IEEE WASPAA 2023

Rajesh R, Padmanabhan Rajan. Indian Institute of Technology, Mandi.

Neural Networks for Interference Reduction in Multi-track Recordings

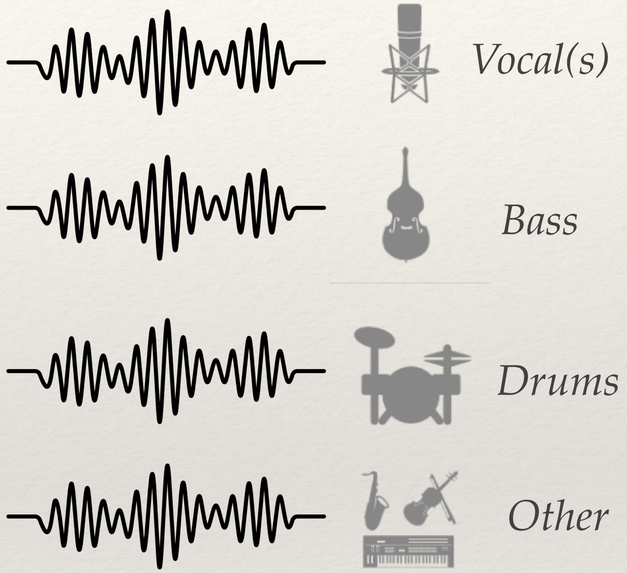
2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 2023, pp. 1-5

Music Source Separation

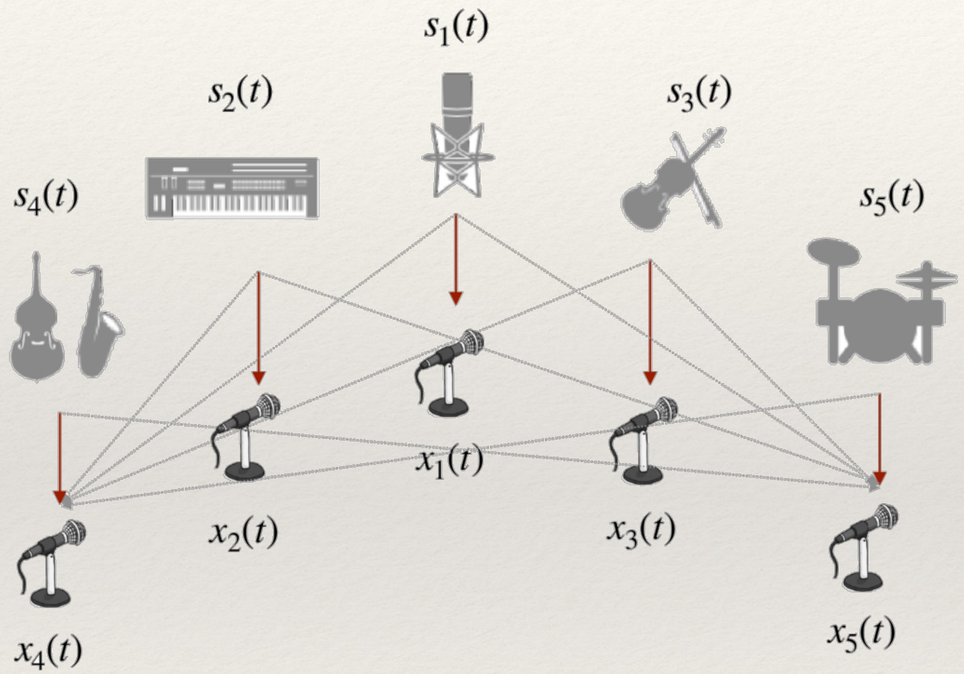
Music

(Vocal(s), Bass, Drums and Other)

