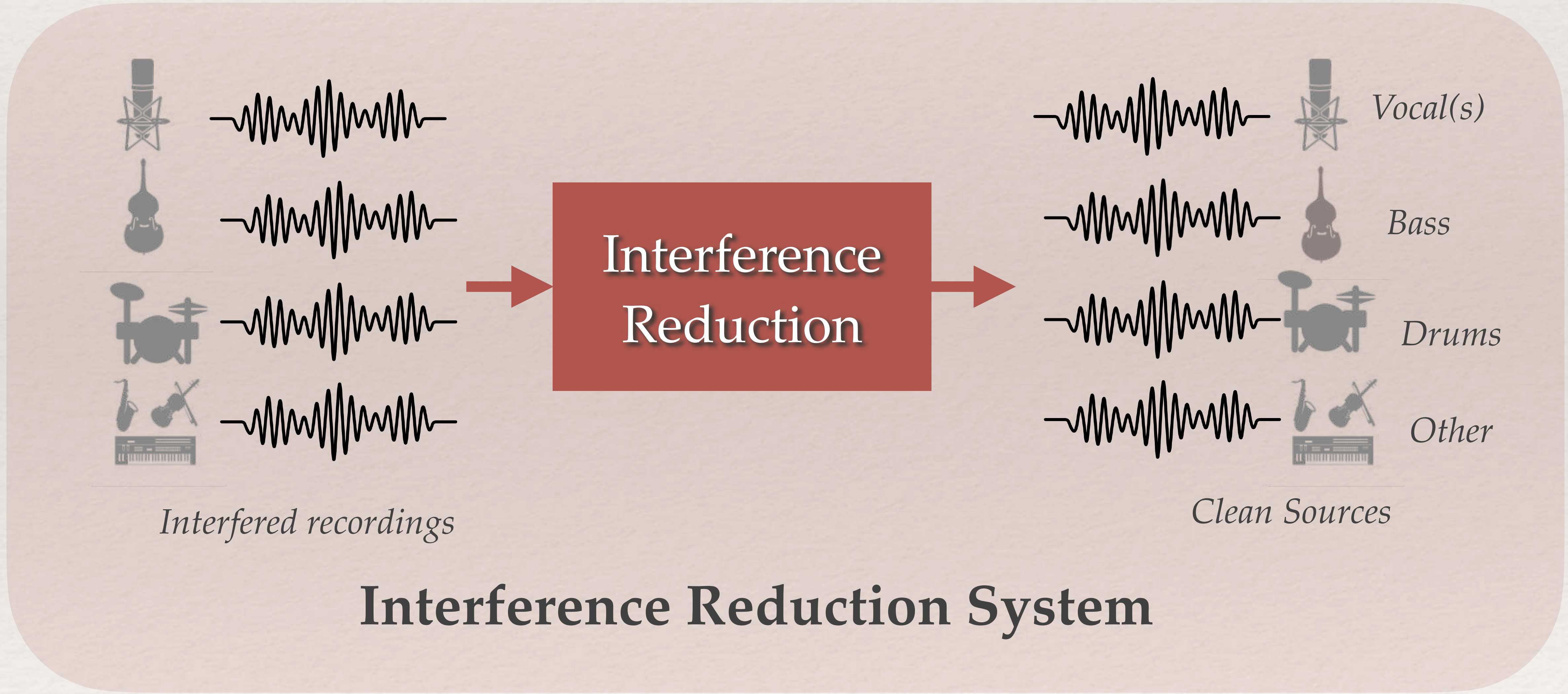


- Live recordings lacks acoustic shielding
- Microphone intended to pick specific source picks up the other sources as well

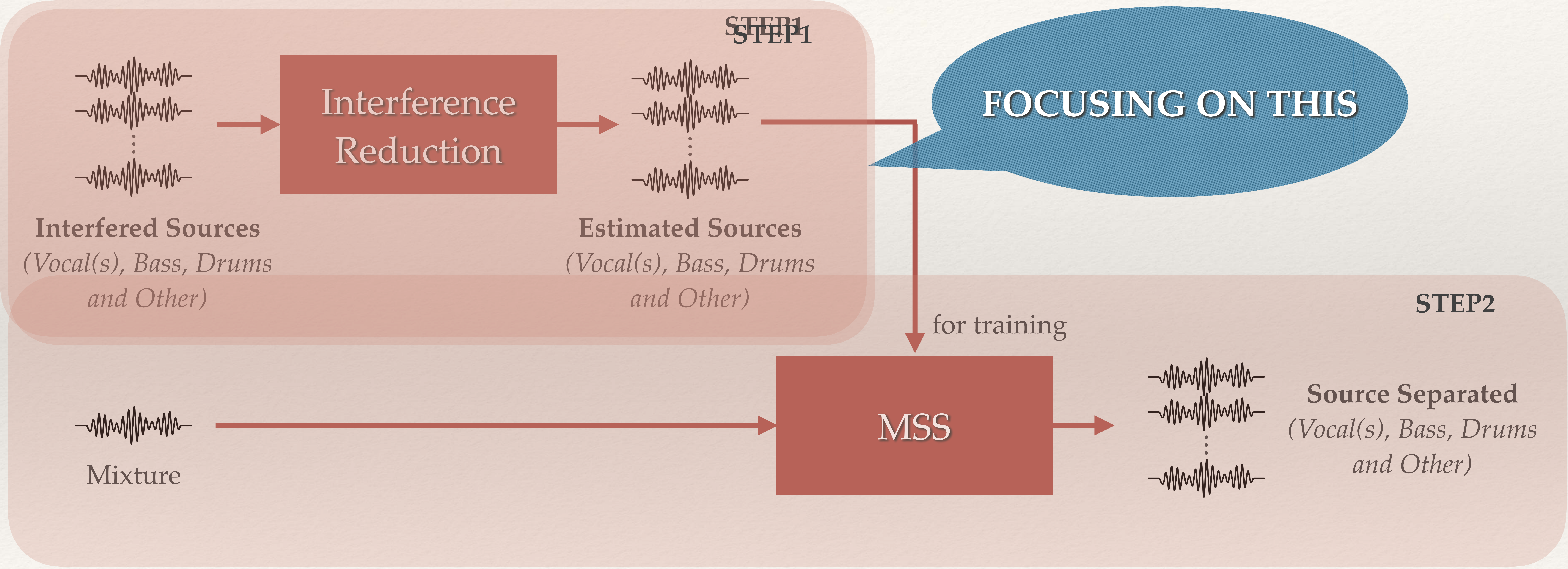
MSS vs Interference Reduction

- Interference reduction: Special type of source separation
- Aim: Clean microphone recordings





Overall Pipeline



Interference Reduction

Trends in Interference Reduction

- No neural network-based techniques proposed, due to dataset?
- DSP Algorithms: **KAMIR** (Kernel Additive Modelling for Interference Reduction) - the state-of-the-art [2015]
- MIRA (Multitrack Interference Reduction Algorithm) & FastMIRA are the advancement of KAMIR

