Model-based audio deep learning

with application to source separation and dereverberation

Gaël RICHARD*

Professor, Telecom Paris, Institut polytechnique de Paris
Scientific director of Hi! PARIS

Conversational AI Reading Group, MILA

*work with collaborators and in particular H. Bai, L. Daudet, L. Bahrman, M. Fontaine, K. Shulze-Forster, B. Torres, G. Peeters,..

With support from *the European Union (ERC, HI-Audio - Hybrid and Interpretable Deep neural audio machines, 101052978).*
Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them*

Content

- A bit about IP Paris and Hi! PARIS
- Our Research group: ADASP
- Hybrid (or model-based) deep learning
- Applications in (unsupervised) music source separation
- Applications in (unsupervised) Dereverberation

Hi! PARIS

PARIS ARTIFICIAL INTELLIGENCE FOR SOCIETY

Hi! PARIS | Center in **Data Science** & **AI** for Science, Business & Society

Hi! PARIS is a multidisciplinary center dedicated to AI and Data Science at the service of **Science**, **Business** and **Society**

Created in **September 2020** by two leading institutions

Joined by **Inria** in **2021**

Backed by leading **corporate donors**

In **2024**, CNRS and UTT joined Hi! PARIS as the center was officially labeled an **AI Cluster** by the french state, securing **€70 million in funding**

Hi! PARIS: Recognized as **a French AI Cluster**

In 2024, Hi! PARIS was designated as one of the nine French AI Clusters, accelerating its growth.

FRANCE 2030

Institut Polytechnique de Paris

70 M€

With this momentum, now is the time to go further!

RESEARCH

250
Faculty members in AI & Data Science

41
Chairs have been funded since 2020
Boosting international attractiveness

13
ERC in AI (active in 2025)

+430
Articles in top-tier journals and conferences in AI

EDUCATION

+250
PhD students in AI & Data Science

8 Top-tier partner schools and universities

2,300
Students involved since 2021 in cross-disciplinary AI/data activities

INSTITUT POLYTECHNIQUE DE PARIS

#1 In France
#10 Worldwide
Graduate Employability (QS 2024)

#41
Worldwide
QS World University Rankings (2026)

HEC PARIS

#2 European Business School (FT 2025)

#2 Executive Education Worldwide (FT 2025)

#1 MSc Data Science for Business X-HEC in Europe (QS 2025)

INNOVATION

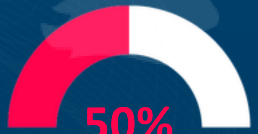
An engineering team to bridge research and development

50+
AI projects delivered

15
Open-source packages

7
Tools built with researchers

(NLP, computer vision, anomaly detection, graphs, audio, deep learning...)



50%
of the French unicorn-founders are alumn from our institutions

171
Startups in AI are founded, incubated, or accelerated within our entrepreneurial ecosystem

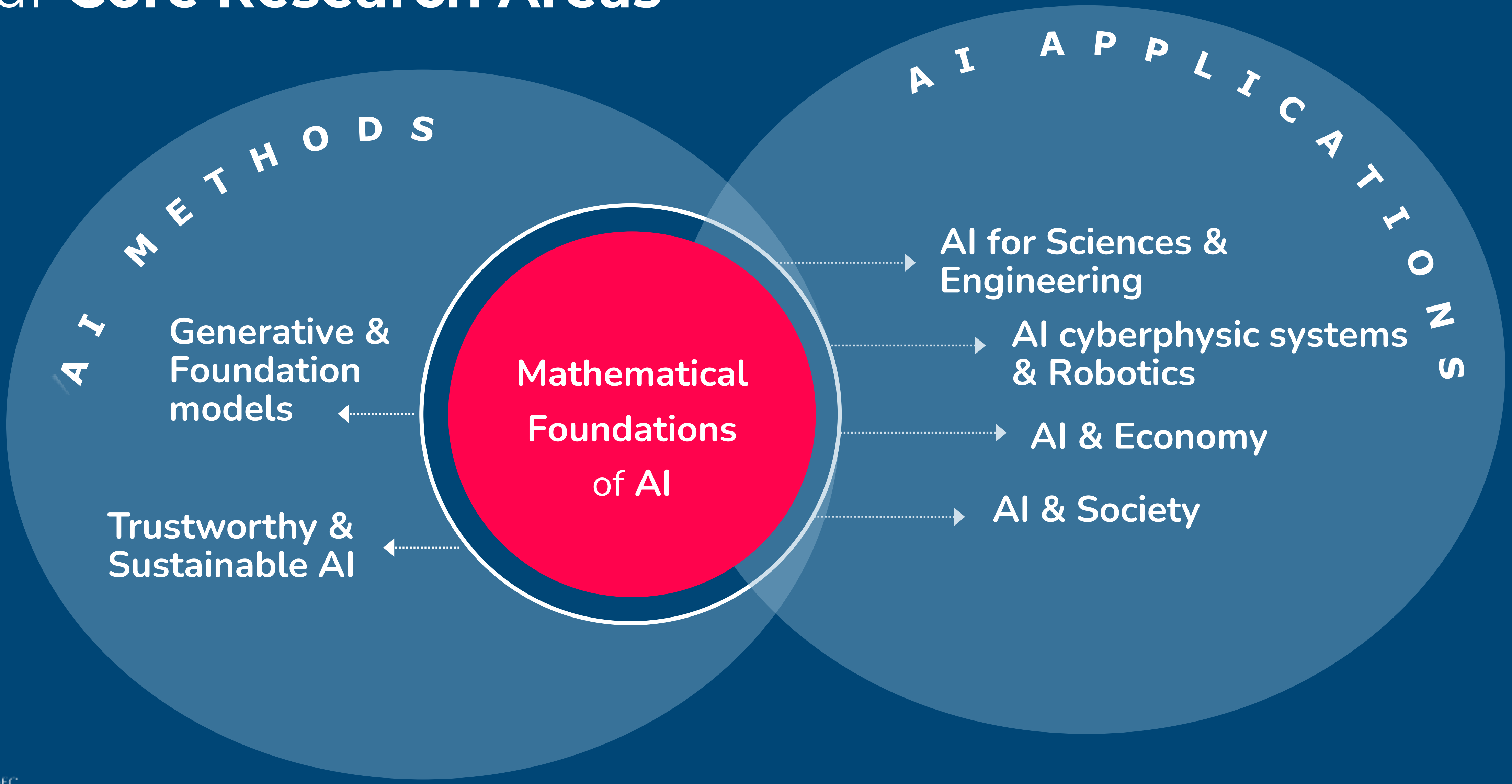
SOCIETY

High-impact public initiatives & events around AI and society
Combining debate, outreach, and inclusion

Non profit
HEC PARIS
&
Public status
INSTITUT POLYTECHNIQUE DE PARIS

Key societal AI priorities
AI in education, AI & democracy, future of work, and ethics...

Our **Core Research Areas**



The ADASP research Group

Audio **D**ata **A**nalysis and **S**ignal **P**rocessing @ Télécom Paris

https://adasp.telecom-paris.fr/

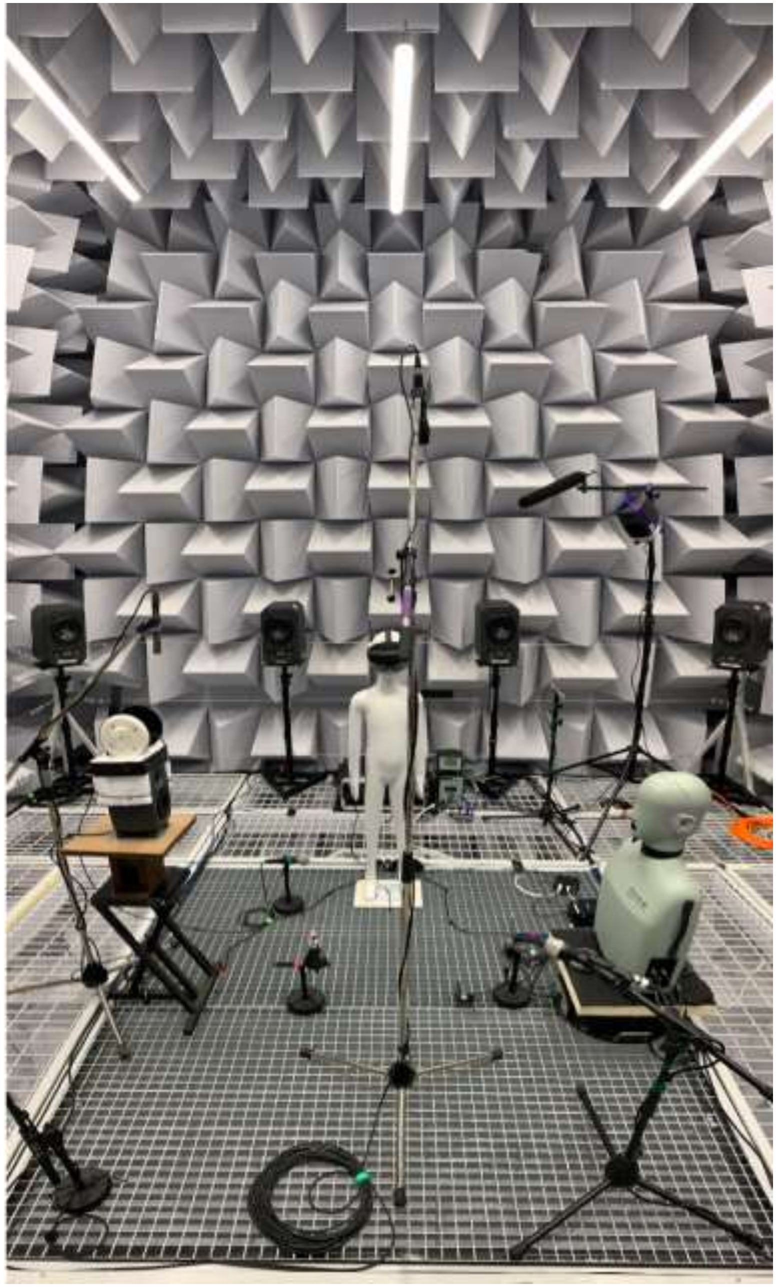
ADASP research group

The group

LTCI lab / IDS department/ S2A team/ ADASP

- 5 Faculty members + 1 Engineer

- 21 PhDs/ 3 Post-Doc/ 2 Research Engineers

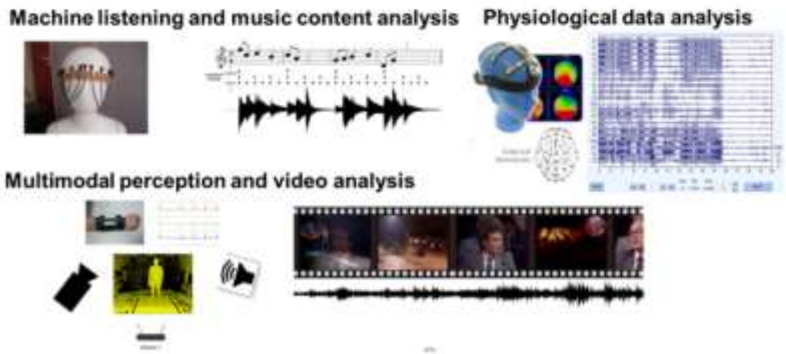


G. Richard

Model-based audio deep learning

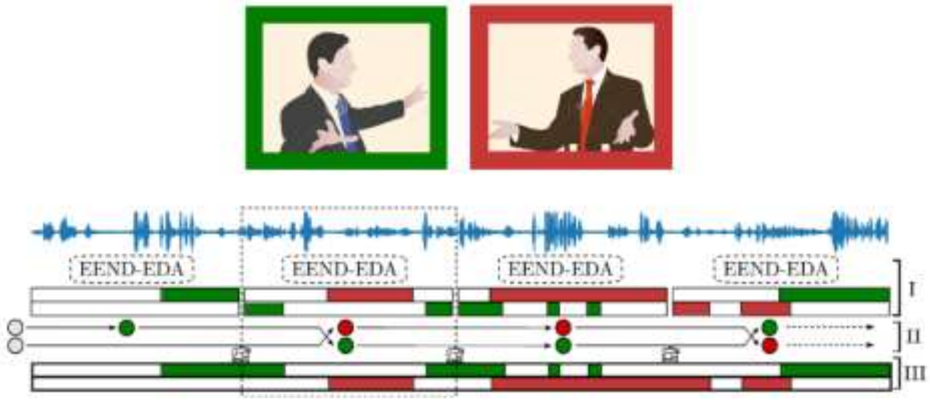
ADASP research group

Research topics

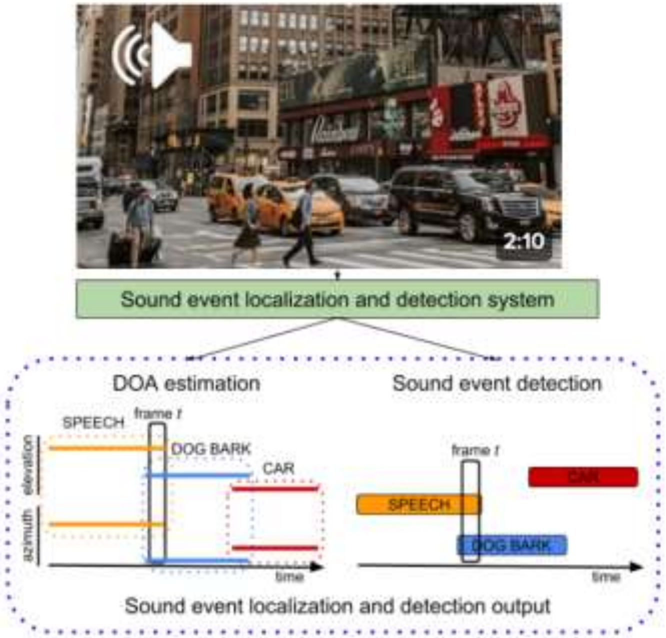


Signal processing, machine learning and **AI** for **analysis** (audio, physiological, multimodal)