MSS System



Interference Effects



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

Assumptions

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

Interference as Noise

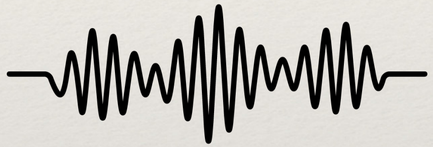
Treating interference as a noise,

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

Microphone recording

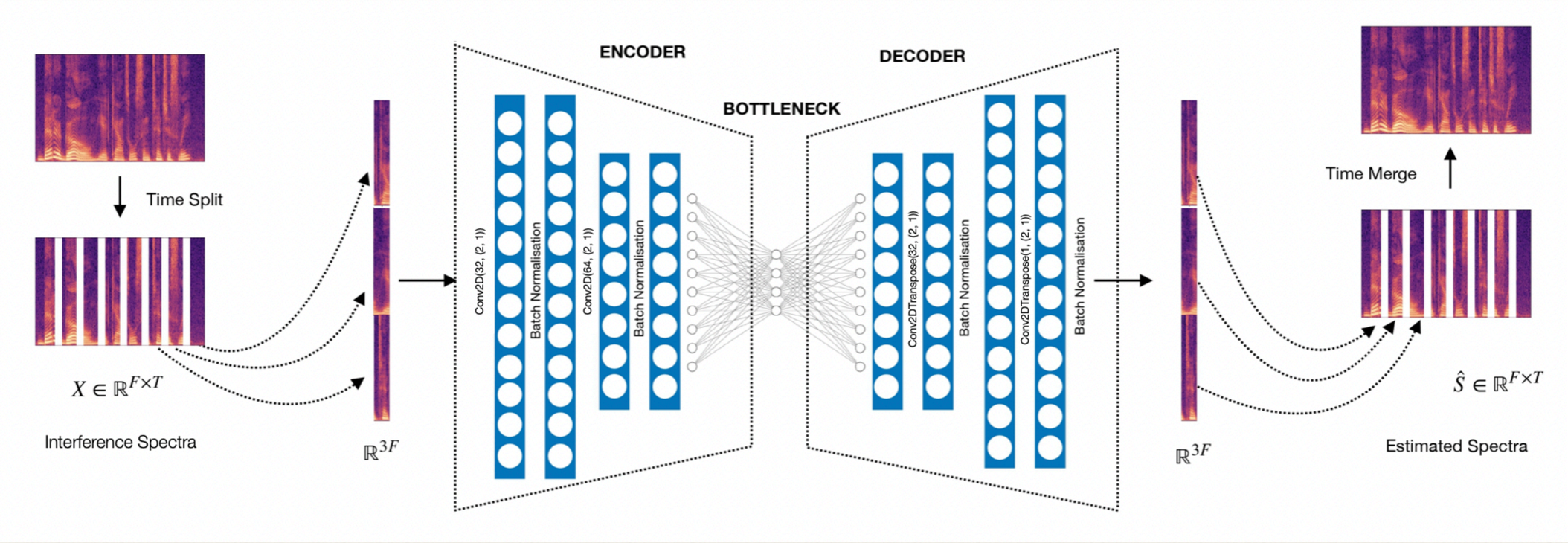
System

Dominant Source



Other Sources (Modelled as noise)

Convolutional Autoencoder (CAE)