Contributions

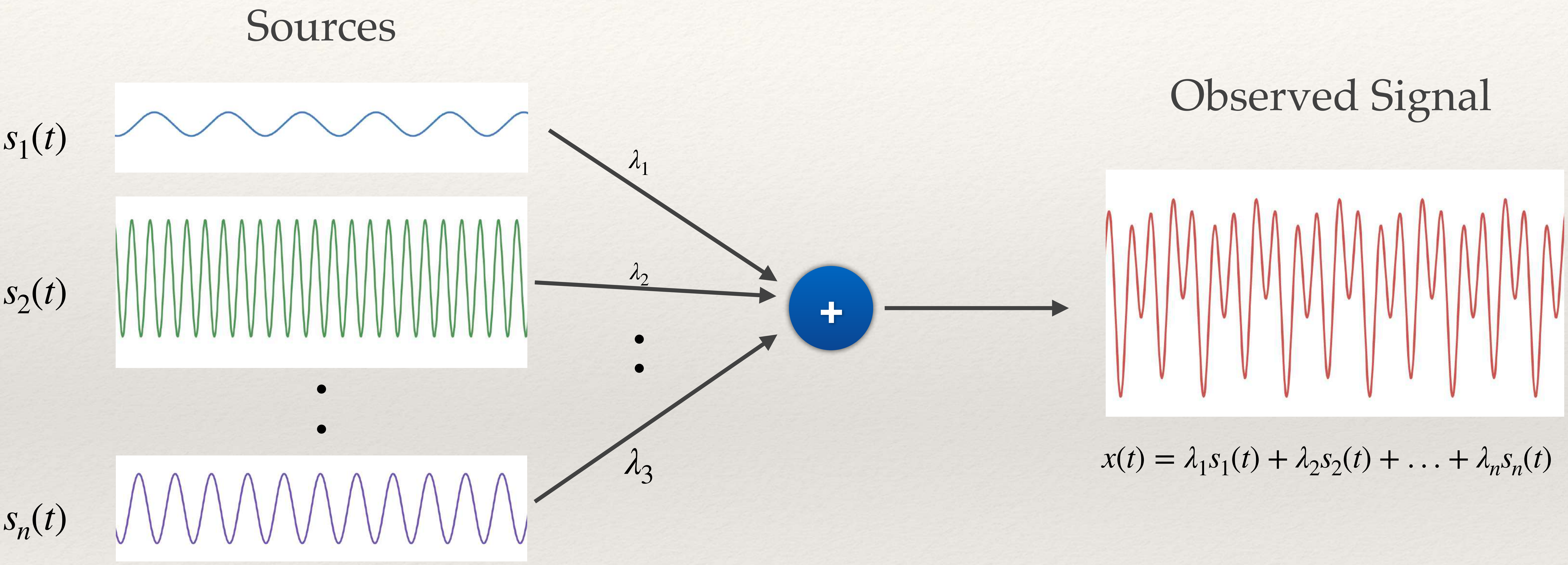
- Learning free Optimisation Algorithm
- Convolutional Autoencoders (CAEs)
- Truncated UNet (t-UNet)
- Dilated full Wave-U-Net (dfUNet) with Graph Attentions

Assumptions

https://images.app.goo.gl/g9MPV2bNE5faJz4M7

- Each source has at least one dedicated microphones.
- At least a single source is dominant in its dedicated microphone.

Mathematical Formulation



$$x(t) = \lambda_1 s_1(t) + \lambda_2 s_2(t) + \ldots + \lambda_n s_n(t)$$

Mathematical Formulation

For k microphones and n sources,

$$x_1(t) = \lambda_{11}s_1(t) + \lambda_{12}s_2(t) + .. + \lambda_{1n}s_n(t)$$

$$x_2(t) = \lambda_{21}s_1(t) + \lambda_{22}s_2(t) + .. + \lambda_{2n}s_n(t)$$

$$\vdots$$

$$x_k(t) = \lambda_{k1}s_1(t) + \lambda_{k2}s_2(t) + .. + \lambda_{kn}s_n(t)$$

$$X = \Lambda S$$

Microphone Recordings (X), Mixing Matrix (Λ), Source Signals (S)

$$X = [x_1(t), x_2(t), \ldots, x_k(t)]^T$$

$$S = [s_1(t), s_2(t), \ldots, s_n(t)]^T$$

Similarly for mixture signal,

$$m(t) = \sum_{i=0}^{n} \beta_i s_i(t) = b^T S$$

Issues with the problem

Equations: $X = \Lambda S$ and $m = b^T S$

- $X = \Lambda S$ is an over-determined or over-constrained problem
- No unique solution, multiple solution exists

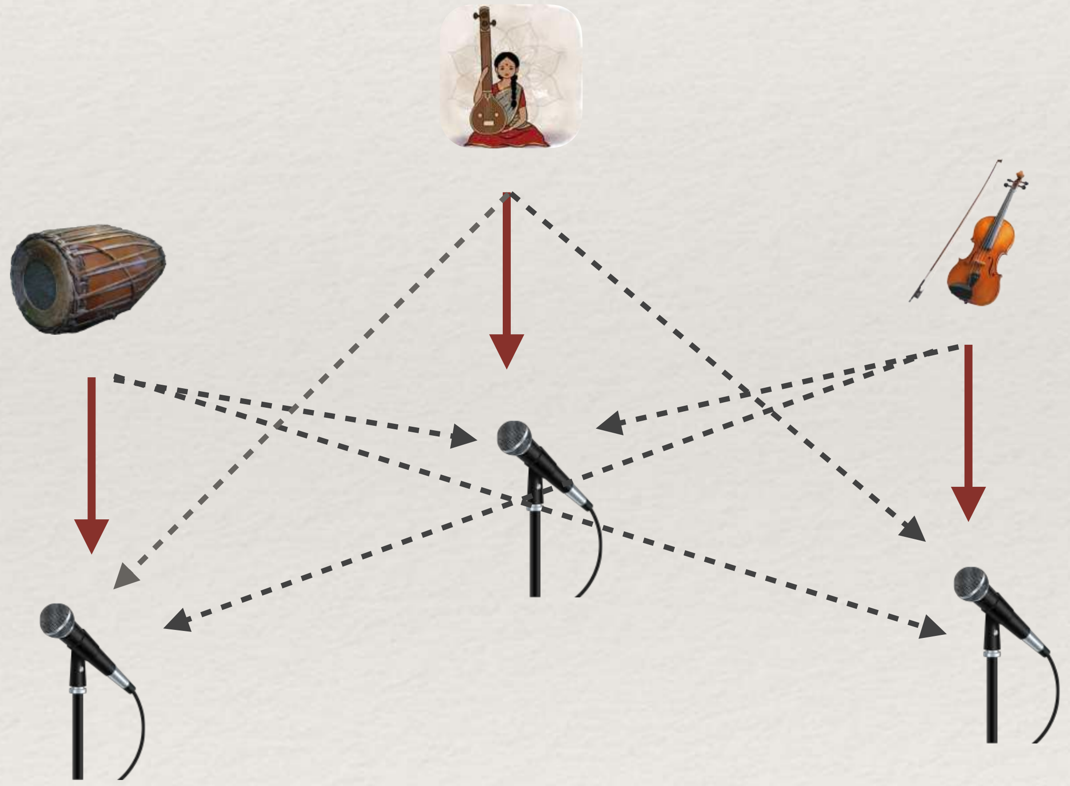
Optimisation Approach

With guidance of
Dr. Siddhartha Sarma

Equations: $X = \Lambda S$ and $m = b^T S$

Problem statement: *minimise* $\|X - \Lambda S\|^2 + \|m - b^T S\|^2$ *with respect to* Λ, S *and* b *subject to constraints:*

1. $\Lambda \neq I$
2. $\lambda_{ii} > \lambda_{ij}$
3. $\gamma_1 \leq \lambda_{ij} \leq \gamma_2, \forall i \neq j$



Alternate Minimisation Solution

- Non convex problem, global minima does not exist
- Alternate minimisation approach
- Derived the update rule for Λ, S *and* b.

Update Rules:

$$\Lambda = (XSS^T)(SS^T + \eta I)^{-1}$$

$$S = (\Lambda^T \Lambda + bb^T)^{-1}(bm + \Lambda^T X)$$

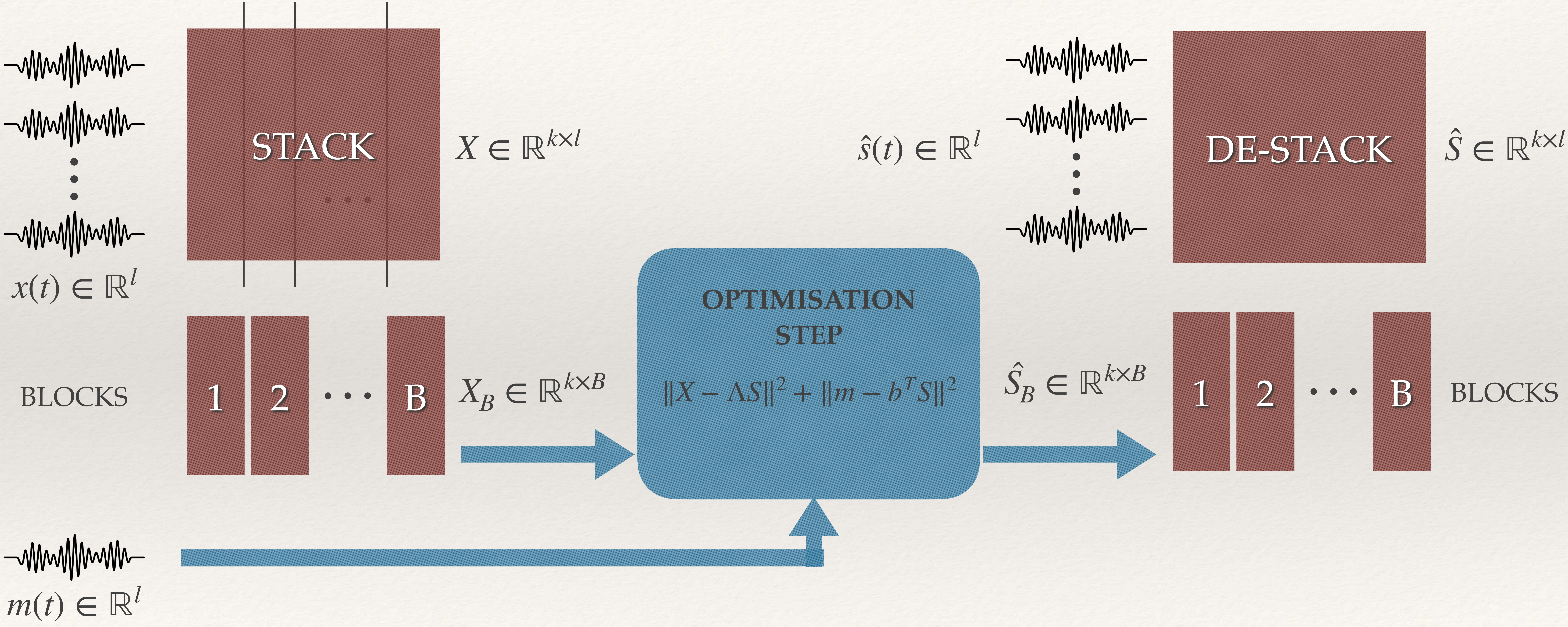
$$b = (SS^T + \eta I)^{-1}(Sm^T)$$

Algorithm

Algorithm 1 Time-domain Optimization Algorithm for Bleed Reduction

Inputs: $X \in \mathbb{R}^{k \times l}$ and $m \in \mathbb{R}^{l}$
Initialize: $\Lambda \leftarrow I$
Initialize: $S \leftarrow X$
Initialize: $b \leftarrow [1, 1, ...1]^T \in \mathbb{R}^{l}$
while $||X - \Lambda S||^2 + ||m - b^T S||^2 \geq \epsilon$ **do**
    $\Lambda \leftarrow (XSS^T)(SS^T + \eta I)^{-1}$ ▷ A update rule
    $\Lambda \leftarrow projection(\Lambda)$
    $S \leftarrow (\Lambda^T \Lambda + bb^T)^{-1}(bm + \Lambda^T X)$ ▷ S update rule
    $b \leftarrow (SS^T + \eta I)^{-1}(Sm^T)$ ▷ b update rule
end while

Overall Procedure



Results

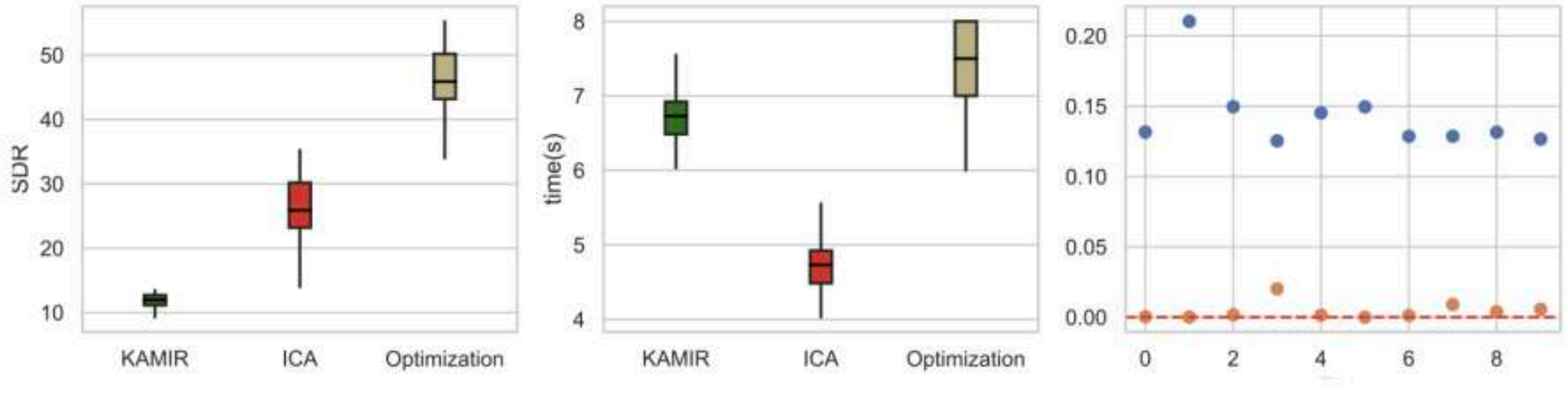


Fig. 3. SDR, time taken and the difference l2 norm is compared with ICA and KAMIR

Results

$$X = \Lambda S$$

True Λ

1	0.1	0.1	0.1
0.1	1	0.1	0.1
0.1	0.1	1	0.1
0.1	0.1	0.1	1

Predicted Λ

1	0.098	0.099	0.099
0.094	1	0.092	0.098
0.094	0.098	1	0.099
0.094	0.098	0.099	1

KAMIR Λ

1.071	0.101	0.1	0.12
0.122	1.07	0.11	0.173
0.284	0.19	1.558	0.564
0.127	0.097	0.104	1.235

Interference Matrix Λ

Shortcomings of the approach

- Linearity: Mixtures in real world follows non-linear mixing.
- High computation time.