Speaker diarization



Sound scene analysis

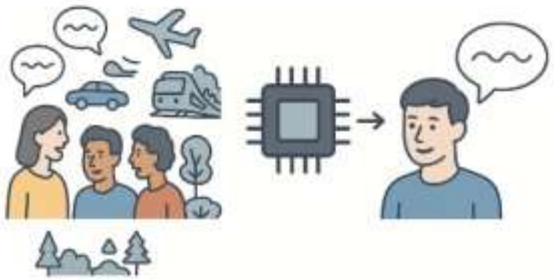


ADASP research group

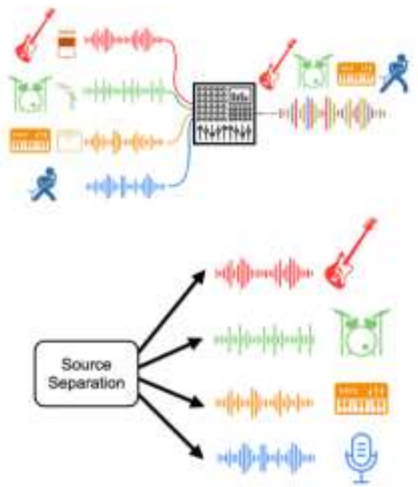
Research topics

Signal processing, machine learning and AI for audio processing

Speech enhancement



Source separation



Neural Audio Coding

G. Richard

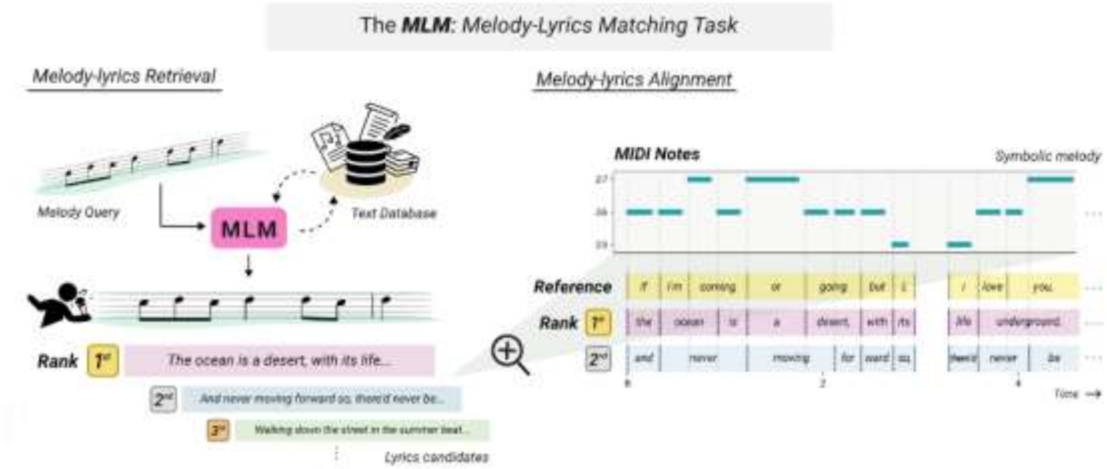
Model-based audio deep learning

ADASP research group

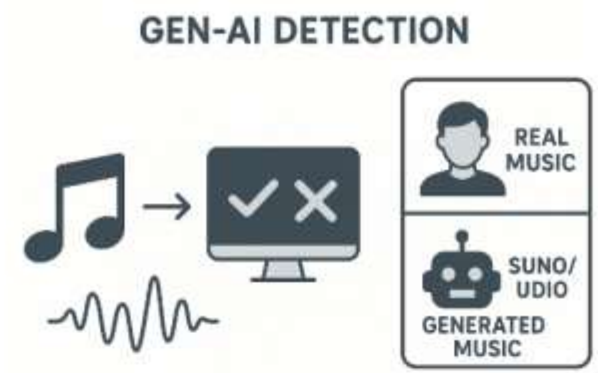
Research topics

Signal processing, machine learning and AI for audio generation

Music generation



Deep-Fake/ Music-AI detection



Hybrid (or Model-based) deep learning

G. Richard

Model-based audio deep learning

Hi-AUDIO: Hybrid and Interpretable Deep Audio machines

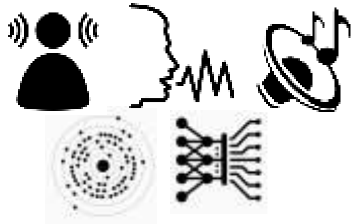
HI-AUDIO is a European Research Council "Advanced Grant" (AdG) project supported by the European Union's Horizon 2020 research and innovation program under Grant Agreement-101052978.

HI-AUDIO project: Context and motivation

- Machine learning: a growing trend towards pure “Data-driven” deep learning approaches
- High performances but some main limitations:
  - *“Knowledge” is learned (only) from data*
  - *Complexity: overparametrized models*
  - Overconsumption regime
  - Non-interpretable/non-controllable
- **The main goal of Hi-Audio :** https://hi-audio.imt.fr/

Audio scene analysis, source separation

Audio representation learning



Sound transformation (style transfer, dereverberation,…)

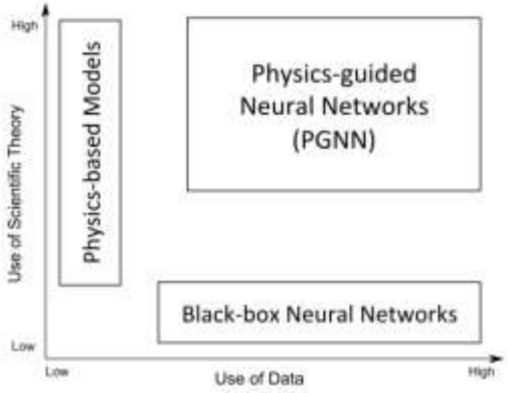
Main goal : To build controllable and frugal machine listening models based on expressive generative modelling

The approach: to build *Hybrid deep learning models*, by **integrating our prior knowledge** about the nature of the processed data.

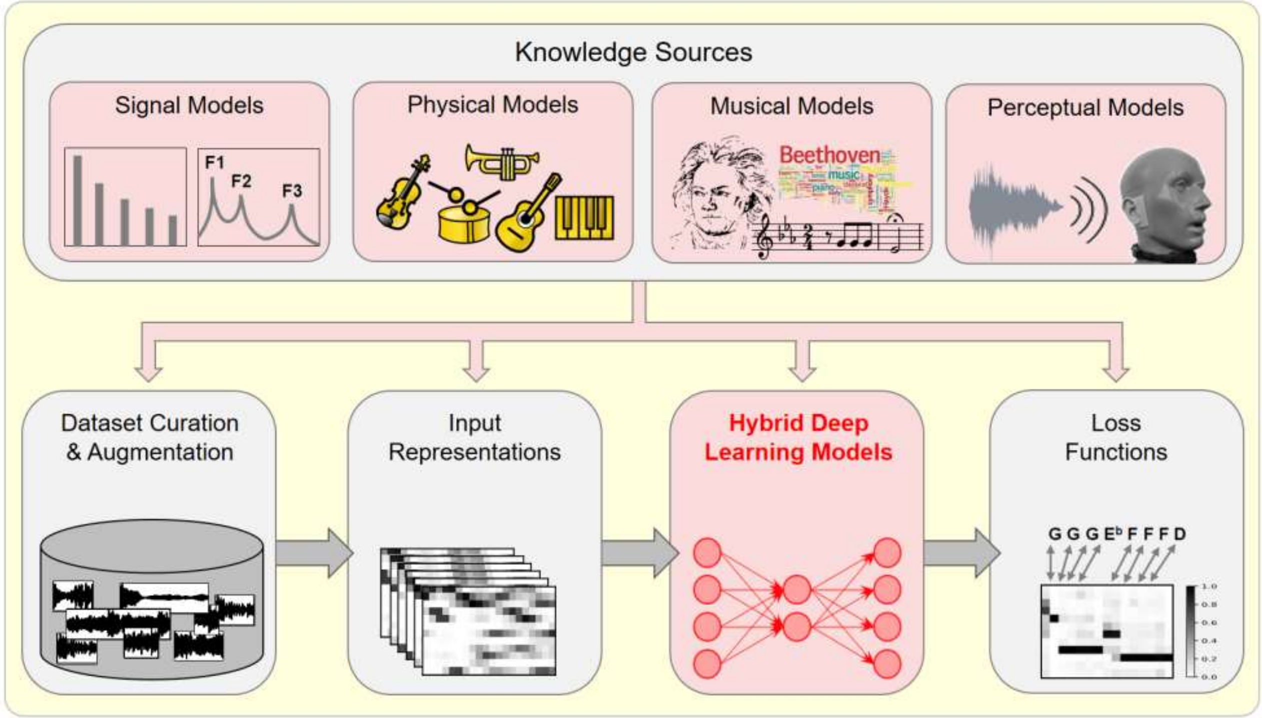
Towards model-based deep learning approaches

G. Richard

- Coupling model-based and deep learning:



Example with Hybrid deep model for Music signals



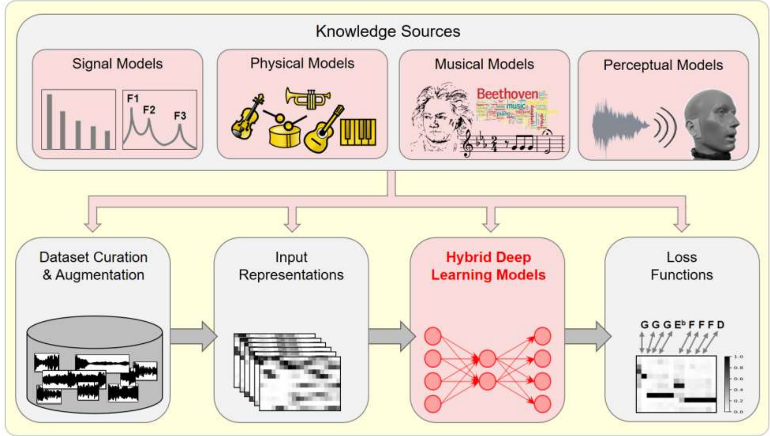
G. Richard, V. Lostanlen, Y.-H. Yang, M. Müller, "Hybrid Deep Learning for Music Information Research", IEEE Signal Processing Magazine - Special Issue on Model-based and Data-Driven Audio Signal Processing, 2025

Hi-Audio, Hybrid and Interpretable Deep neural audio machines, European Research Council "Advanced Grant" (AdG) project - https://hi-audio.imt.fr/

G. Richard

Model-based audio deep learning

Some results



- **Model-based deep learning for audio signals [1]**
- **Music generation, Style transfer, sound transformation:**
  - Novel Structure-informed Positional Encoding (PE) methods for using transformers of linear complexity [2,3]
  - Interpretable music synthesis and sound transformation algorithms exploiting diffusion models [4]
  - Unsupervised model-based deep learning for musical source separation (singing voice, drums) [5,6]
  - New disentangled discrete representations for sound transformation or joint audio coding and source separation [7,8]
- **Deep Hybrid dereverberation** : combining differentiable physical model of reverberation with deep learning for speech dereverberation [9]
- Development and launch of the **HI-AUDIO platform** for distributed music recordings (to gather a large, varied, multi-genre, multi-track, multi-instruments annotated music database) : https://hiaudio.fr/ [10]

[1] G..Richard, V. Lostanlen, Y.-H. Yang, M. Müller, "Model-based Deep Learning for Music Information Research", IEEE Signal Processing Magazine, 2024

[2] M. Agarwal C. Wang, G. Richard. F-StrIPE: Fast Structure-Informed Positional Encoding for Symbolic Music Generation, ICASSP 2025.

[3] M. Agarwal C. Wang, G. Richard. Of All StrIPEs: Investigating Structure-informed Positional Encoding for Efficient Music Generation, https://arxiv.org/pdf/2504.05364

[4] T. Baoueb, X. Bie, H. Janati, G. Richard. WaveTransfer: A Flexible End-to-end Multi-instrument Timbre Transfer with Diffusion. MLSP 2024.

[5] K Schulze-Forster, G. Richard, L. Kelley, C. Doire, R Badeau Unsupervised Music Source Separation Using Differentiable Parametric Source Models, IEEE Trans. On AASP, 2023

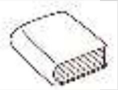
[6] B. Torres, G. Peeters, G. Richard, "The Inverse Drum Machine: Source Separation Through Joint Transcription and Analysis-by-Synthesis", https://arxiv.org/abs/2505.03337

[7] X. Bie, X. Liu, G. Richard. Learning Source Disentanglement in Neural Audio Codec. ICASSP 2025

[8] B. Ginies, X. Bie, O. Fercoq, G. Richard, Soft Disentanglement in Frequency Bands for \\ Neural Audio Codecs, Eusipco 2025

[9] Louis Bahrman, Mathieu Fontaine, Gaël Richard, U-DREAM: Unsupervised Dereverberation guided by a Reverberation Model, 2025, preprint https://hal.science/hal-05158698v1

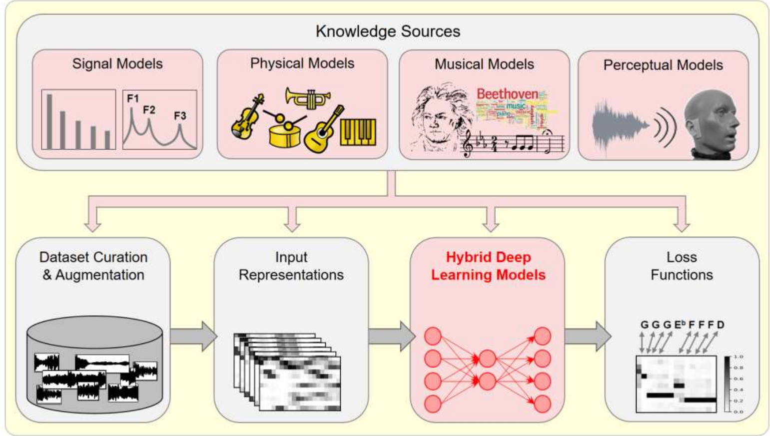
[10] J. Gil Panal, A. David, G. Richard, "The Hi-Audio online platform for distributed music recordings", Submitted to the Eurasip Journal on Audio, Speech and Music Processing, 2025



G. Richard

Model-based audio deep learning

Some results



- **Model-based deep learning for audio signals [1]**
- **Music generation, Style transfer, sound transformation:**
  - Novel Structure-informed Positional Encoding (PE) methods for using transformers of linear complexity [2,3]
  - Interpretable music synthesis and sound transformation algorithms exploiting diffusion models [4]
  - **Unsupervised model-based deep learning for musical source separation (singing voice, drums) [5,6]**
  - New disentangled discrete representations for sound transformation or joint audio coding and source separation [7,8]
- **Deep Hybrid dereverberation** : combining differentiable physical model of reverberation with deep learning for speech dereverberation [9]
- Development and launch of the **HI-AUDIO platform** for distributed music recordings (to gather a large, varied, multi-genre, multi-track, multi-instruments annotated music database) : https://hiaudio.fr/ [10]