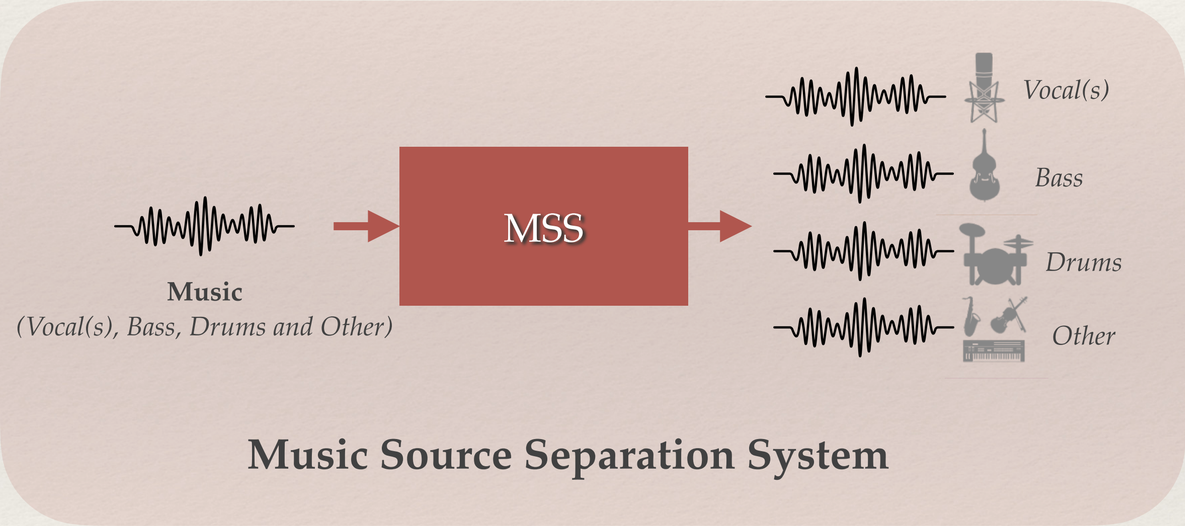


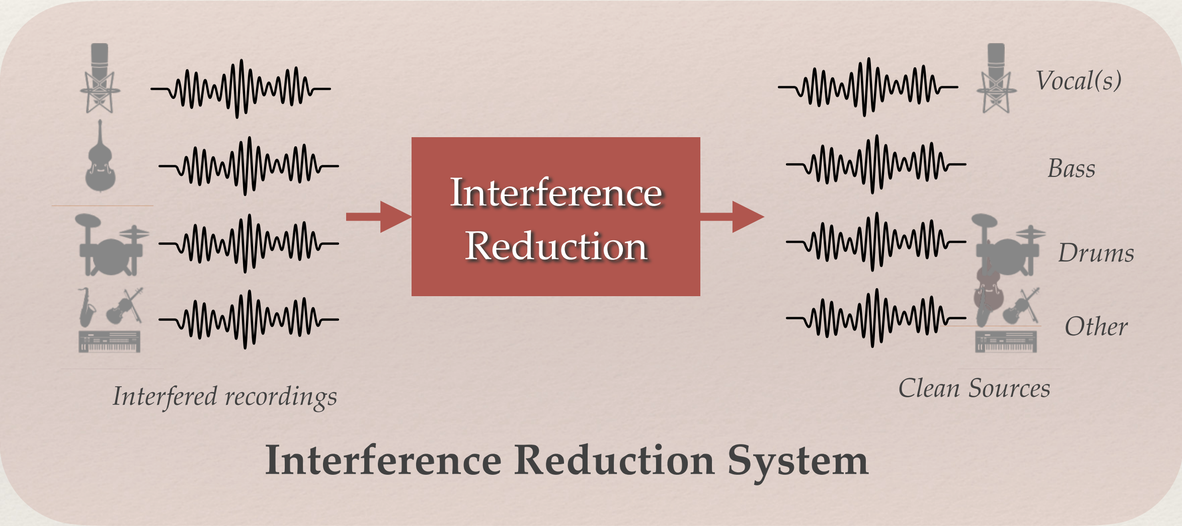
CAE Limitations

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

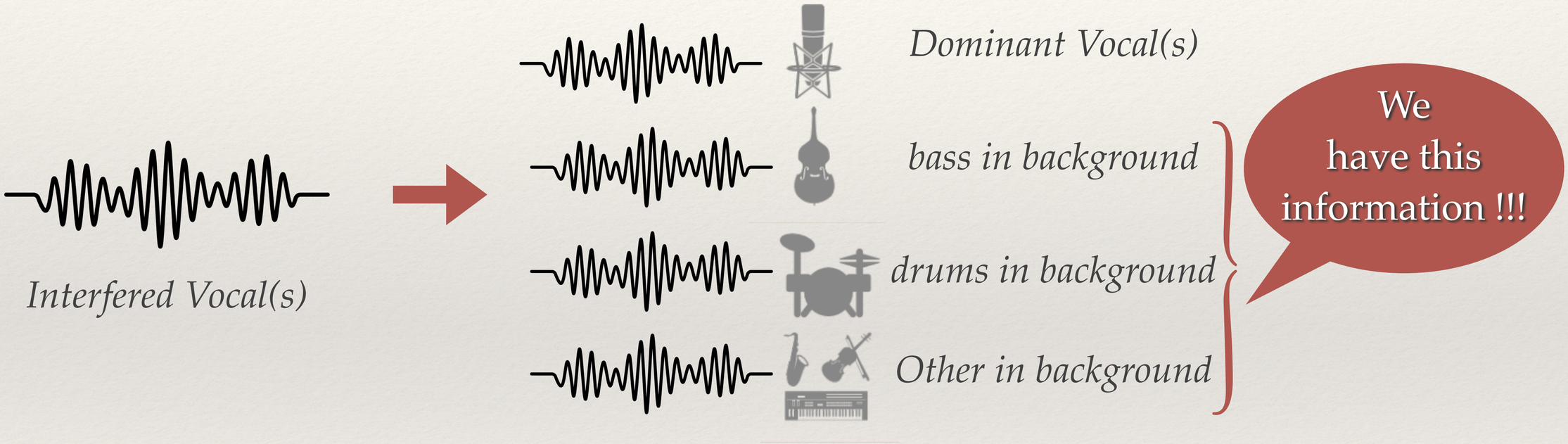
Interference Reduction is MSS

- Interference reduction problem is a special type of source separation
- We have interfered sources and the goal is to clean them





Interference Reduction is MSS



Interference Learning based Reduction

Let us have K microphones capturing N sources,

$$x_k(t) = \lambda_{k1}s_1(t) + \lambda_{k2}s_2(t) + \ldots + \lambda_{kN}s_N(t)$$

λ_{kn} represents gain of the acoustic path from n^{th} source to k^{th} microphone