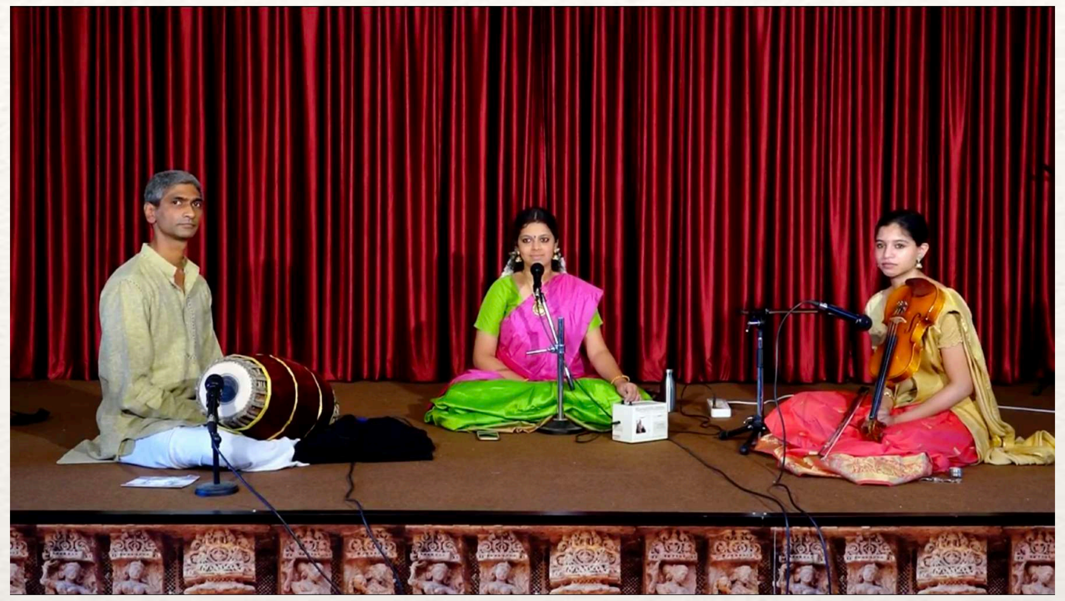
Learning based Interference Reduction

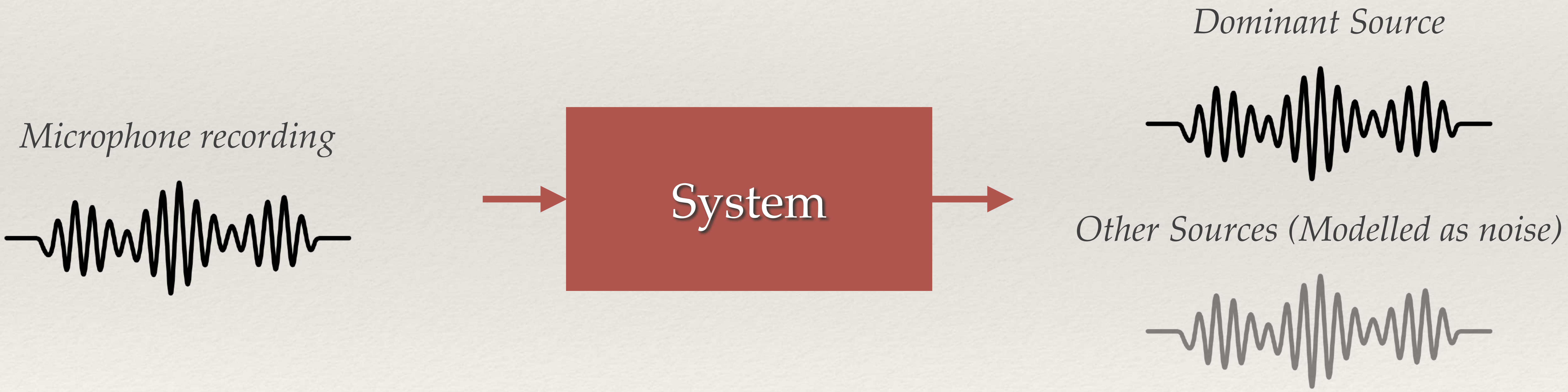
- Why?
- Datasets?
- Generalisability?

Interference as Noise

Treating interference as a noise,

$$x(t) = s(t) + n(t)$$

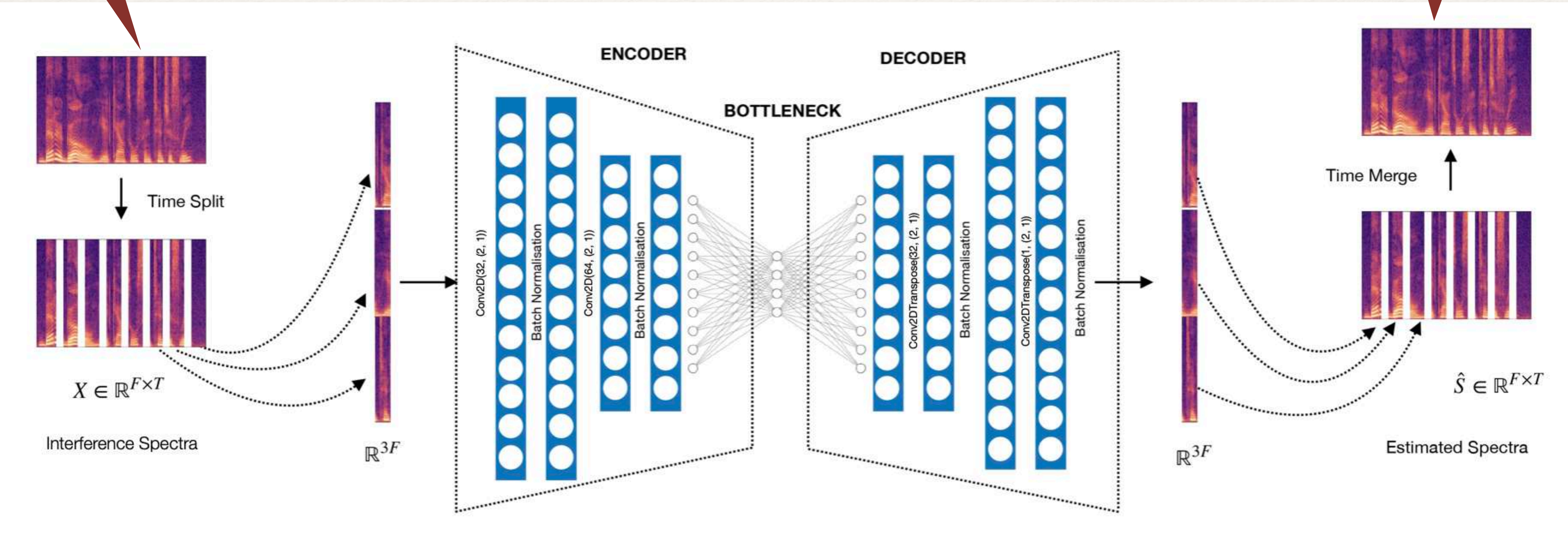




Convolutional Autoencoder (CAE)

Microphone Recordings

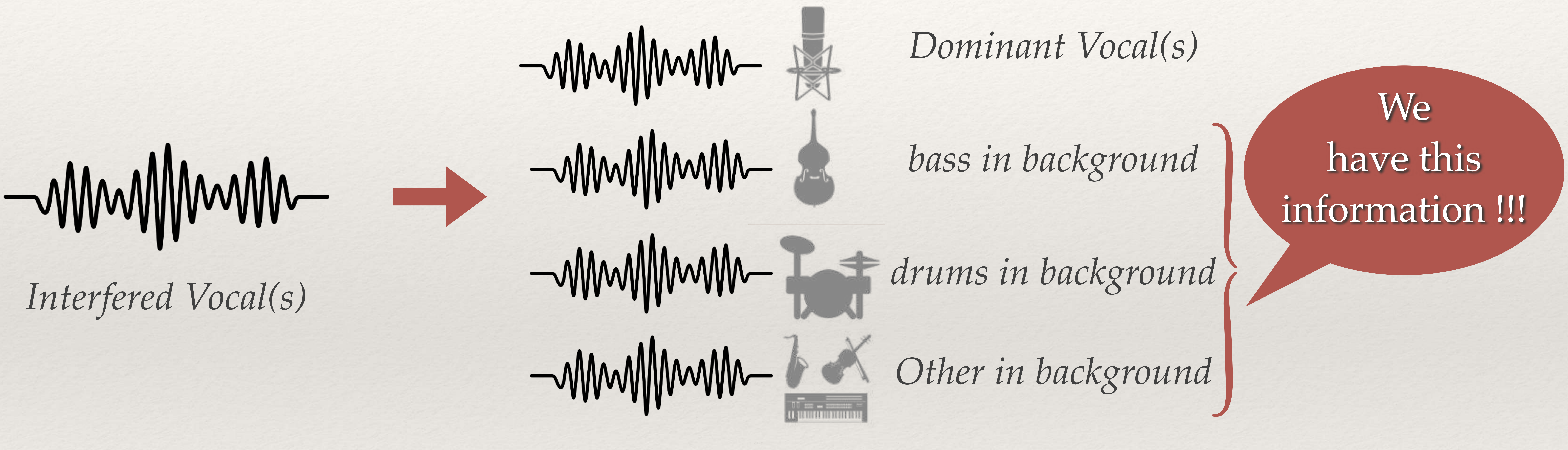
Estimated Sources



CAE Limitations

- Poor generalisability
- Thus, for each source there should be dedicated trained CAEs
- Phase information issues

Hidden Information



Interference Learning based Reduction

In general, let $X \in \mathbb{R}^{K\times L}$ be the time-aligned received by the K microphones corresponding to an audio of length L.

let $X \in \mathbb{R}^{K\times L}$ be the true sources, then the relationship between X and S can be modelled as,

$$X = \Lambda S$$

$$\Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \dots & \lambda_{1N} \\ \lambda_{21} & \lambda_{22} & \dots & \lambda_{2N} \\ \vdots & & \vdots & \\ \lambda_{K1} & \lambda_{K2} & \dots & \lambda_{KN} \end{pmatrix} \quad X = \begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_K(t) \end{pmatrix} \quad S = \begin{pmatrix} s_1(t) \\ s_2(t) \\ \vdots \\ s_N(t) \end{pmatrix}$$