[1] G..Richard, V. Lostanlen, Y.-H. Yang, M. Müller, "Model-based Deep Learning for Music Information Research", IEEE Signal Processing Magazine, 2024
[2] M. Agarwal C. Wang, G. Richard. F-StrIPE: Fast Structure-Informed Positional Encoding for Symbolic Music Generation, ICASSP 2025.
[3] M. Agarwal C. Wang, G. Richard. Of All StrIPEs: Investigating Structure-informed Positional Encoding for Efficient Music Generation, https://arxiv.org/pdf/2504.05364
[4] T. Baoueb, X. Bie, H. Janati, G. Richard. WaveTransfer: A Flexible End-to-end Multi-instrument Timbre Transfer with Diffusion. MLSP 2024.
[5] K Schulze-Forster, G. Richard, L. Kelley, C. Doire, R Badeau Unsupervised Music Source Separation Using Differentiable Parametric Source Models, IEEE Trans. On AASP, 2023
[6] B. Torres, G. Peeters, G. Richard, "The Inverse Drum Machine: Source Separation Through Joint Transcription and Analysis-by-Synthesis", https://arxiv.org/abs/2505.03337
[7] X. Bie, X. Liu, G. Richard. Learning Source Disentanglement in Neural Audio Codec. ICASSP 2025
[8] B. Ginies, X. Bie, O. Fercoq, G. Richard, Soft Disentanglement in Frequency Bands for \\ Neural Audio Codecs, Eusipco 2025
[9] Louis Bahrman, Mathieu Fontaine, Gaël Richard, U-DREAM: Unsupervised Dereverberation guided by a Reverberation Model, 2025, preprint https://hal.science/hal-05158698v1
[10] J. Gil Panal, A. David, G. Richard, "The Hi-Audio online platform for distributed music recordings", Submitted to the Eurasip Journal on Audio, Speech and Music Processing, 2025

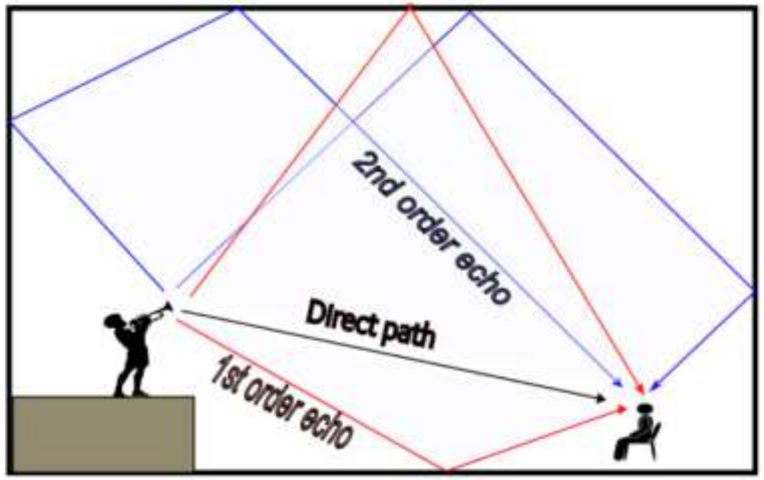
G. Richard

Model-based audio deep learning

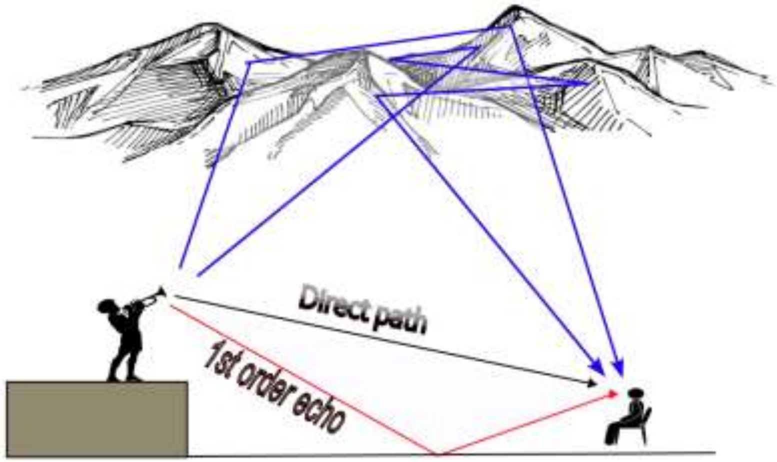
Deep hybrid De-reverberation

Reverberation : definition

- "In acoustics, **reverberation** is a persistence of sound after it is produced" [1]
- It is often created when a sound is reflected on surfaces, causing multiple reflections that build up and then decay as the sound is absorbed by the surfaces of objects in the space [2]



Reverberation in a room

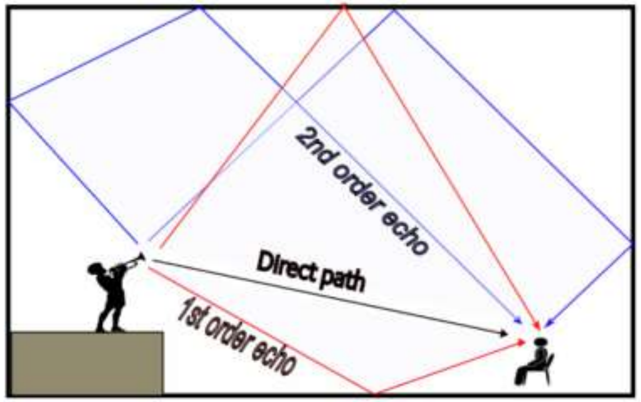


Reverberation in an open space

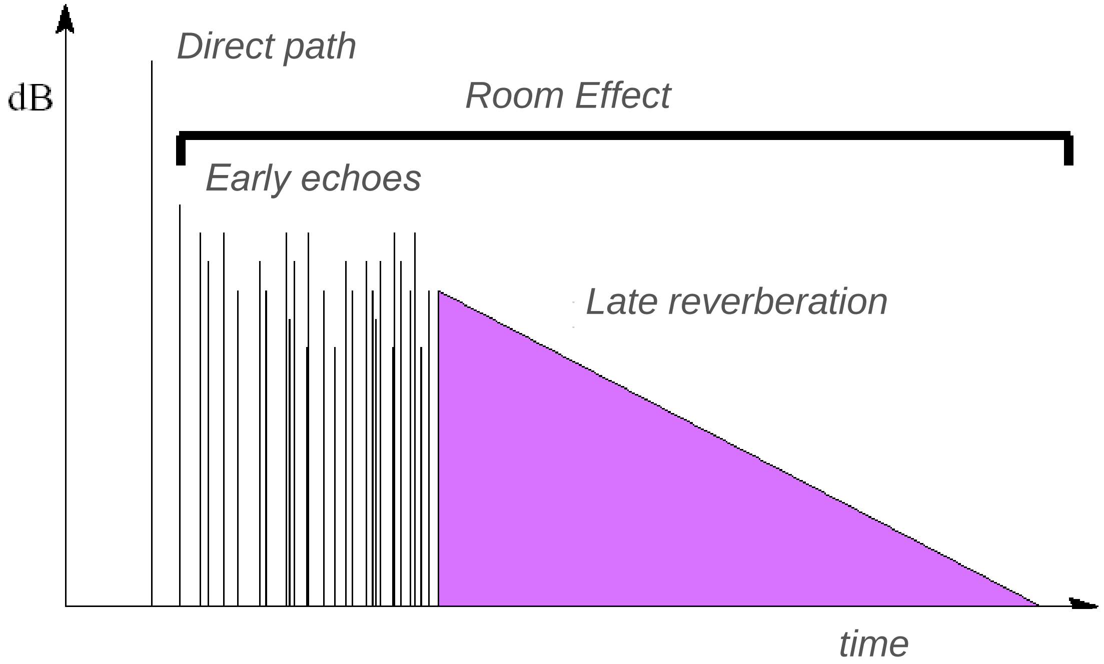
[1] Wikipedia, from Valente, Michael; Holly Hosford-Dunn; Ross J. Roeser (2008). Audiology. Thieme. pp. 425–426. ISBN 978-1-58890-520-8.
[2] Wikipedia, from Lloyd, Llewelyn Southworth (1970). Music and Sound. Ayer Publishing. pp. 169. ISBN 978-0-8369-5188-2.

Reverberation: Room effect

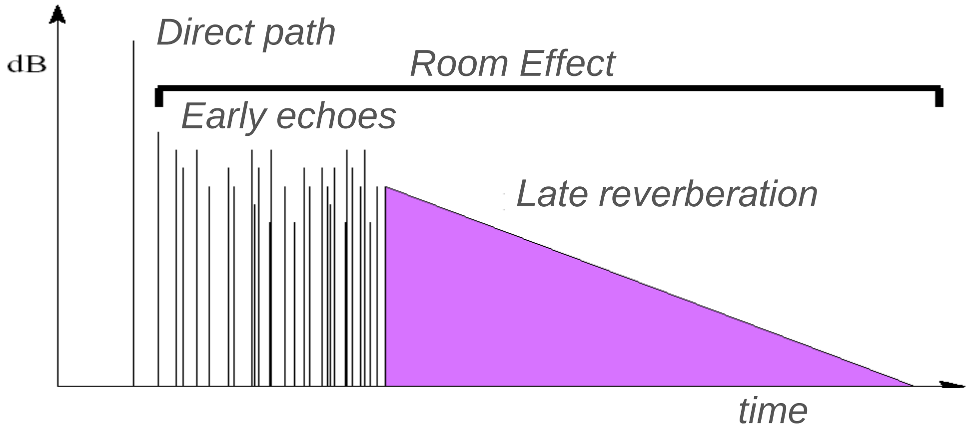
- Room effect can be decomposed in:
  - A contribution due to **early echoes** or early reflexions (which depends on the room geometry and on the positions of the source and microphone)
  - A contribution due to **late reverberation** (which mainly depends on the volume and global absorption of the room)



The Room Impulse Response (RIR)



Reverberation: Room effect



- Room effect = filtering effect

$$y(t) = \int_0^\infty x(t-u)h(u)du$$

- or

The Room Impulse Response (RIR)
(or acoustic channel)

$$y(n) = \sum_{i=0}^\infty x(n-i)h(i)$$

Applications: Reverberation and Dereverberation

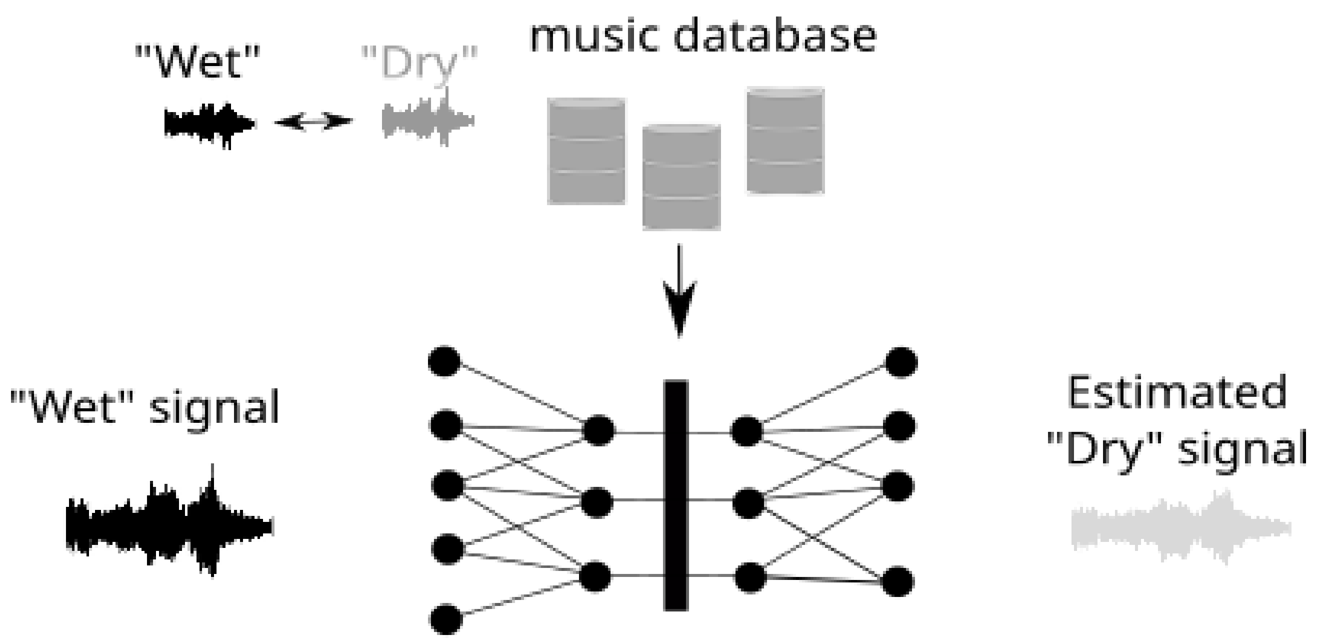
- **Dereverberation:** removing the reverberation effect to retrieve the original source (or « dry » signal)

$$\text{"Recovering } \hat{x}(n) \text{ from the reverberated signal } y(n)\text{"}$$

  - Applications:
    - Speech enhancement (especially late reverberation removal to increase intelligibility)
    - Robust speech recognition
    - Acoustic transfer

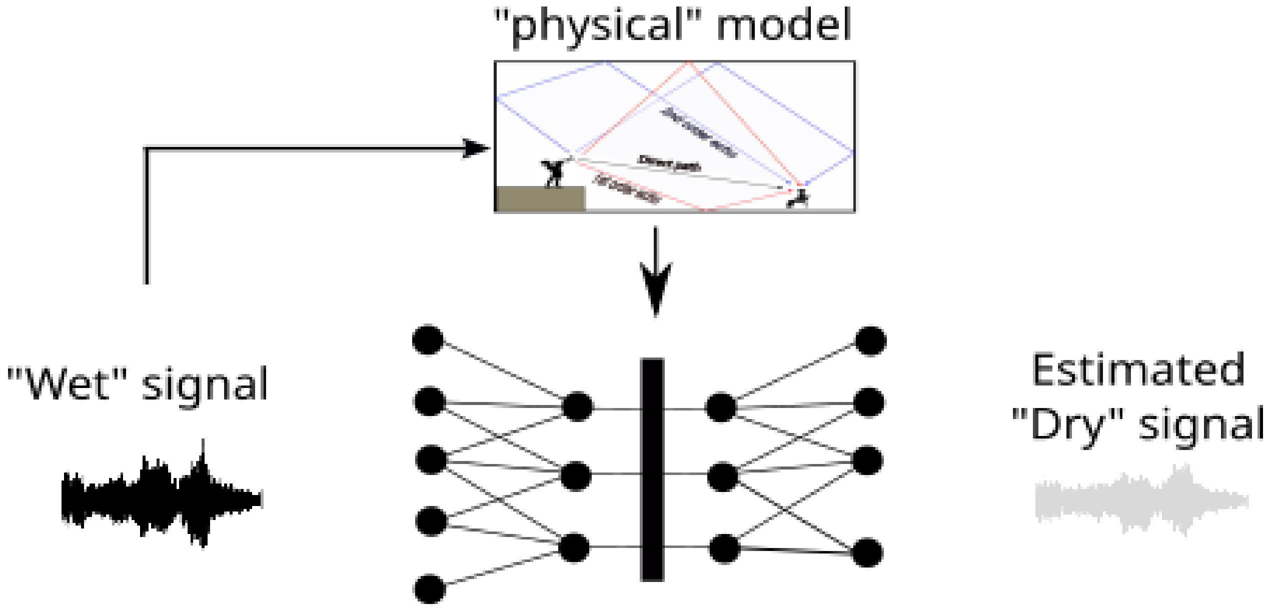
Towards model-based deep learning approaches