$s_n(t)$ represents gain of the n^{th} true source

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,

$$X = \Lambda S$$

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

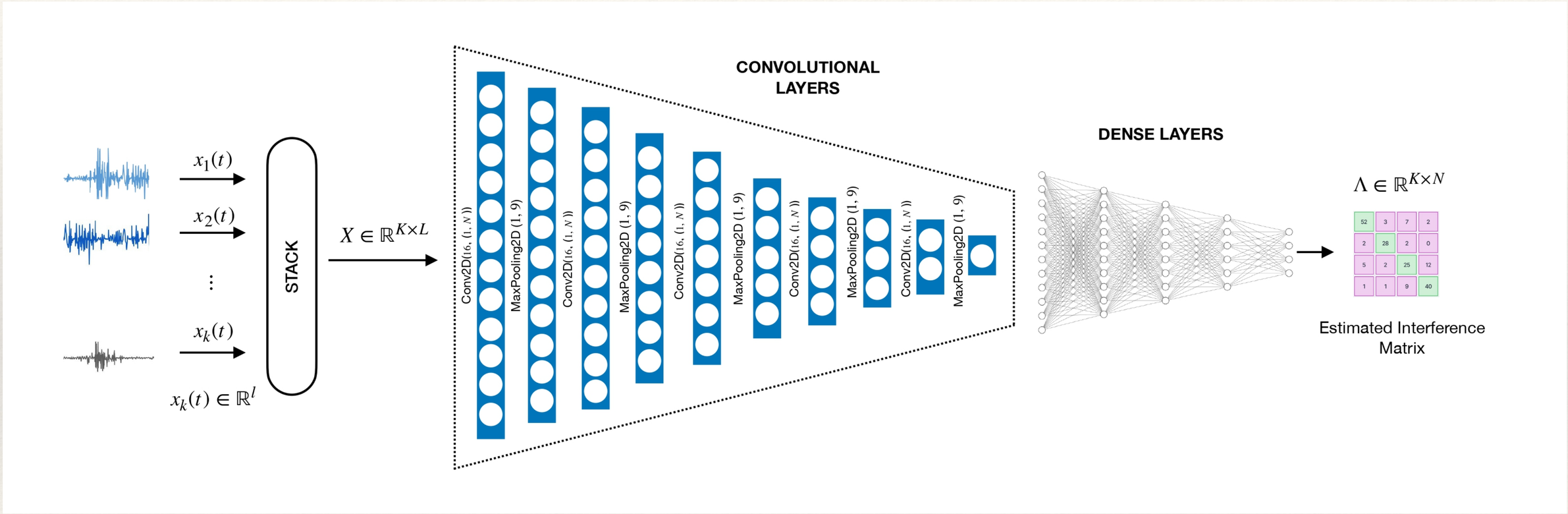
Interference Learning based Reduction

The interference reduced sources can be estimated by,

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

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

t-UNet Architecture



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. Dokmanić, "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