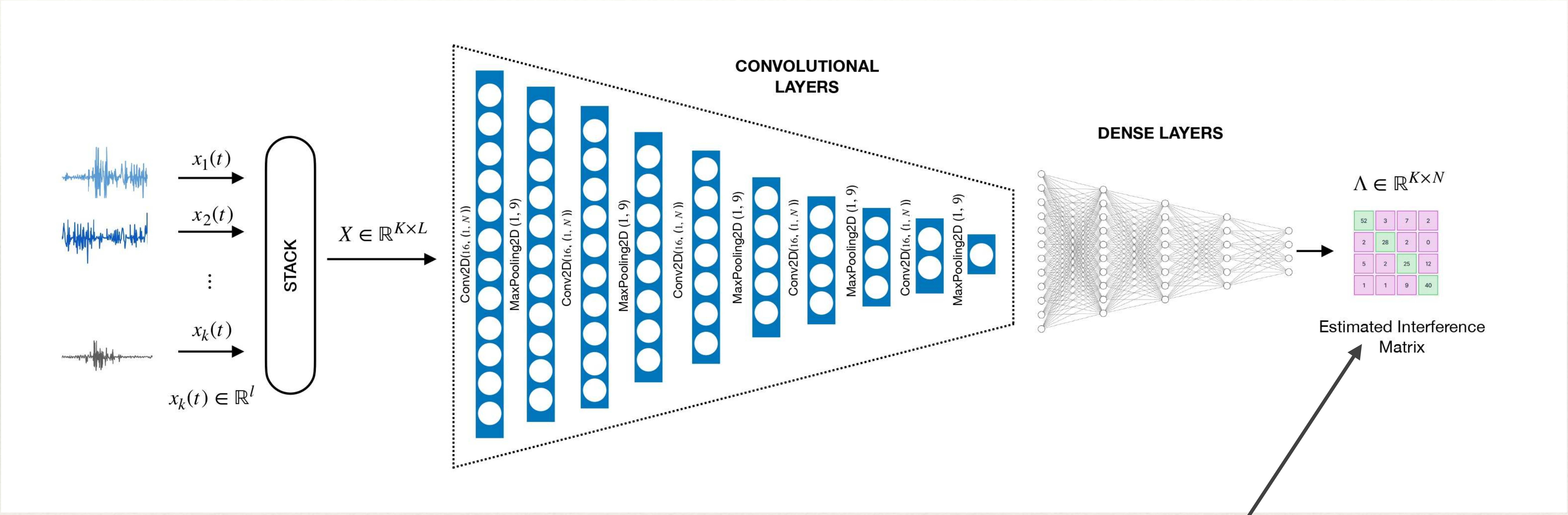
Interference Learning based Reduction

The interference reduced sources can be estimated by,

$$\hat{S} = \Lambda^{\dagger} X$$

Where $\dagger$ is the pseudo inverse of Λ.

t-UNet Architecture



$$X = \Lambda S$$

Datasets

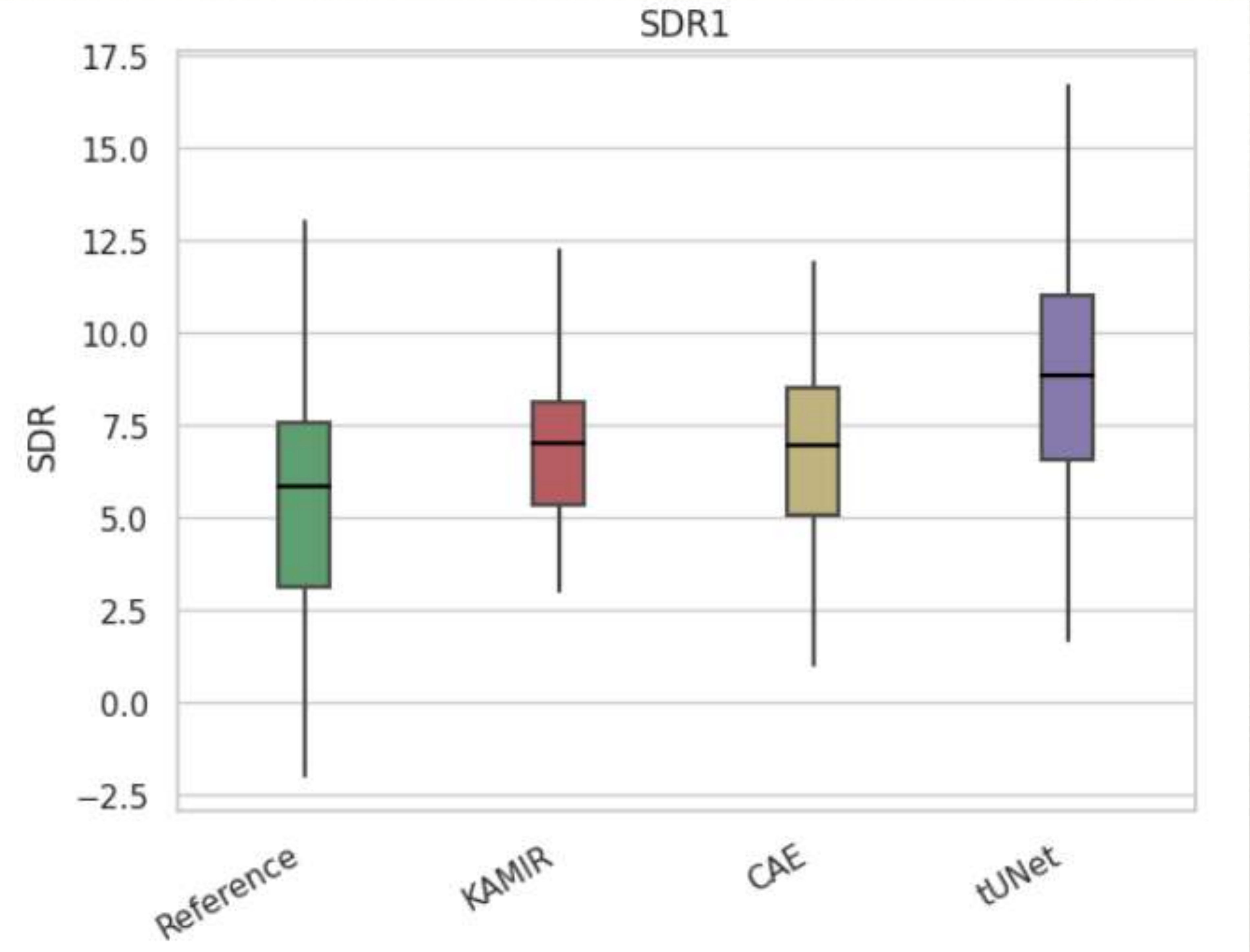
- Artificially created the bleeding with MUSDB18HQ[1] dataset
- **MUSDB**: Linear Mixtures - Mixup the stem within the track using randomly generated interference matrix Λ
- **MUSDBR**: Convolute Mixtures: Introducing room impulse responses and time delays using pyroomacoustics[2]

---

[1] *Z. Rafii, A. Liutkus, F.-R. Stoter, S. I. Mimilakis and R. Bittner, "Musdb18-HQ - an uncompressed version of MUSDB18," Aug. 2019. [online] Available: https://doi.org/10.5281/zenodo.3338373.*

[2] *R. Scheibler, E. Bezzam, and I. Dokmani´c, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) IEEE, 2018, pp. 351–355.*

Results



LINEAR MIXTURES

Fig: SDR for the proposed models compared with KAMIR[3] under linear mixtures dataset

[3] *T. Pratzlich, R. M. Bittner, A. Liutkus, and M. Muller, "Kernel additive modeling for interference reduction in multi-channel music recordings," in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015, pp. 584–588*