- Machine learning: a growing trend towards pure “Data-driven” deep learning approaches



Towards model-based deep dereverberation

- Exploiting a physical model of reverberation



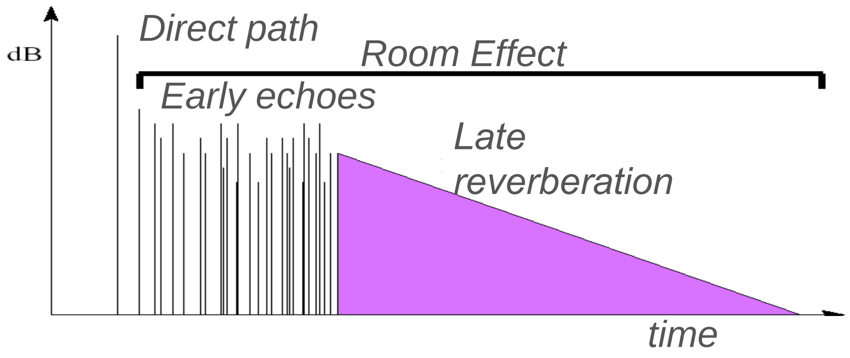
Louis Bahrman, Mathieu Fontaine, Gaël Richard, U-DREAM: Unsupervised Dereverberation guided by a Reverberation Model, 2025, preprint https://hal.science/hal-05158698v1

G. Richard

Model-based audio deep learning

Towards model-based deep dereverberation

Exploiting a room impulse response model



- The RIR model: important parameters:
  - **Direct-to-Reverberant ratio (DRR):** quantifies the energy balance between the direct path and the reverberant tail

$$\mathrm{DRR}_{dB} = 10\log_{10}\left(\frac{\sum_{n=0}^{n_d} h^2(n)}{\sum_{n=n_d+1}^{\infty} h^2(n)}\right)$$

  - **Reverberation time RT_{60} :** can be estimated (Under idealized conditions) from the slope of the energy decay curve (EDC)

$$\mathrm{EDC}_h(t) = \int_t^{+\infty} h(u)du,$$

P. A. Naylor and N. D. Gaubitch, Eds., Speech Dereverberation, ser. Signals and Communication Technology. London: Springer London, 2010.

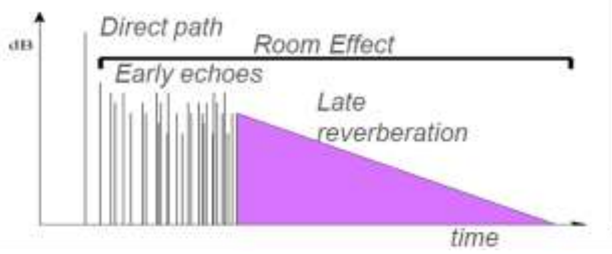
The statistical Polack model

G. Richard

Model-based audio deep learning

- DRR and RT60 are sufficient to characterize the Polack (late) reverberation model [1]

$$h_r(n) = b(n)e^{-n/\tau},$$



  - With $b(n) \sim \mathcal{N}(0, \sigma^2)$ and $\tau = \frac{\mathrm{RT}_{60} f_s}{3\ln(10)}$.

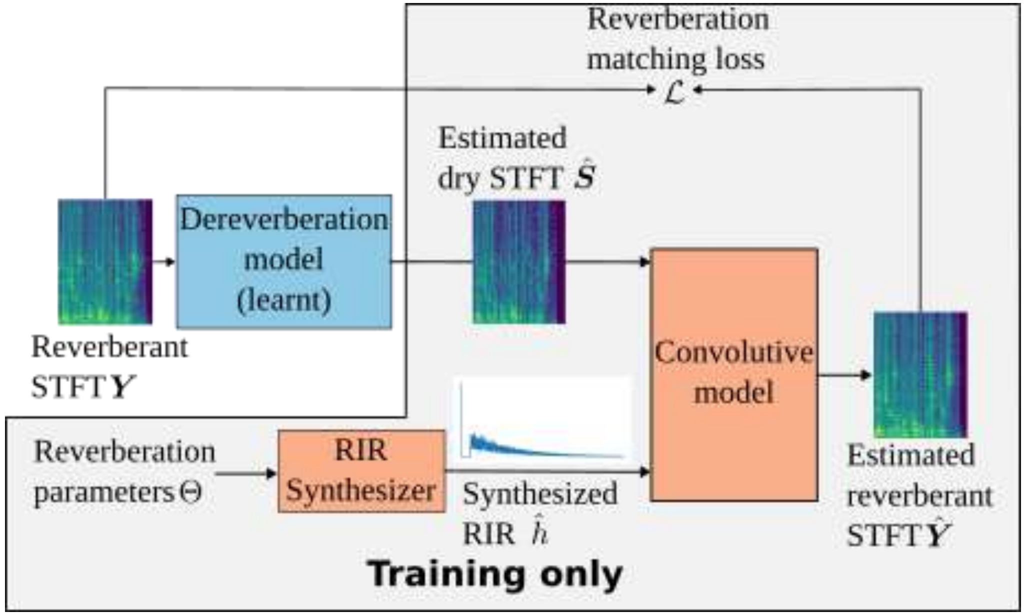
- For reverberation, the polack model is valid after the « mixing time » $n_m = (4V f_s)/(cA)$, where V , f_s, c, A are respectively the room volume, the sampling frequency, the speed of sound and the area of the walls.

[1] J.-D. Polack, “La transmission de l’energie sonore dans les salles,” Ph.D. dissertation (in French), Université du Maine, 1988

G. Richard

Model-based audio deep learning

Towards model-based deep dereverberation
Exploiting a room impulse response model



Regularization term
(for low amplitudes)

- Reverberation Loss used: $\mathcal{L} = \sum_{f,t} \left[|\hat{Y}_{f,t} - Y_{f,t}|^2 + \lambda \left| \log \left(\frac{1+\gamma|\hat{Y}_{f,t}|}{1+\gamma|Y_{f,t}|} \right) \right|^2 \right]$

L. Bahrman, M. Fontaine, and G. Richard, "A Hybrid Model for Weakly Supervised Speech Dereverberation," in ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing 873 (ICASSP), Apr. 2025,
S. Schwär and M. Müller, "Multi-Scale Spectral Loss Revisited," IEEE Signal Process. Lett., vol. 30, pp. 1712–1716, 2023.

Towards model-based deep dereverberation

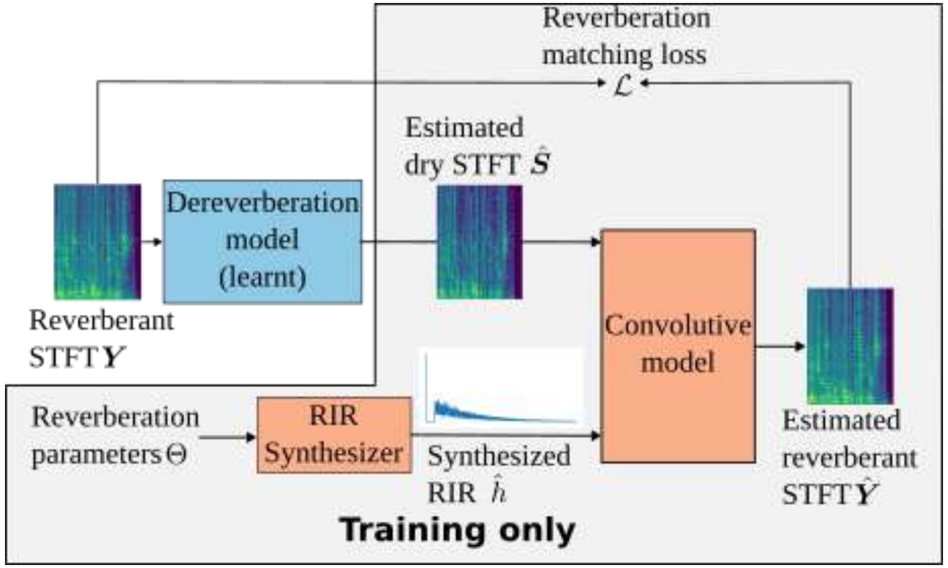
Exploiting a room impulse response model

- Main advantages of the model

  - Can be trained in an unsupervised way
(no needs of pairs Wet- dry of signals)

  - The dereverberation model is more interpretable and controllable (e.g. use « physical » constraints)

  - Smaller network may be sufficient to obtain similar performances than bigger networks trained in a supervised way



L. Bahrman, M. Fontaine, and G. Richard, "A Hybrid Model for Weakly Supervised Speech Dereverberation," in ICASSP 2025 - 2025 IEEE 872 International Conference on Acoustics, Speech and Signal Processing 873 (ICASSP), Apr. 2025,

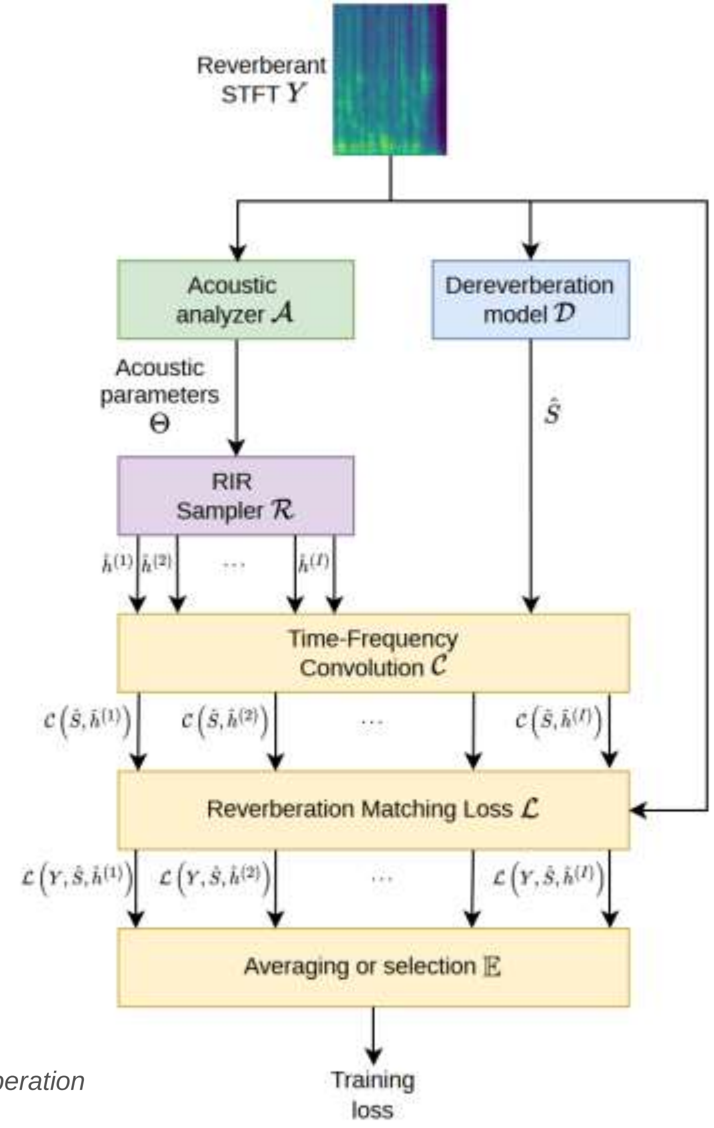
U-DREAM: the extension to “Unsupervised Dereverberation” guided by a Reverberation Model

- The optimization problem

$$\hat{\boldsymbol{S}}, \hat{\Theta} = \underset{\boldsymbol{S},\Theta}{\operatorname{argmin}} \, \mathbb{E}_{p(h|\Theta)} \left[\|\boldsymbol{Y} - \mathcal{C}(\boldsymbol{S}, h)\|_F^2 \right]$$

- An **Acoustic Analyzer** to estimate acoustic parameters for sampling candidate Room Impulse Responses

- **RIR sampler**, using Polack’s model as previously, but several draws possible



Louis Bahrman, Mathieu Fontaine, Gaël Richard, U-DREAM: Unsupervised Dereverberation guided by a Reverberation Model, 2025, preprint https://hal.science/hal-05158698v1

G. Richard

Model-based audio deep learning

Towards model-based deep dereverberation

Exploiting a room impulse response model

Some results

- **Dataset used:** EARS-ISM (synthetic RIR) - EARS-Reverb (Real RIRs)
- **Dereverberation model used**: BiLSTM *(2-layer 599 bidirectional LSTM model followed by a linear layer, performing subband processing of the STFT magnitudes).*
- **Pre-trained Acoustic Analyzer:** Parameter MSE loss, trained with 100 samples of couple (y, $\Theta = \{\mathrm{DRR}, \mathrm{RT}_{60}\}$)
- **Evaluation (objective) metrics**
  - SI-SDR (« signal distorsion »),
  - PESQ (« perceptual quality »
  - STOI (« intelligibility »),
  - SRMR (« reverberation »)

L. Bahrman, M. Fontaine, and G. Richard, "A Hybrid Model for Weakly Supervised Speech Dereverberation," in ICASSP 2025,Apr. 2025,

L. Bahrman, M. Fontaine, G. Richard, U-DREAM: Unsupervised Dereverberation guided by a Reverberation Model, 2025, preprint https://hal.science/hal-05158698v1

(EARS): J. Richter, Y.-C. Wu, S. Krenn, S. Welker, B. Lay, S. Watanabe, A. Richard, and T. Gerkmann, "EARS: An Anechoic Fullband Speech 1001 Dataset Benchmarked for Speech Enhancement and Dereverberation," 1002 in *Interspeech 2024*.

(BiLSTM): F. Weninger & al. "Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR," in Latent Variable Analysis and Signal Separation, E. Vincent, A. Yeredor, Z. Koldovsk ́and P. Tichavsk ́y, Eds. Cham:Springer International Publishing, 2015, pp. 91–99.

(WPE) T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech Dereverberation Based on Variance-Normalized Delayed Linear Prediction," IEEE Trans. ASLP, vol. 18, no. 7, Sep. 2010.