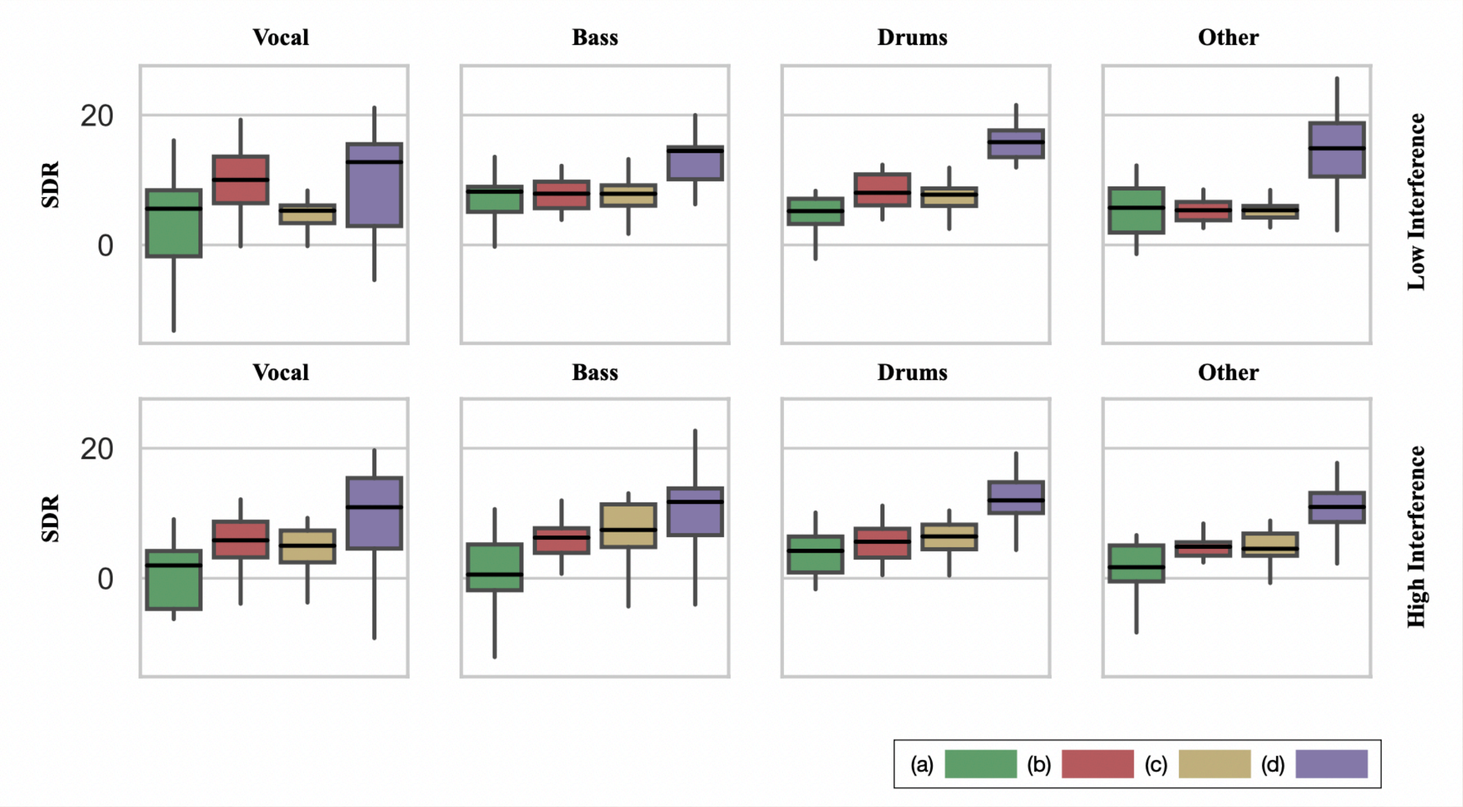


Fig: SDR for the proposed models compared with KAMIR[3] under linear mixtures dataset on high and low interference conditions.

a) Reference SDR, (b) KAMIR, (c) CAE, and (d) t-UNet

[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

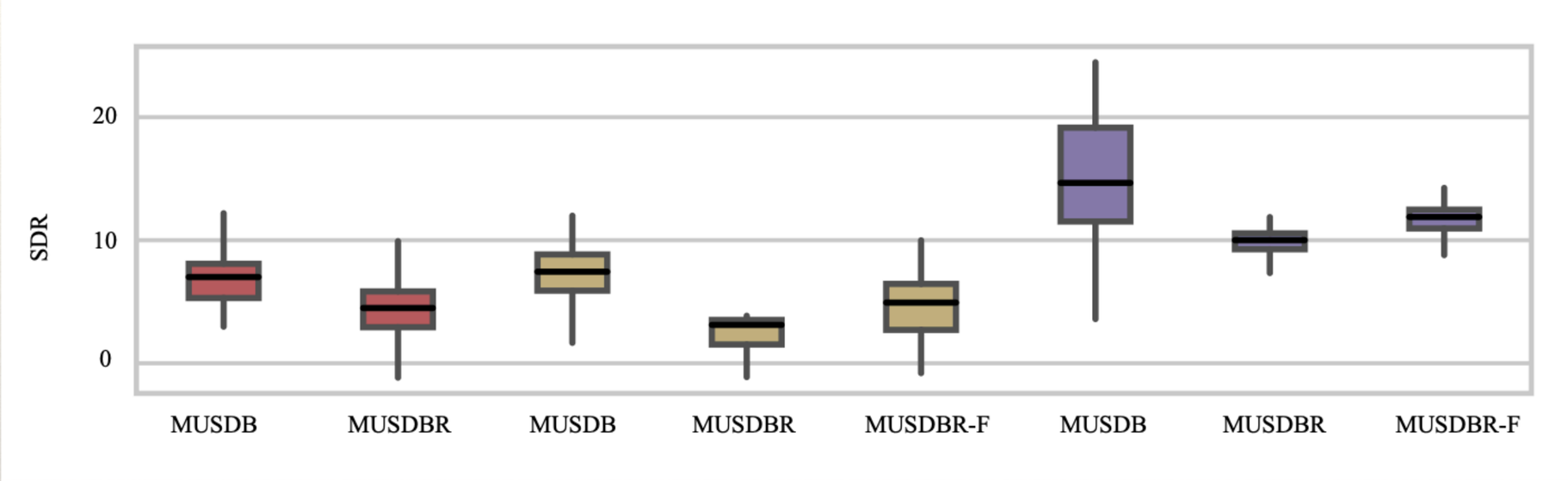


Fig: Average SDR for the proposed models with convolute mixtures under matched and mismatched case