Results

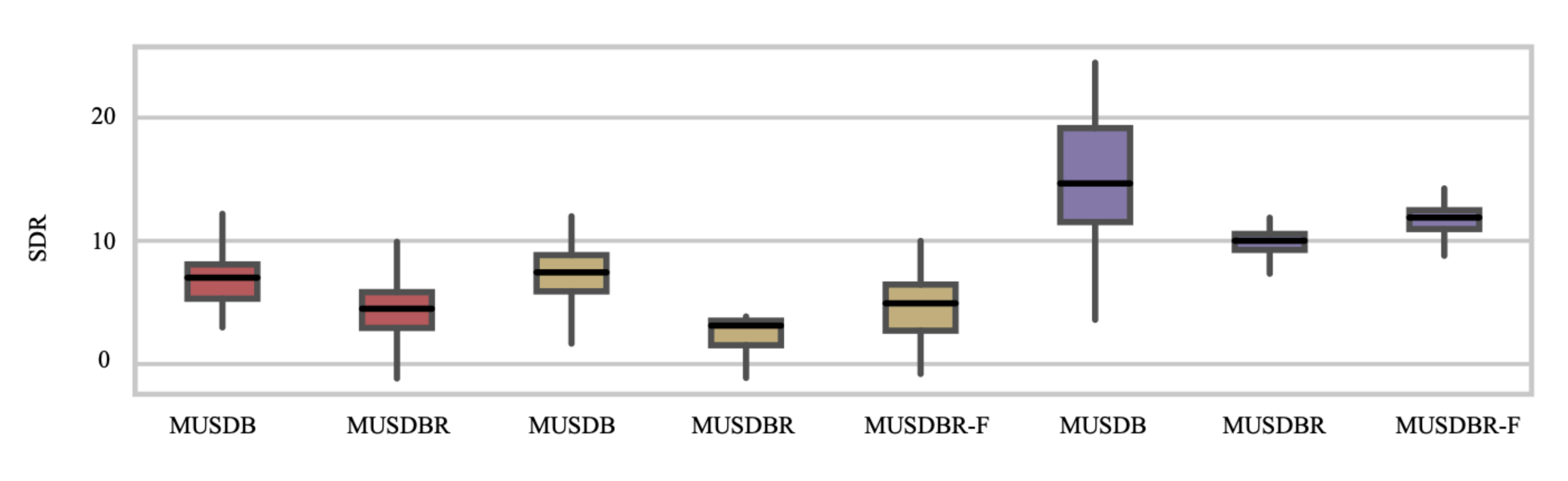


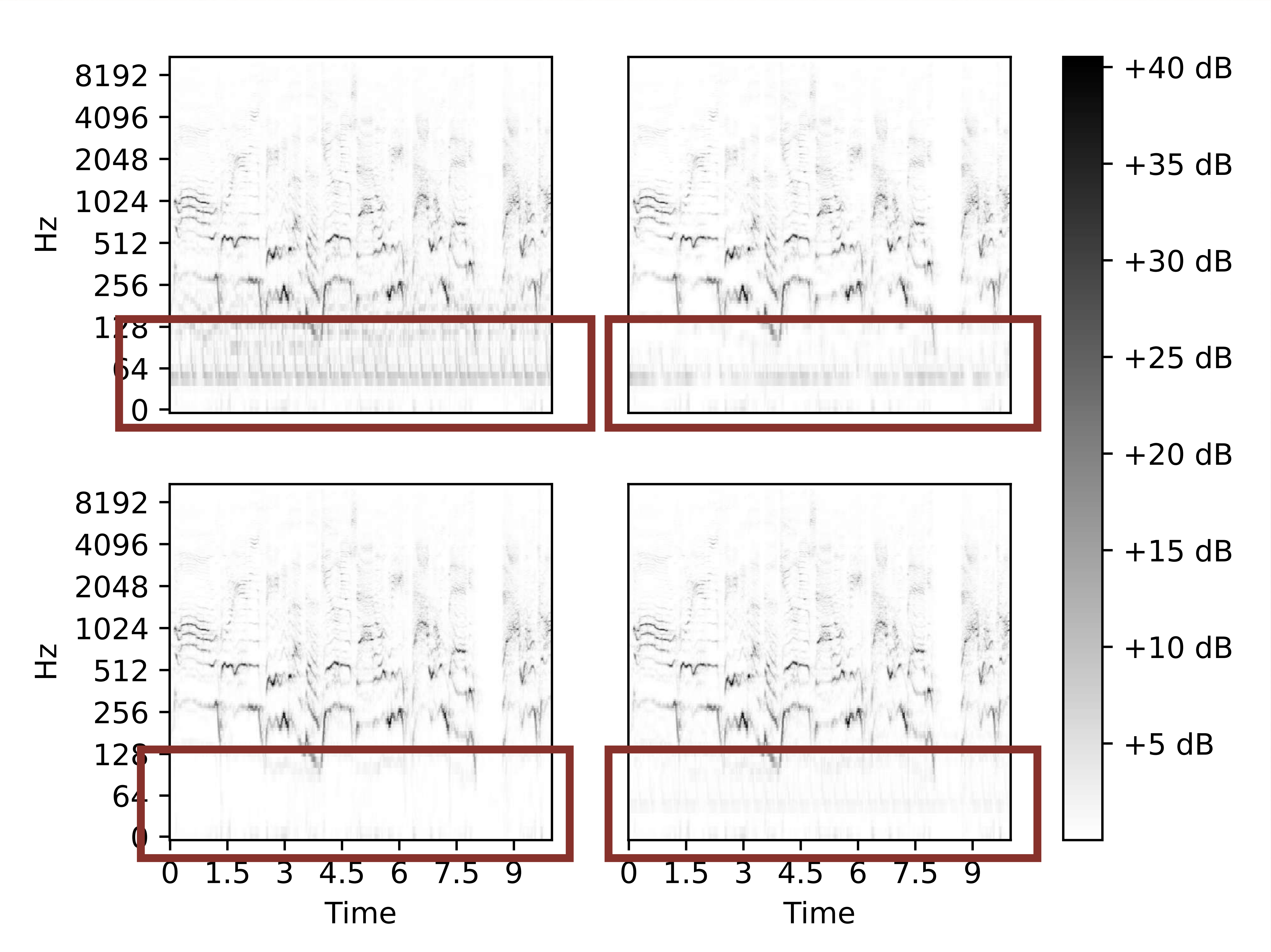
Fig: Average SDR for the proposed models with convolute mixtures under matched and mismatched case

KAMIR, CAE, and t-UNet are represented in Red, Yellow, and Magenta respectively.

Suffix F represents models fine-tuned with MUSDBR

CONVOLUTE MIXTURES

Results



8192
4096
2048
1024
512
256
128
64
0
Hz
0 1.5 3 4.5 6 7.5 9
Time
+40 dB
+35 dB
+30 dB
+25 dB
+20 dB
+15 dB
+10 dB
+5 dB