Towards model-based deep dereverberation

Exploiting a room impulse response model

Some results

		Synthetic RIRs				Real RIRs			
Supervision type	Supervision	↑ SISDR	ESTOI	WB-PESQ	SRMR	↑ SISDR	ESTOI	WB-PESQ	SRMR
strong	Dry speech	-2.0 ± 6.1	0.75 ± 0.12	2.15 ± 0.64	7.7 ± 3.6	-14.5 ± 9.2	0.61 ± 0.13	1.73 ± 0.41	6.5 ± 2.9
	Exact RIR	-2.3 ± 5.8	0.72 ± 0.13	1.99 ± 0.66	8.5 ± 3.6	-15.6 ± 10.6	0.61 ± 0.14	1.75 ± 0.46	6.5 ± 2.8
weak	Oracle parameters	-1.7 ± 5.4	0.67 ± 0.15	1.74 ± 0.62	6.4 ± 3.0	-14.5 ± 8.1	0.58 ± 0.13	1.64 ± 0.39	5.4 ± 2.6
unsupervised	Pretrained Acoustic Analyzer	-3.6 ± 5.1	0.64 ± 0.12	1.62 ± 0.43	8.0 ± 3.4	-14.5 ± 8.7	0.57 ± 0.12	1.58 ± 0.31	6.2 ± 2.9
	WPE	-2.1 ± 5.0	0.72 ± 0.14	1.94 ± 0.76	6.9 ± 3.4	-15.8 ± 9.1	0.54 ± 0.17	1.54 ± 0.43	5.2 ± 3.2
	Reverberant	-6.7 ± 6.4	0.67 ± 0.15	1.79 ± 0.64	8.2 ± 5.9	-16.1 ± 9.3	0.52 ± 0.17	1.48 ± 0.36	4.8 ± 2.9

- **All methods perform some level of dereverberation**

L. Bahrman, M. Fontaine, and G. Richard, “A Hybrid Model for Weakly Supervised Speech Dereverberation,” in ICASSP 2025,Apr. 2025,
L. Bahrman, M. Fontaine, G. Richard, U-DREAM: Unsupervised Dereverberation guided by a Reverberation Model, 2025, preprint https://hal.science/hal-05158698v1
(EARS): J. Richter, Y.-C. Wu, S. Krenn, S. Welker, B. Lay, S. Watanabe, A. Richard, and T. Gerkmann, “EARS: An Anechoic Fullband Speech 1001 Dataset Benchmarked for Speech Enhancement and Dereverberation,” 1002 in *Interspeech 2024*.
(BiLSTM): F. Weninger & al. “Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR,” in Latent Variable Analysis and Signal Separation, E. Vincent, A. Yeredor, Z. Koldovsk ́and P. Tichavsk ́y, Eds. Cham:Springer International Publishing, 2015, pp. 91–99.
(WPE) T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, “Speech Dereverberation Based on Variance-Normalized Delayed Linear Prediction,” IEEE Trans. ASLP, vol. 18, no. 7, Sep. 2010.

Towards model-based deep dereverberation

Exploiting a room impulse response model

Some results

Supervision type	Supervision	Synthetic RIRs ↑ SISDR	ESTOI	WB-PESQ	SRMR	Real RIRs ↑ SISDR	ESTOI	WB-PESQ	SRMR
strong	Dry speech	-2.0 ± 6.1	0.75 ± 0.12	2.15 ± 0.64	7.7 ± 3.6	-14.5 ± 9.2	0.61 ± 0.13	1.73 ± 0.41	6.5 ± 2.9
	Exact RIR	-2.3 ± 5.8	0.72 ± 0.13	1.99 ± 0.66	8.5 ± 3.6	-15.6 ± 10.6	0.61 ± 0.14	1.75 ± 0.46	6.5 ± 2.8
weak	Oracle parameters	-1.7 ± 5.4	0.67 ± 0.15	1.74 ± 0.62	6.4 ± 3.0	-14.5 ± 8.1	0.58 ± 0.13	1.64 ± 0.39	5.4 ± 2.6
unsupervised	Pretrained Acoustic Analyzer	-3.6 ± 5.1	0.64 ± 0.12	1.62 ± 0.43	8.0 ± 3.4	-14.5 ± 8.7	0.57 ± 0.12	1.58 ± 0.31	6.2 ± 2.9
	WPE	-2.1 ± 5.0	0.72 ± 0.14	1.94 ± 0.76	6.9 ± 3.4	-15.8 ± 9.1	0.54 ± 0.17	1.54 ± 0.43	5.2 ± 3.2
	Reverberant	-6.7 ± 6.4	0.67 ± 0.15	1.79 ± 0.64	8.2 ± 5.9	-16.1 ± 9.3	0.52 ± 0.17	1.48 ± 0.36	4.8 ± 2.9

- **Weakly-supervised method outperforms the baseline WPE on most metrics (especially on real RIRs)**

L. Bahrman, M. Fontaine, and G. Richard, “A Hybrid Model for Weakly Supervised Speech Dereverberation,” in ICASSP 2025,Apr. 2025,
L. Bahrman, M. Fontaine, G. Richard, U-DREAM: Unsupervised Dereverberation guided by a Reverberation Model, 2025, preprint https://hal.science/hal-05158698v1
(EARS): J. Richter, Y.-C. Wu, S. Krenn, S. Welker, B. Lay, S. Watanabe, A. Richard, and T. Gerkmann, “EARS: An Anechoic Fullband Speech 1001 Dataset Benchmarked for Speech Enhancement and Dereverberation,” 1002 in *Interspeech 2024*.
(BiLSTM): F. Weninger & al. “Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR,” in Latent Variable Analysis and Signal Separation, E. Vincent, A. Yeredor, Z. Koldovsk ́and P. Tichavsk ́y, Eds. Cham:Springer International Publishing, 2015, pp. 91–99.
(WPE) T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, “Speech Dereverberation Based on Variance-Normalized Delayed Linear Prediction,” IEEE Trans. ASLP, vol. 18, no. 7, Sep. 2010.

Towards model-based deep dereverberation

Exploiting a room impulse response model

Some results

		Synthetic RIRs				Real RIRs			
Supervision type	Supervision	↑ SISDR	ESTOI	WB-PESQ	SRMR	↑ SISDR	ESTOI	WB-PESQ	SRMR
strong	Dry speech	-2.0 ± 6.1	0.75 ± 0.12	2.15 ± 0.64	7.7 ± 3.6	-14.5 ± 9.2	0.61 ± 0.13	1.73 ± 0.41	6.5 ± 2.9
	Exact RIR	-2.3 ± 5.8	0.72 ± 0.13	1.99 ± 0.66	8.5 ± 3.6	-15.6 ± 10.6	0.61 ± 0.14	1.75 ± 0.46	6.5 ± 2.8
weak	Oracle parameters	-1.7 ± 5.4	0.67 ± 0.15	1.74 ± 0.62	6.4 ± 3.0	-14.5 ± 8.1	0.58 ± 0.13	1.64 ± 0.39	5.4 ± 2.6
unsupervised	Pretrained Acoustic Analyzer	-3.6 ± 5.1	0.64 ± 0.12	1.62 ± 0.43	8.0 ± 3.4	-14.5 ± 8.7	0.57 ± 0.12	1.58 ± 0.31	6.2 ± 2.9
	WPE	-2.1 ± 5.0	0.72 ± 0.14	1.94 ± 0.76	6.9 ± 3.4	-15.8 ± 9.1	0.54 ± 0.17	1.54 ± 0.43	5.2 ± 3.2
	Reverberant	-6.7 ± 6.4	0.67 ± 0.15	1.79 ± 0.64	8.2 ± 5.9	-16.1 ± 9.3	0.52 ± 0.17	1.48 ± 0.36	4.8 ± 2.9

- **Unsupervised method is efficient, in particular on Real RIRs**

L. Bahrman, M. Fontaine, and G. Richard, “A Hybrid Model for Weakly Supervised Speech Dereverberation,” in ICASSP 2025,Apr. 2025,
L. Bahrman, M. Fontaine, G. Richard, U-DREAM: Unsupervised Dereverberation guided by a Reverberation Model, 2025, preprint https://hal.science/hal-05158698v1
(EARS): J. Richter, Y.-C. Wu, S. Krenn, S. Welker, B. Lay, S. Watanabe, A. Richard, and T. Gerkmann, “EARS: An Anechoic Fullband Speech 1001 Dataset Benchmarked for Speech Enhancement and Dereverberation,” 1002 in *Interspeech 2024*.
(BiLSTM): F. Weninger & al. “Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR,” in Latent Variable Analysis and Signal Separation, E. Vincent, A. Yeredor, Z. Koldovsk ́and P. Tichavsk ́y, Eds. Cham:Springer International Publishing, 2015, pp. 91–99.
(WPE) T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, “Speech Dereverberation Based on Variance-Normalized Delayed Linear Prediction,” IEEE Trans. ASLP, vol. 18, no. 7, Sep. 2010.

Towards model-based deep dereverberation

Exploiting a room impulse response model

- Some sounds (weak-supervision results)

	Wet input	Ground truth	FSN (proposed)	FSN	BiLSTM (proposed)	BiLSTM	Baseline
WS			✓	✗	✓	✗	✓
RT60=0.6							

- More audio demo at https://louis-bahrman.github.io/Hybrid-WSSD/

L. Bahrman, M. Fontaine, and G. Richard, "A Hybrid Model for Weakly Supervised Speech Dereverberation," in ICASSP 2025 - 2025 IEEE 872 International Conference on Acoustics, Speech and Signal Processing 873 (ICASSP), Apr. 2025,

(WPE) T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech Dereverberation Based on Variance-Normalized Delayed Linear Prediction," IEEE Trans. ASLP, vol. 18, no. 7, Sep. 2010.

G. Richard

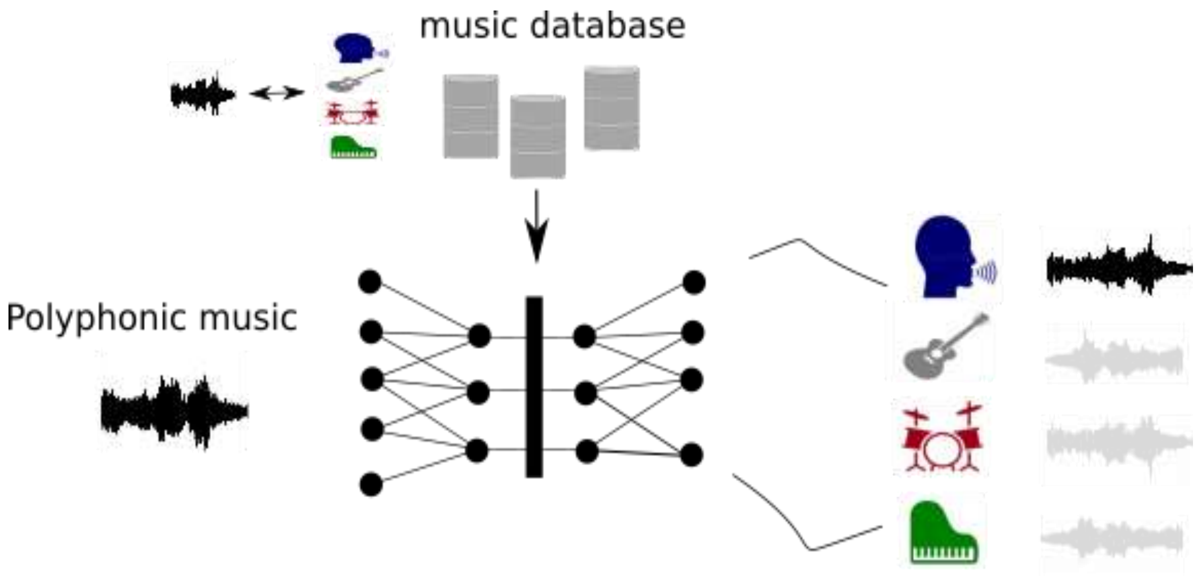
Music Source Separation

G. Richard

Towards Hybrid deep learning

… by **integrating our prior knowledge** about the nature of the processed data.

- For example in music source separation

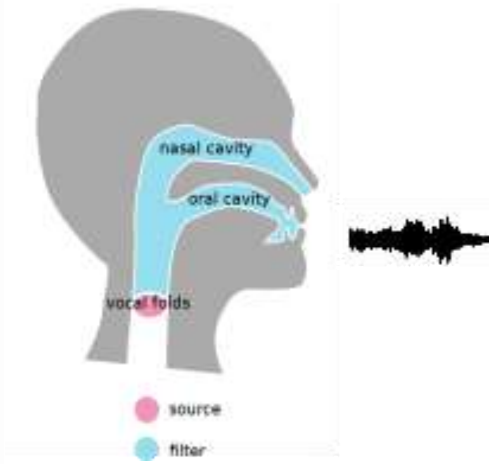


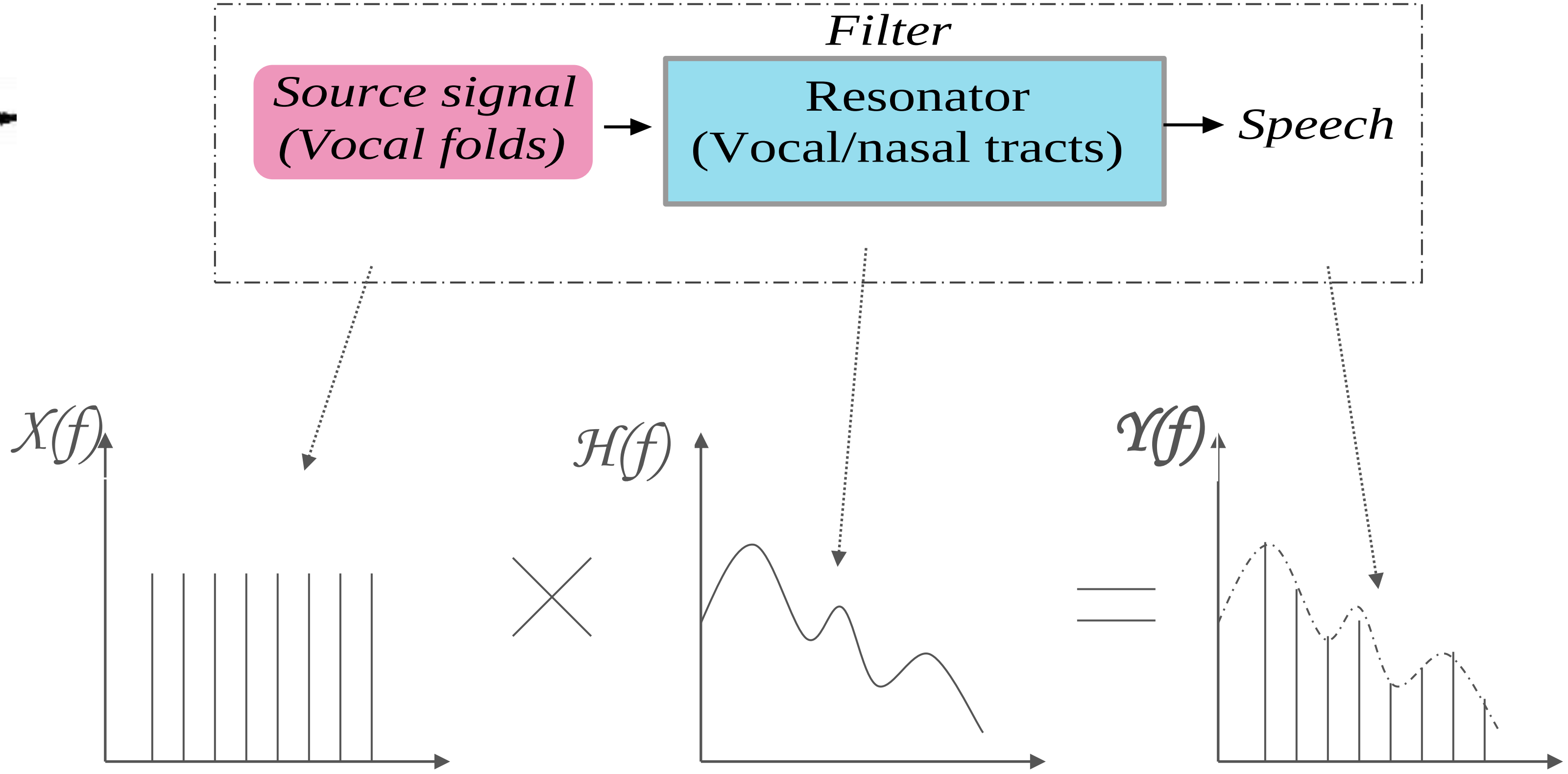
Main limitations:

- *Difficulty to obtain « aligned » data*
- *Knowledge learned (only) from data*
- *Complexity: overparametrized models*
- Overconsumption regime
- **Non-interpretable/non-controllable**

The source filter model

an efficient speech production model



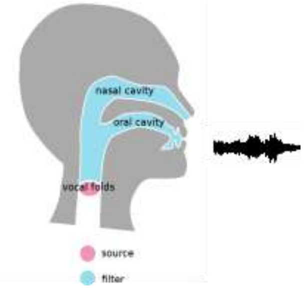


Fant, G. Acoustic theory of speech production, 1960, The Hague, The Netherlands, Mouton.

Towards Hybrid deep learning

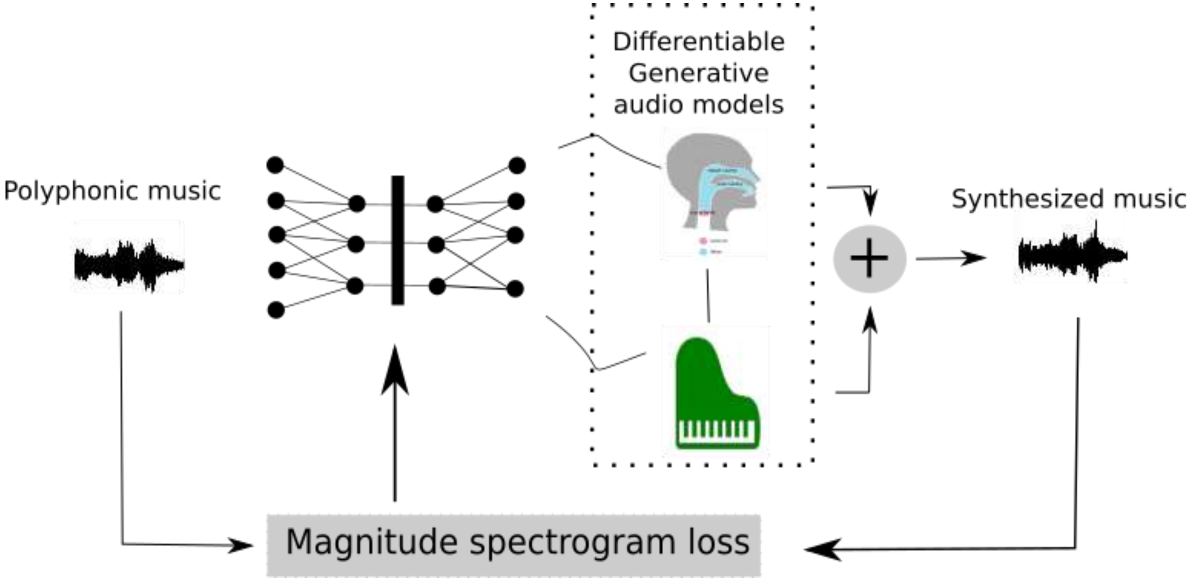
… by **integrating our prior knowledge** about the nature of the processed data.

Knowledge about « how the sound is produced « (e.g. sound production models)



Singing voice as a source / filter model :

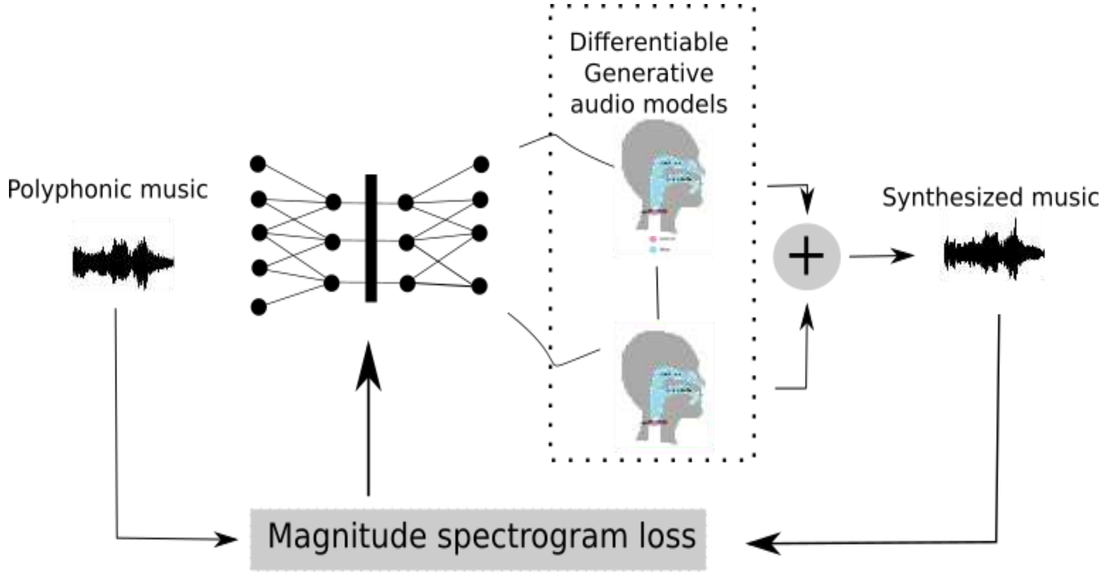
- source = vibration of vocal folds
- Filter = resonances of vocal/nasal cavities



Towards Hybrid deep learning

… by **integrating our prior knowledge** about the nature of the processed data.

- Application for unsupervised audio source separation (choir singing)



Highlights

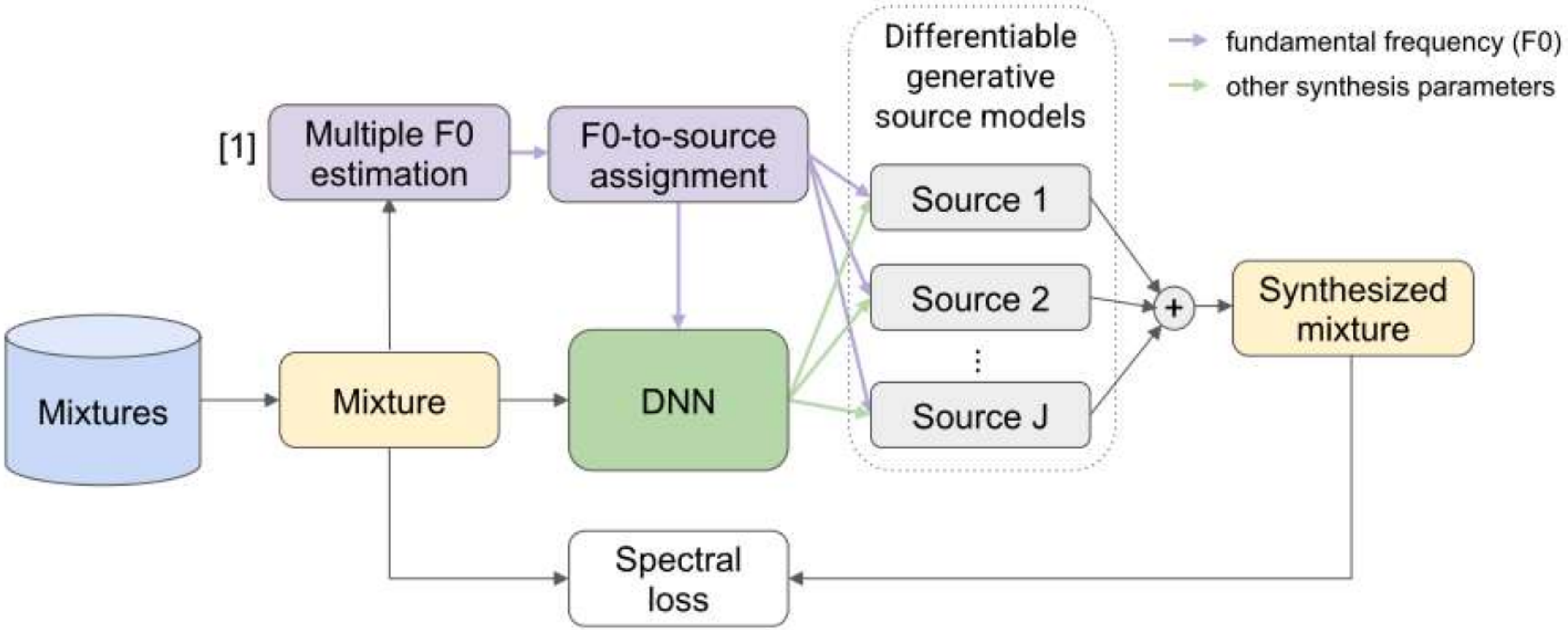
- Unsupervised :
  - Learning only from the polyphonic recording (*no need of the true individual tracks*)
- Homogeneous sources :
  - All sources have similar acoustic properties

K. Schulze-Forster, G. Richard, L. Kelley, C. S. J. Doire and R. Badeau, "Unsupervised Music Source Separation Using Differentiable Parametric Source Models," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1276-1289, 2023, doi: 10.1109/TASLP.2023.3252272. (Open Access)

G. Richard

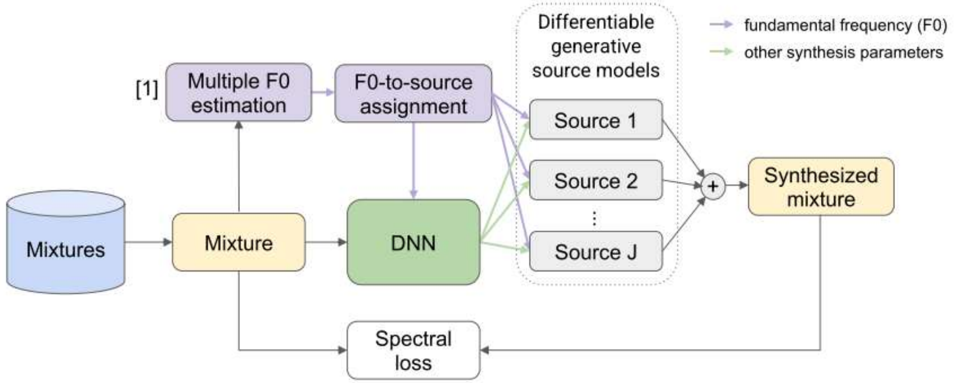
Unsupervised learning strategy

(e.g. no need of the individual source signals)



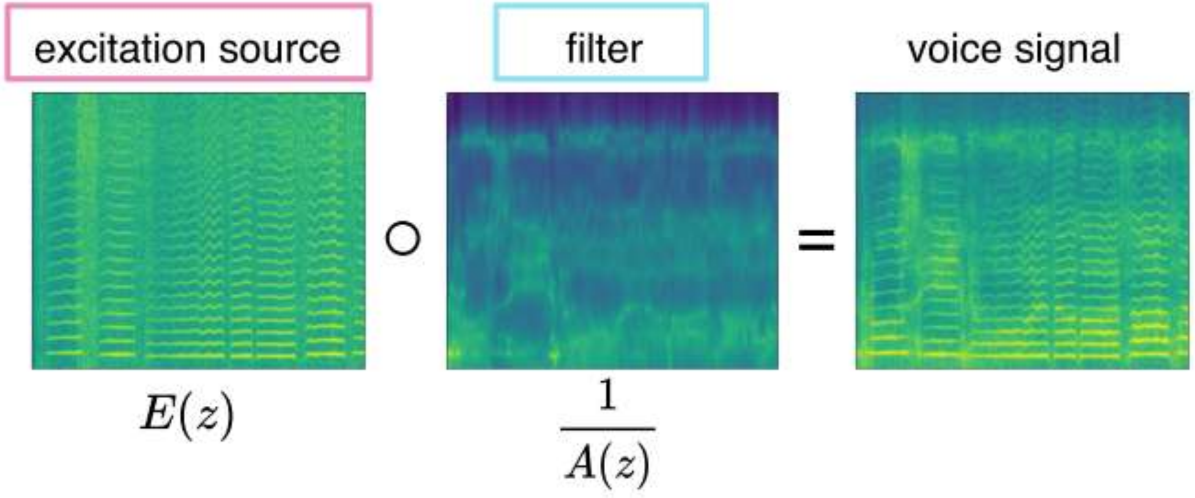
[1] H. Cuesta, B. McFee, and E. Gómez. Multiple f0 estimation in vocal ensembles using convolutional neural networks. ISMIR, 2020.