KAMIR, CAE, and t-UNet represented in Red, Yellow, and Magenta respectively.
Suffix F represents models fine-tuned with MUSDBR

Results

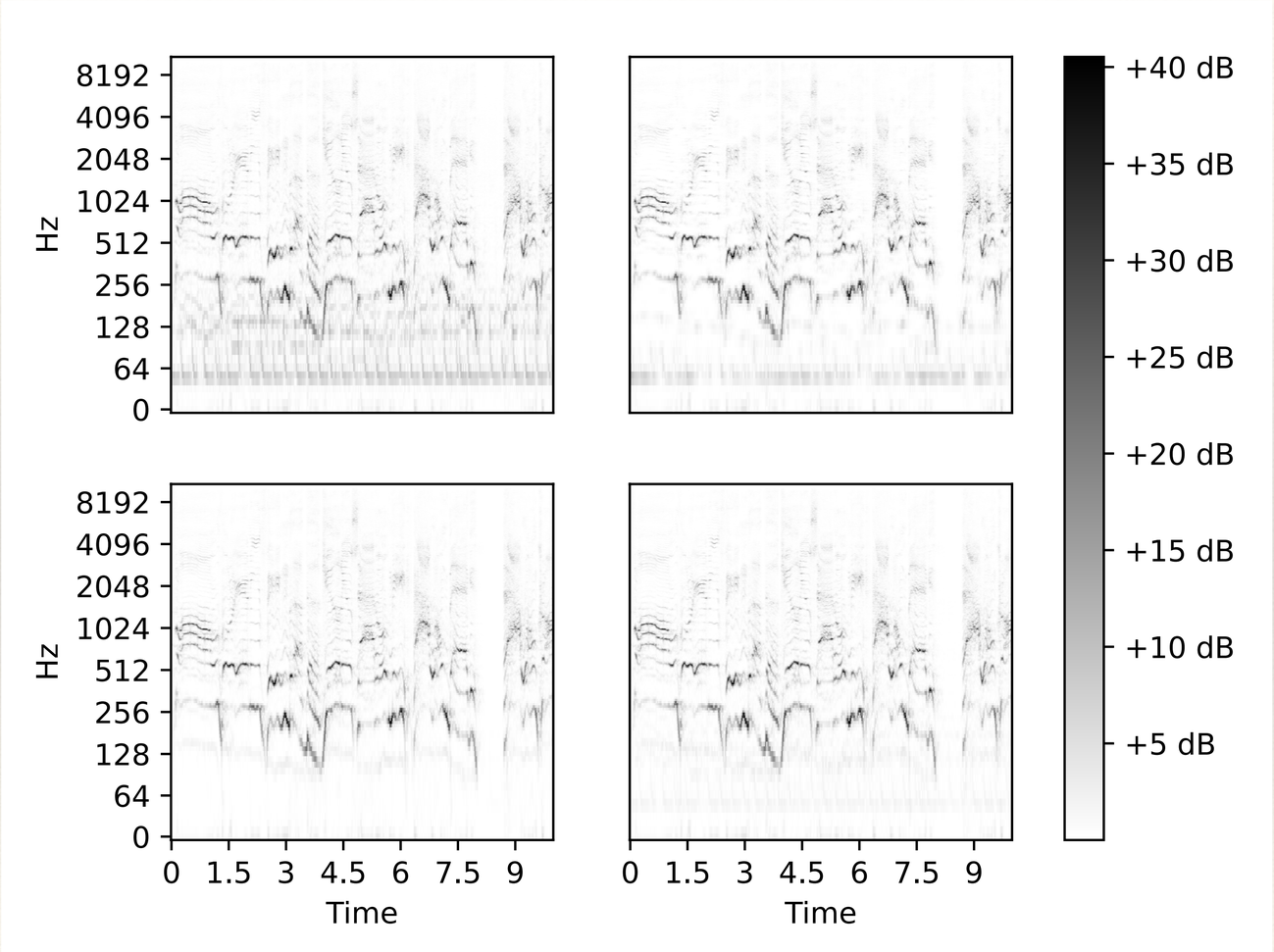


Fig: Spectrogram for a specific vocal source. From top left clockwise: vocal with interference from bass, drums and others; KAMIR prediction; CAE prediction; and t-UNet prediction.