Results

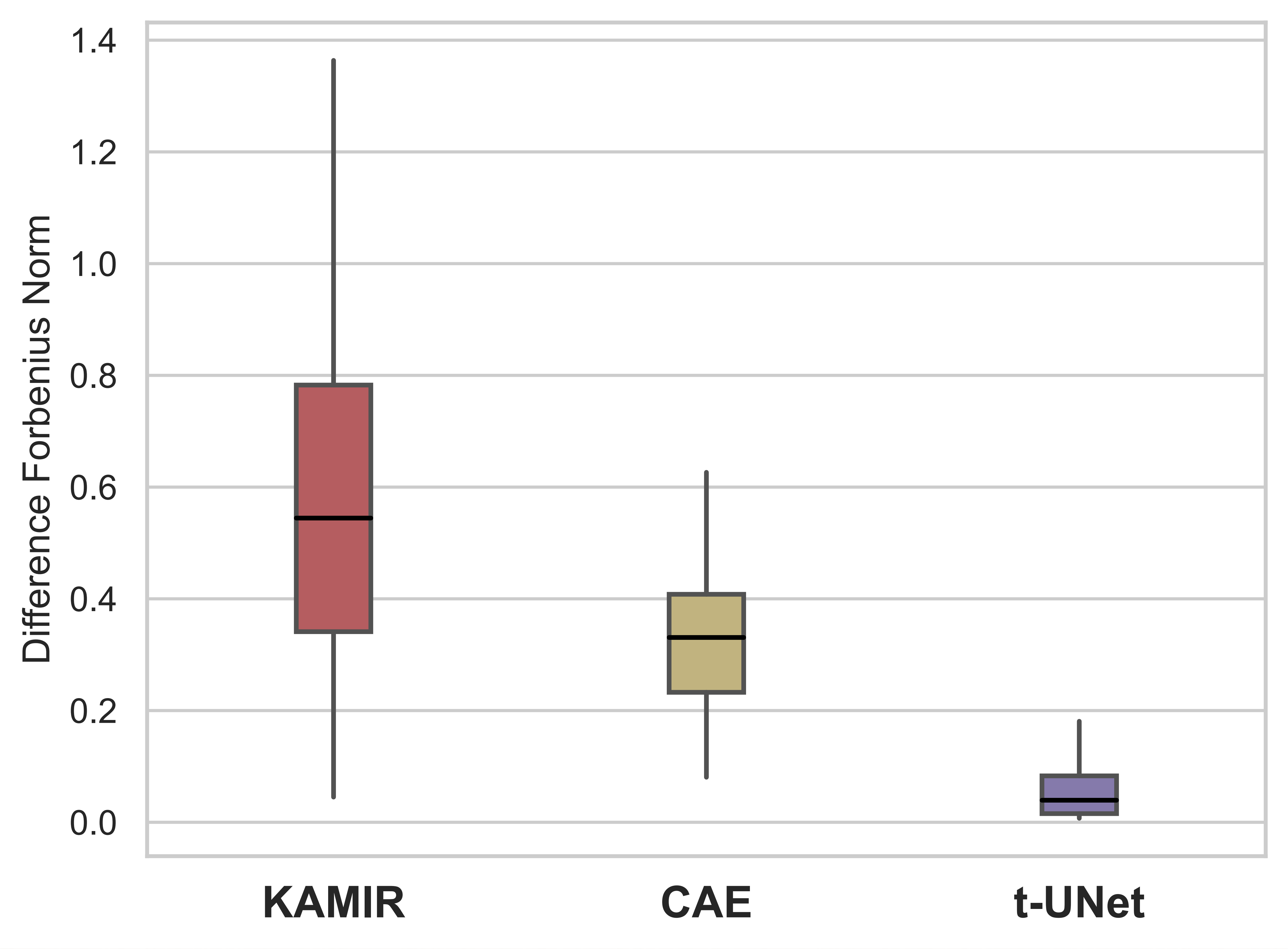


Fig: Difference of Frobenius norm of the true Λ with the predicted $\hat{\Lambda}$.

MSS Performance

On Wave-U-Net with MUSDB18HQ dataset,

	Clean	Interference	CAE Cleaned	t-UNet Cleaned
SDR	2.32	0.96	1.72	2.03

Table: Music Source Separation Performance

Computational Complexity:

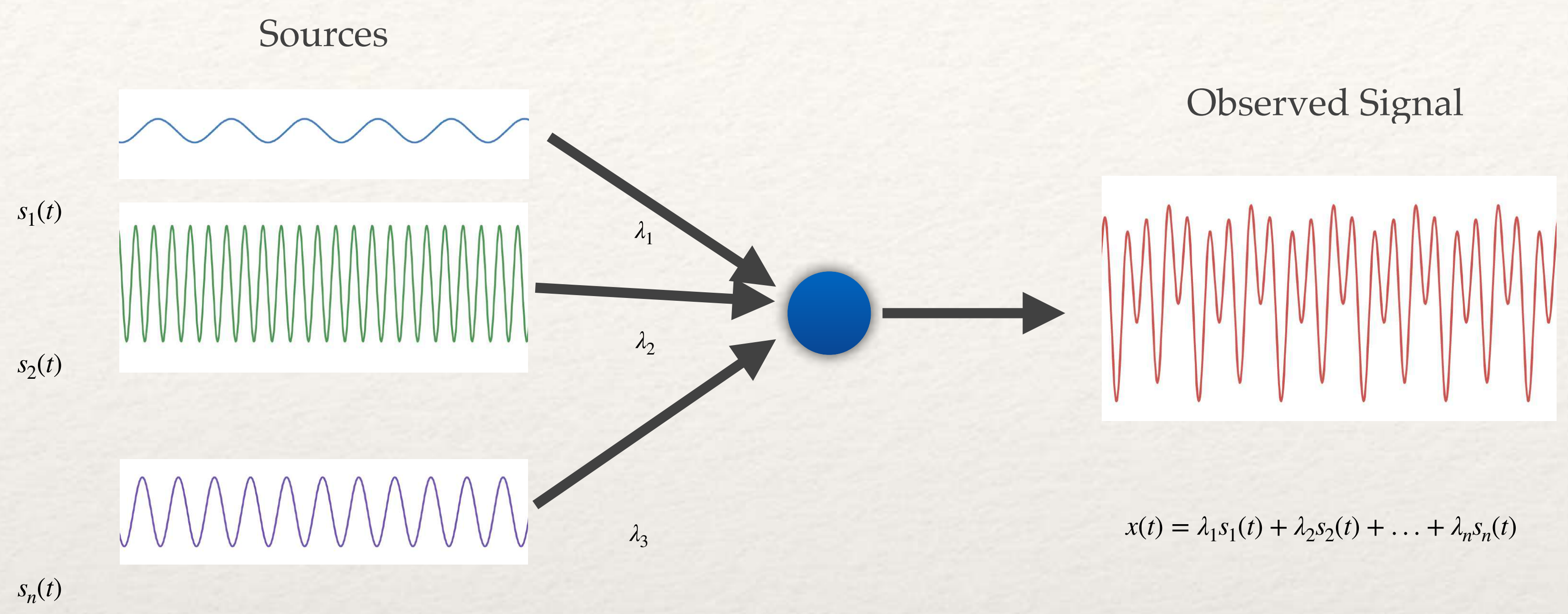
	KAMIR	CAEs	tUNet
Average	660.4	2.4	2.19

Table: Time taken in seconds for 100 test tracks

Conclusion of CAEs & t-UNet

- Proposed two neural networks for interference reduction: CAEs and t-UNet, both performing better than KAMIR
- CAEs has difficulties in generalising and works in TF domain where t-UNet reduces interference directly by learning interference matrix.
- t-UNet outperforms all the models in-terms of SDR and computationally faster
- Interference reduction improves the source separation performance

Disadvantages



- tUNet built with the mathematical approximation of the problem as $X = \Lambda S$ which is still **linear!**
- Initial evaluations of the live recordings reveals the t-UNet is not effective.