Parametric source models



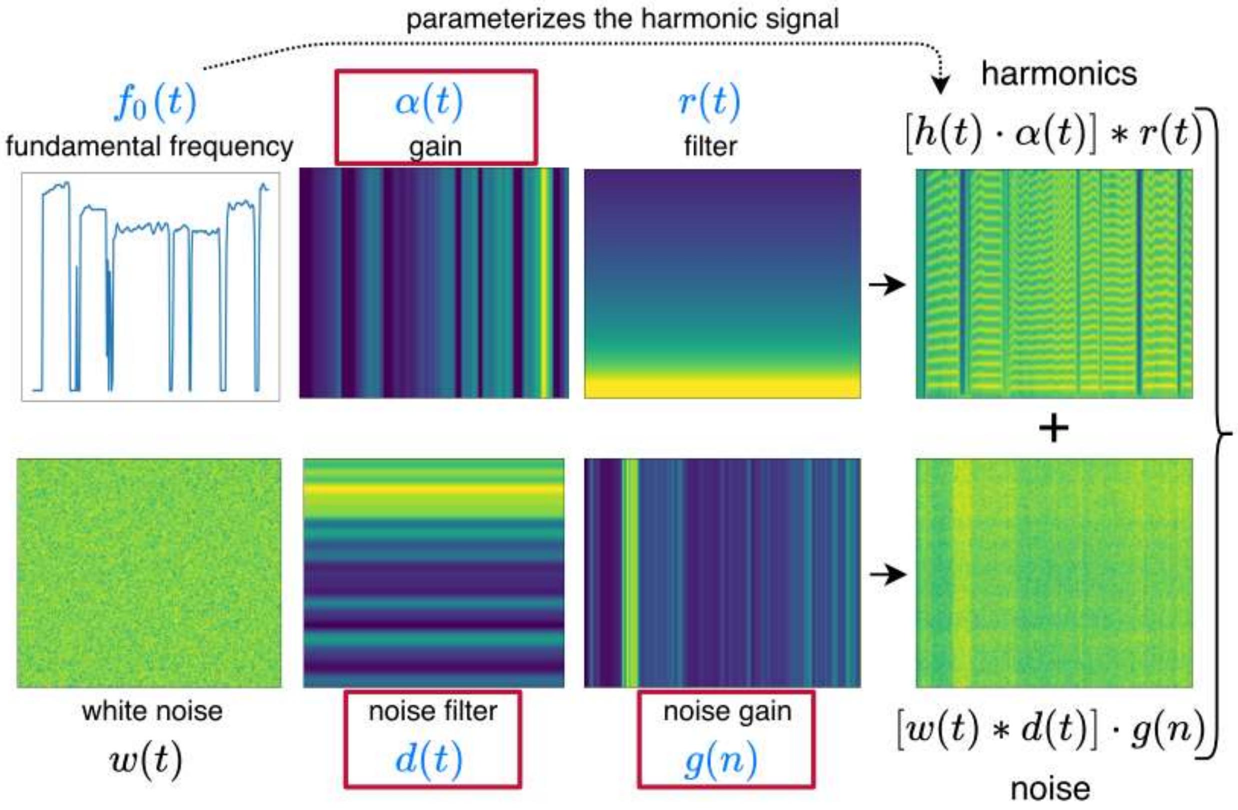
Singing voice as a source / filter model :

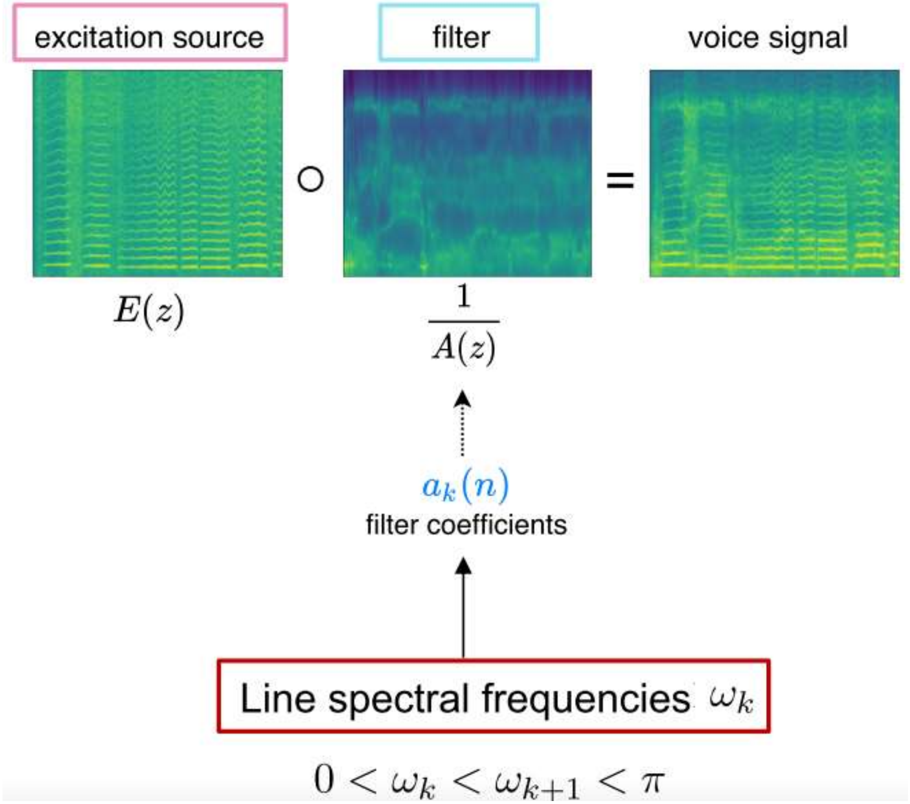
- source = vibration of vocal folds
- Filter = resonances of vocal/nasal cavities



G. Richard

Parametric source models





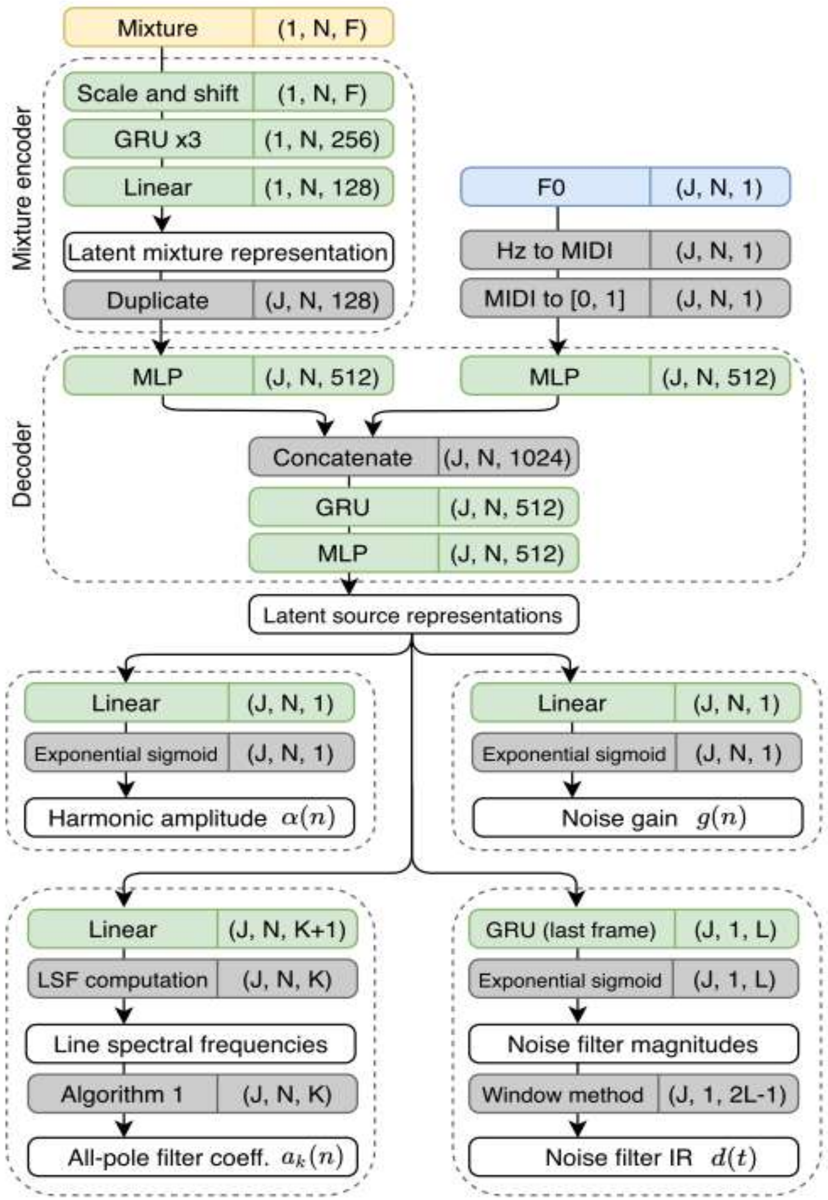
TELECOM Paris

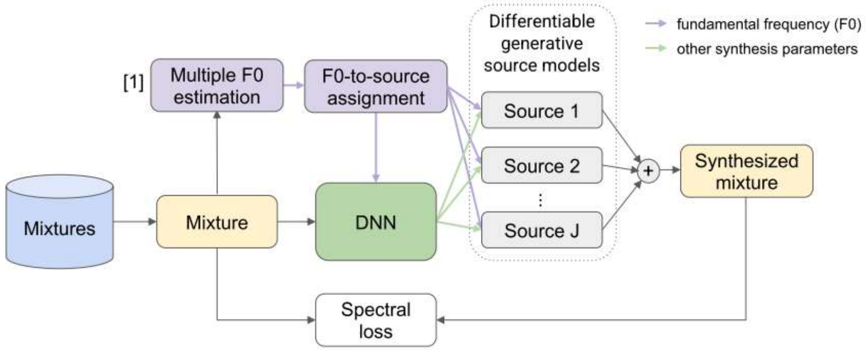
IP PARIS

G. Richard

Model-based audio deep learning

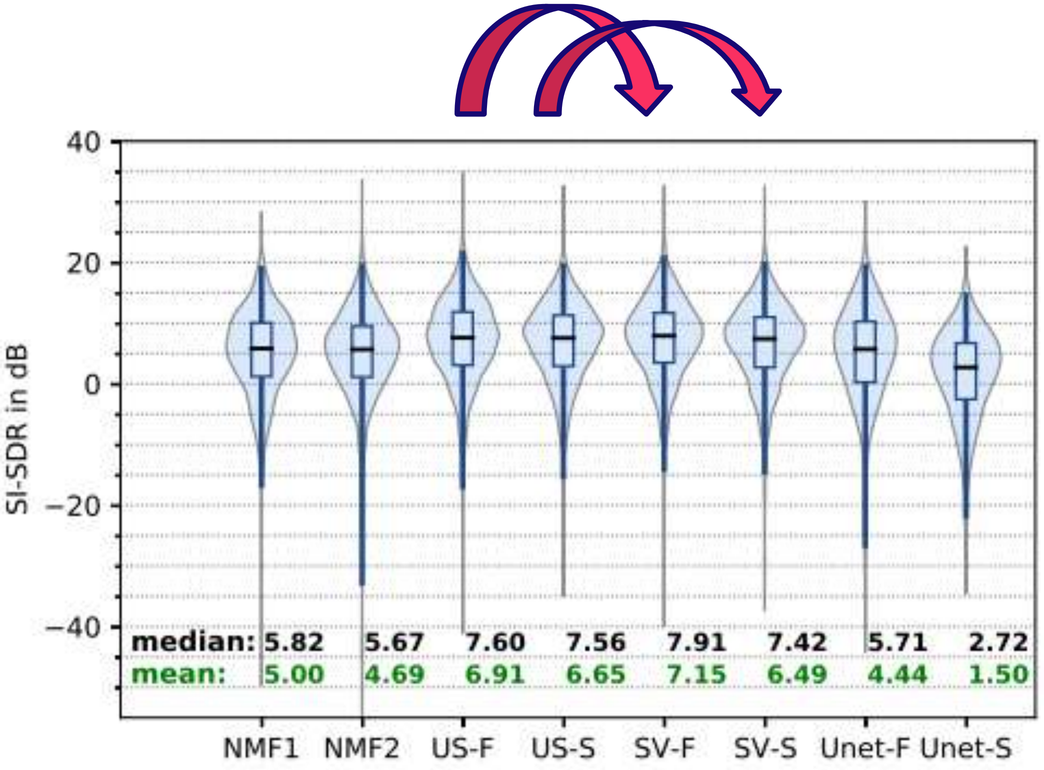
Global architecture overview





Some results

- Unsupervised (US) ≈ supervised (SU)



(b) $J = 4$ sources

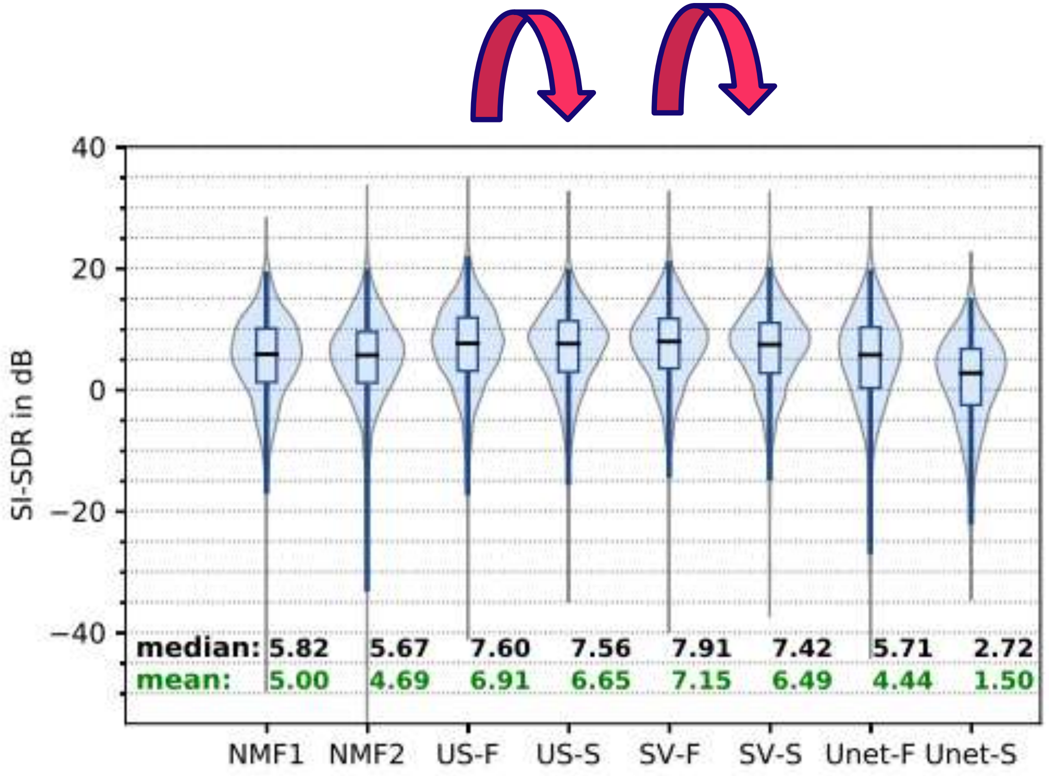
NMF1: S. Ewert and M. M¨uller, "Using score-informed constraints for NMF- based source separation," in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing. IEEE, 2012, pp. 129–132.