Results

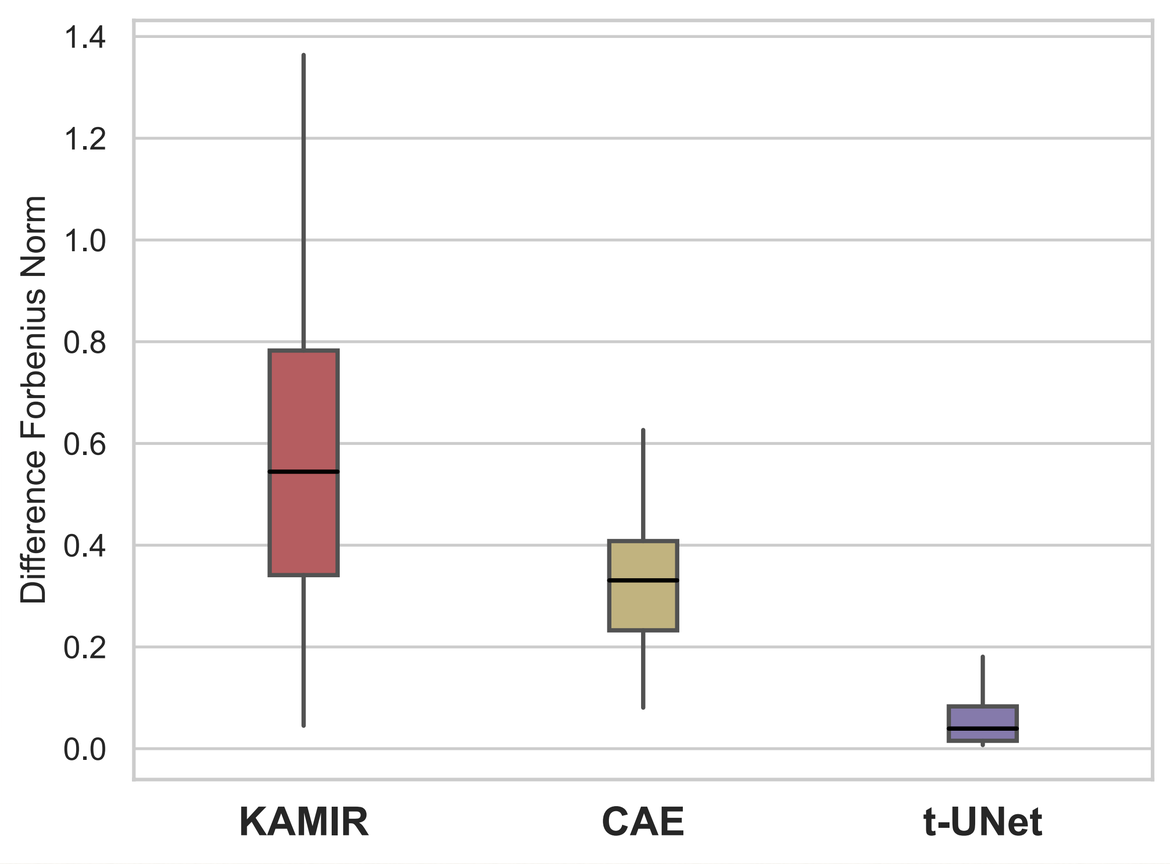


Fig: Difference of Frobenius norm of the true Λ with the predicted $\hat{\Lambda}$.

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

- 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

“Thanks for your attention”