Acoustic Treated vs Normal Room

https://images.app.goo.gl/oMMMJN7VJ4inwNnq8

https://images.app.goo.gl/65HCSCiKP55FfWVMA

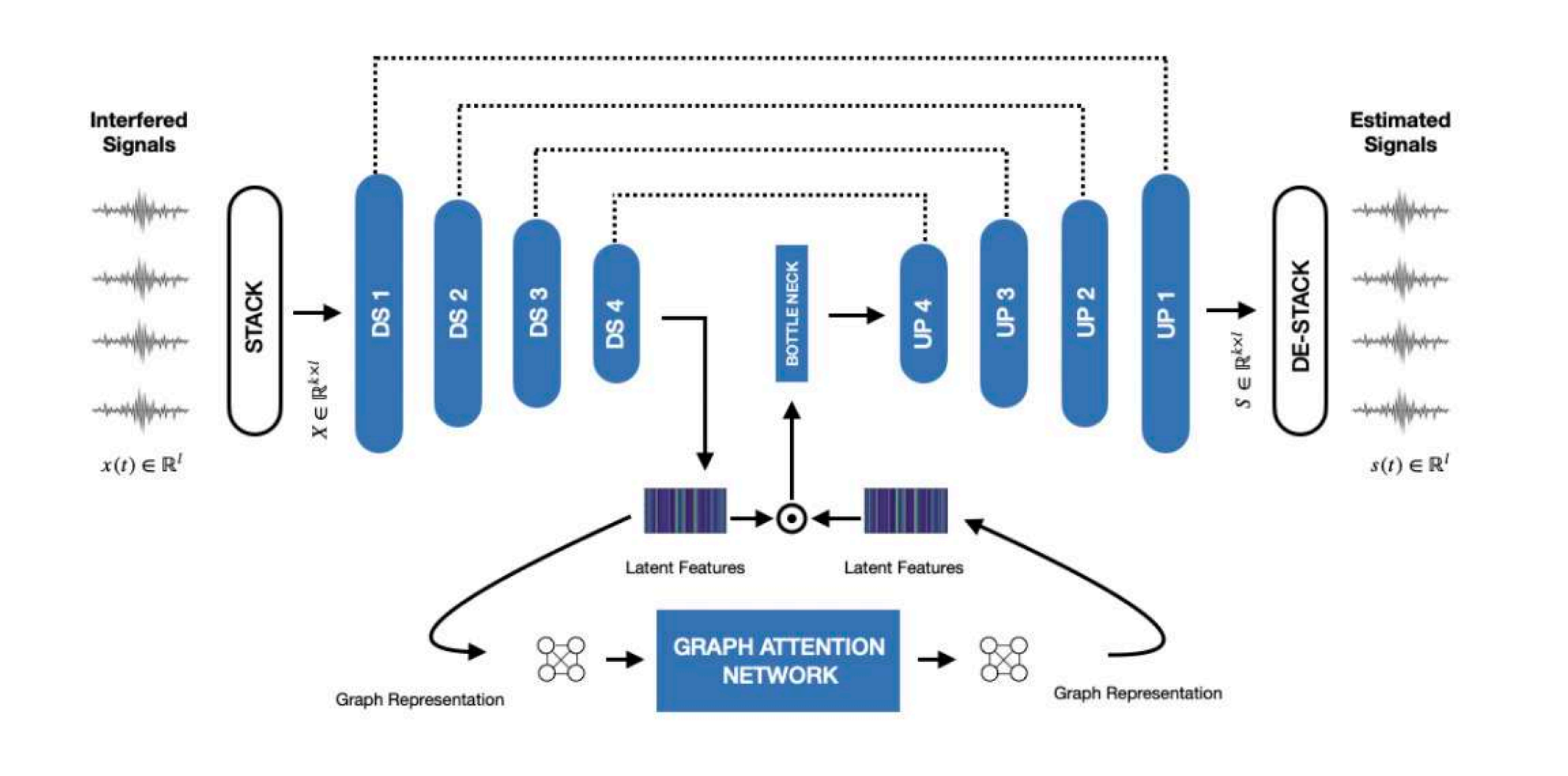
Extending the problem to Non-linearity

For k microphones and n sources,

$$
\begin{aligned}
x_1(t) &= f(\boldsymbol{s_1(t)}, s_2(t), \ldots, s_n(t)) \\
x_2(t) &= g(s_1(t), \boldsymbol{s_2(t)}, \ldots, s_n(t)) \\
&\vdots \\
x_k(t) &= h(s_1(t), s_2(t), \ldots, \boldsymbol{s_n(t)})
\end{aligned}
$$

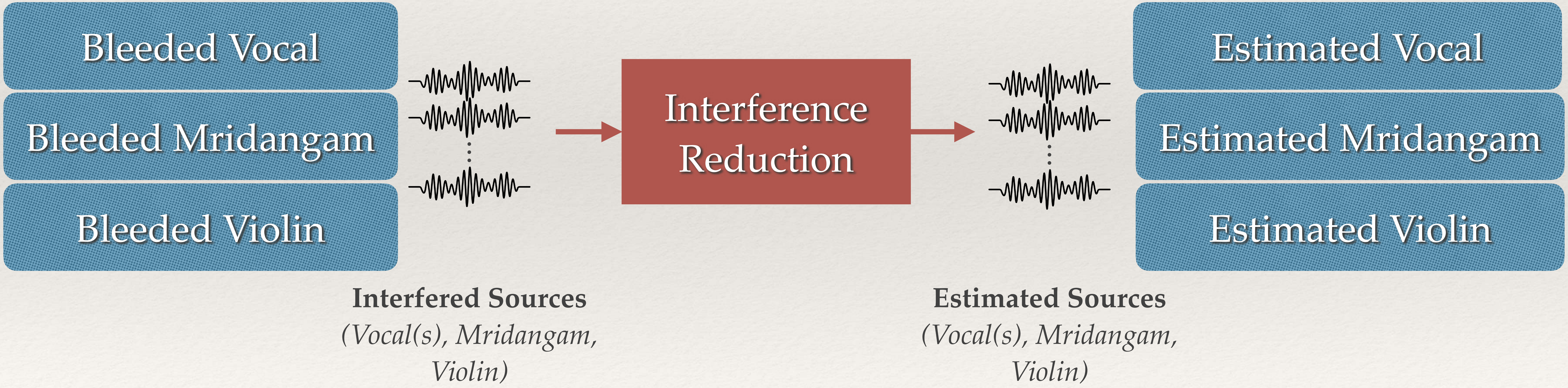
Where $f(\,.\,)$, $g(\,.\,)$, and $h(\,.\,)$ are some unknown functions

Dilated Wave-U-Net with Graph Attentions

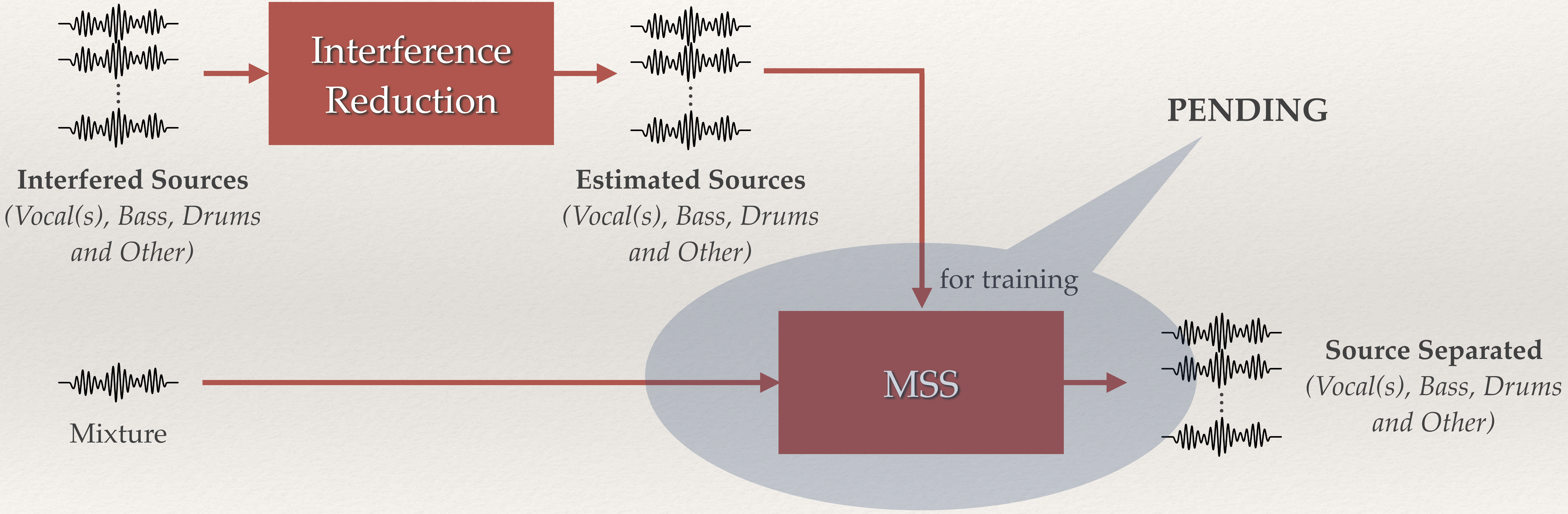


Testing on live recordings

- The Saraga Dataset: Vocal(s), mridangam, and violin
- Extending to out-of-domain samples thru post processing



Future work



Interference Reduction
Interfered Sources
(Vocal(s), Bass, Drums and Other)
Estimated Sources
(Vocal(s), Bass, Drums and Other)
PENDING
for training
MSS
Mixture
Source Separated
(Vocal(s), Bass, Drums and Other)

Publications

- Rajesh R and Padmanabhan Rajan, "Neural Networks for Interference Reduction in Multi-Track Recordings," *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2023, pp. 1-5.

“Thank you all for your time and attention”

Indian Institute of Technology Mandi