NMF2: J.-L. Durrieu, B. David, and G. Richard, "A musically motivated mid- evel representation for pitch estimation and musical audio source separation," IEEE J. Selected Topics in Signal Processing, vol. 5, no. 6, pp. 1180–1191, 2011.

UNET: D. Petermann, P. Chandna, H. Cuesta, J. Bonada, and E. Gomez, "Deep learning based source separation applied to choir ensembles," in Proc. Int. Soc. Music Inf. Retrieval Conf., 2020, pp. 733–739.

Some results

- Unsupervised (US) ≈ supervised (SU)
- Almost no drop of performances when using only 3% of the training data (US-F vs. US-S and SV-F vs. SV-S)

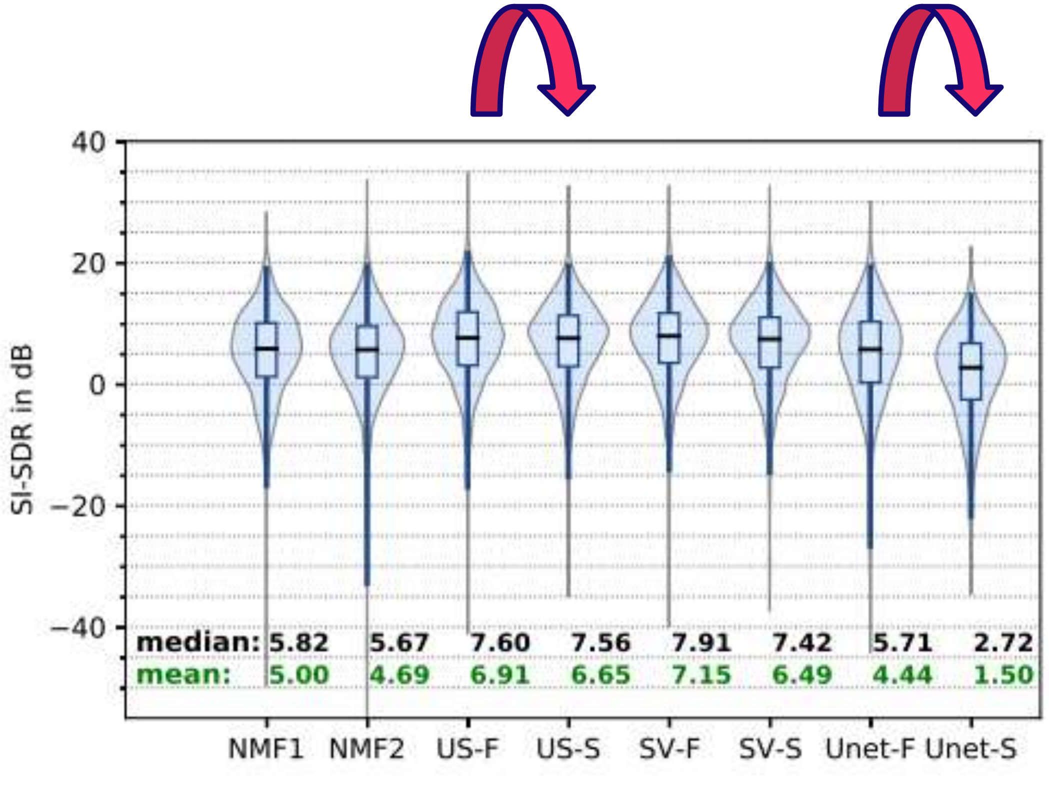


(b) $J = 4$ sources

NMF1: S. Ewert and M. M¨uller, "Using score-informed constraints for NMF- based source separation," in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing. IEEE, 2012, pp. 129–132.
NMF2: J.-L. Durrieu, B. David, and G. Richard, "A musically motivated mid- evel representation for pitch estimation and musical audio source separation," IEEE J. Selected Topics in Signal Processing, vol. 5, no. 6, pp. 1180–1191, 2011.
UNET: D. Petermann, P. Chandna, H. Cuesta, J. Bonada, and E. Gomez, "Deep learning based source separation applied to choir ensembles," in Proc. Int. Soc. Music Inf. Retrieval Conf., 2020, pp. 733–739.

Some results

- Unsupervised (US) ≈ supervised (SU)
- Almost no drop of performances when using only 3% of the training data (US-F vs. US-S and SV-F vs. SV-S)
- ..much larger drop of performances of the supervised baseline model (Unet)

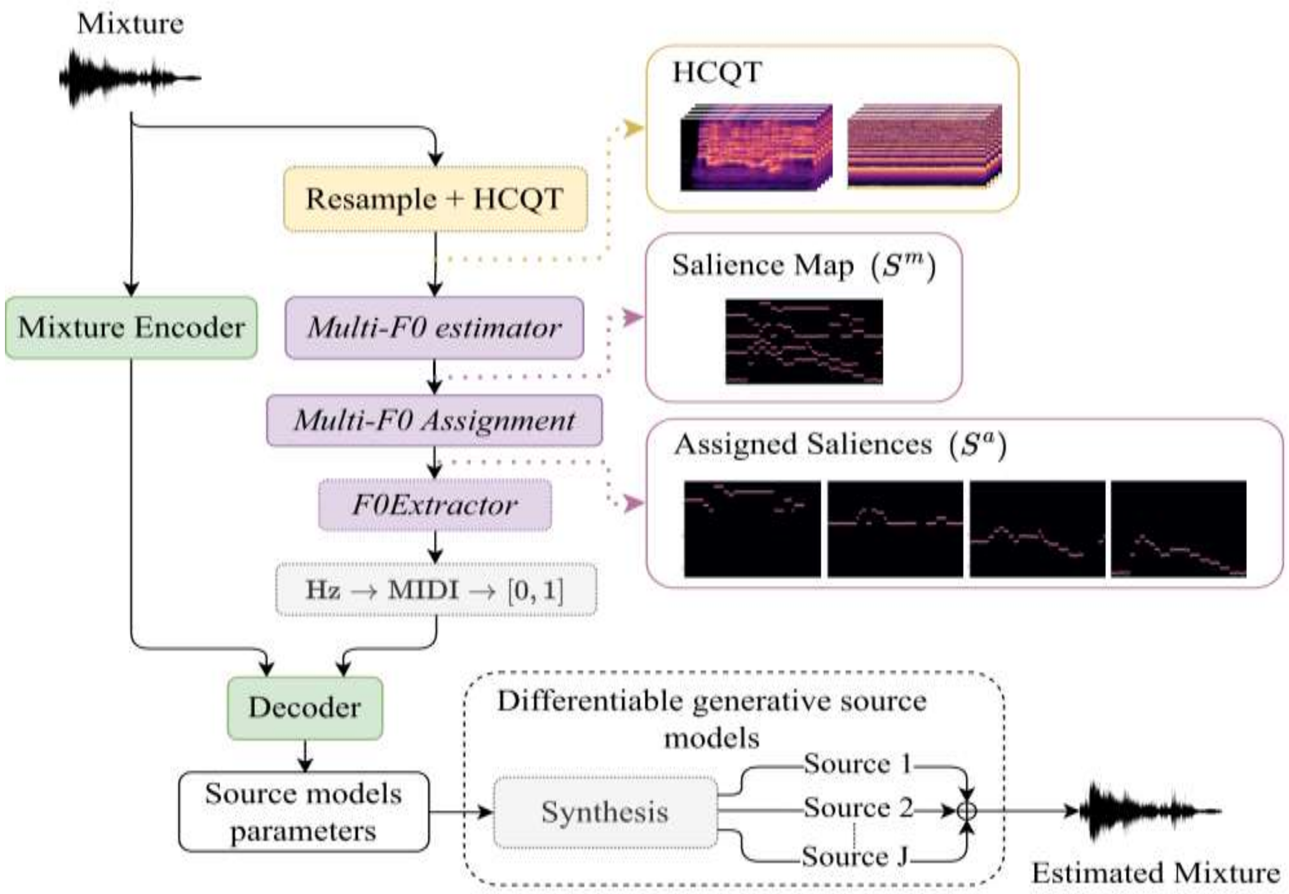


(b) $J = 4$ sources

NMF1: S. Ewert and M. Mueller, "Using score-informed constraints for NMF- based source separation," in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing. IEEE, 2012, pp. 129–132.
NMF2: J.-L. Durrieu, B. David, and G. Richard, "A musically motivated mid- evel representation for pitch estimation and musical audio source separation," IEEE J. Selected Topics in Signal Processing, vol. 5, no. 6, pp. 1180–1191, 2011.
UNET: D. Petermann, P. Chandna, H. Cuesta, J. Bonada, and E. Gomez, "Deep learning based source separation applied to choir ensembles," in Proc. Int. Soc. Music Inf. Retrieval Conf., 2020, pp. 733–739.

Towards a fully differentiable model for unsupervised singing voice separation

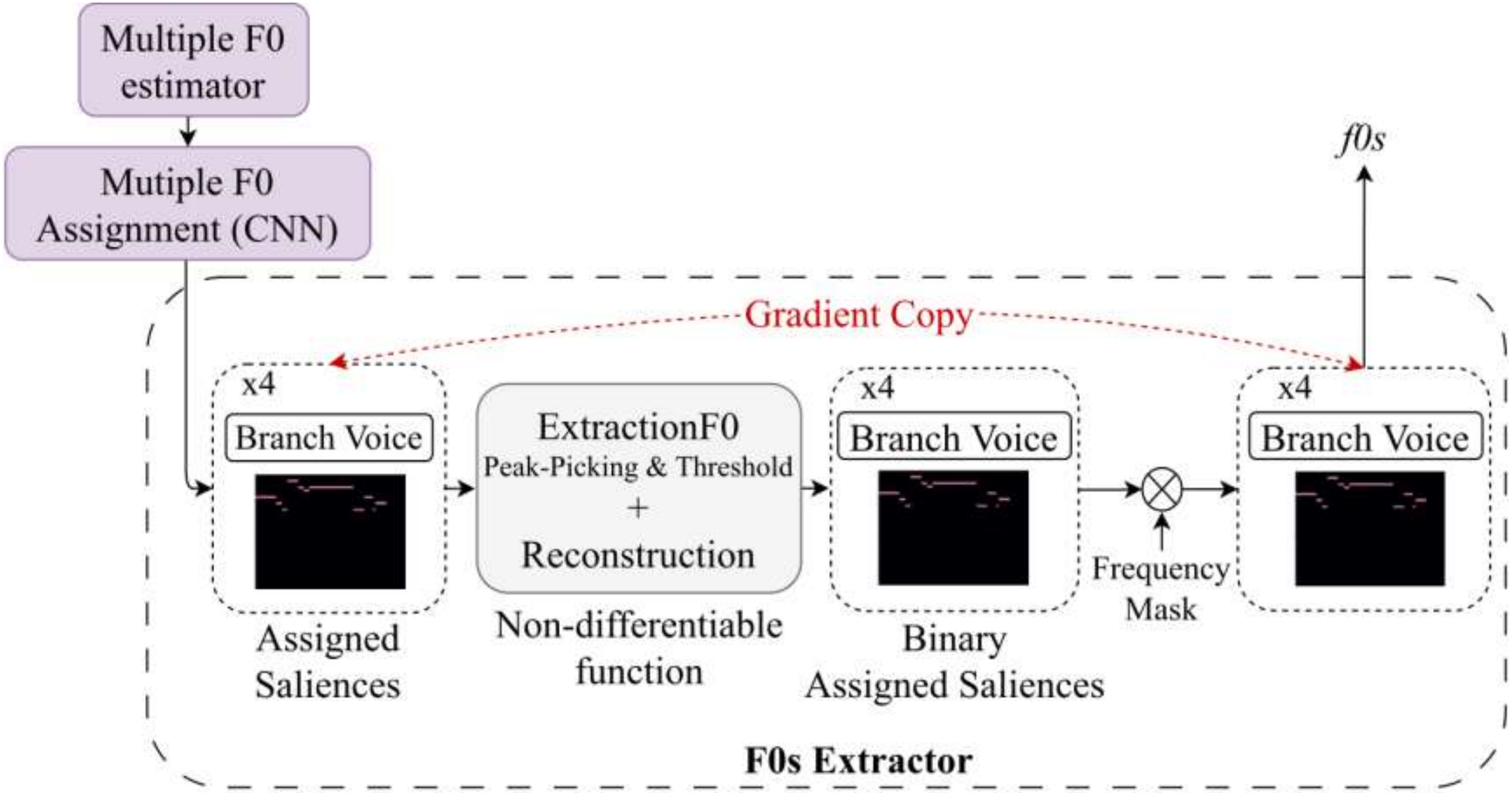
- Integration of multi-F0 extractor and automatic voice assignment



G. Richard, P. Chouteau, B. Torres A fully differentiable model for unsupervised singing voice separation, . *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr 2024, Seoul, South Korea.

Towards a fully differentiable model for unsupervised singing voice separation

- Extraction of F0 sequences from assigned salience maps.



G. Richard, P. Chouteau, B. Torres A fully differentiable model for unsupervised singing voice separation, . *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr 2024, Seoul, South Korea.

G. Richard

Model-based audio deep learning

Towards a fully differentiable model for unsupervised singing voice separation

- End-to-end approach less accurate than the baseline semi-integrated approach
  - *Train data: Bach Chorales-Barbershop Quartet (BCBSQ)*
  - *Test data: Choral Singing Dataset (CSD)*

Model	SI_SDR [dB]		OA [%]		RPA [%]		RCA [%]	
	μ	Md	μ	Md	μ	Md	μ	Md
UMSS [1]	**6.91**	**7.60**	-	-	-	-	-	-
U-Net [21]	4.44	5.71	-	-	-	-	-	-
$S_F S_F$	2.93	3.59	66	68	72	75	73	77
$S_{FT} S_{FT}$	4.81	6.07	73	79	80	87	82	88
$S_F S_{FT}$	5.77	6.46	78	82	85	90	85	89
W_{UP}	6.20	6.91	79	84	87	91	88	92

- … but much more robust on out of domain data
  - *Train data: Bach Chorales-Barbershop Quartet (BCBSQ) or BC1Song (e.G. reduced BCBSQ)*
  - *Test data: Cantoria*

Model	BC1Song		BCBSQ	
	μ	Md	μ	Md
UMSS [1]	0.31	0.73	0.86	1.38
U-Net [21]	-2.31	-2.07	0.97	1.47
W_{UP}	**1.93**	**2.61**	**3.29**	**3.79**

G. Richard, P. Chouteau, B. Torres A fully differentiable model for unsupervised singing voice separation, . *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr 2024, Seoul, South Korea.

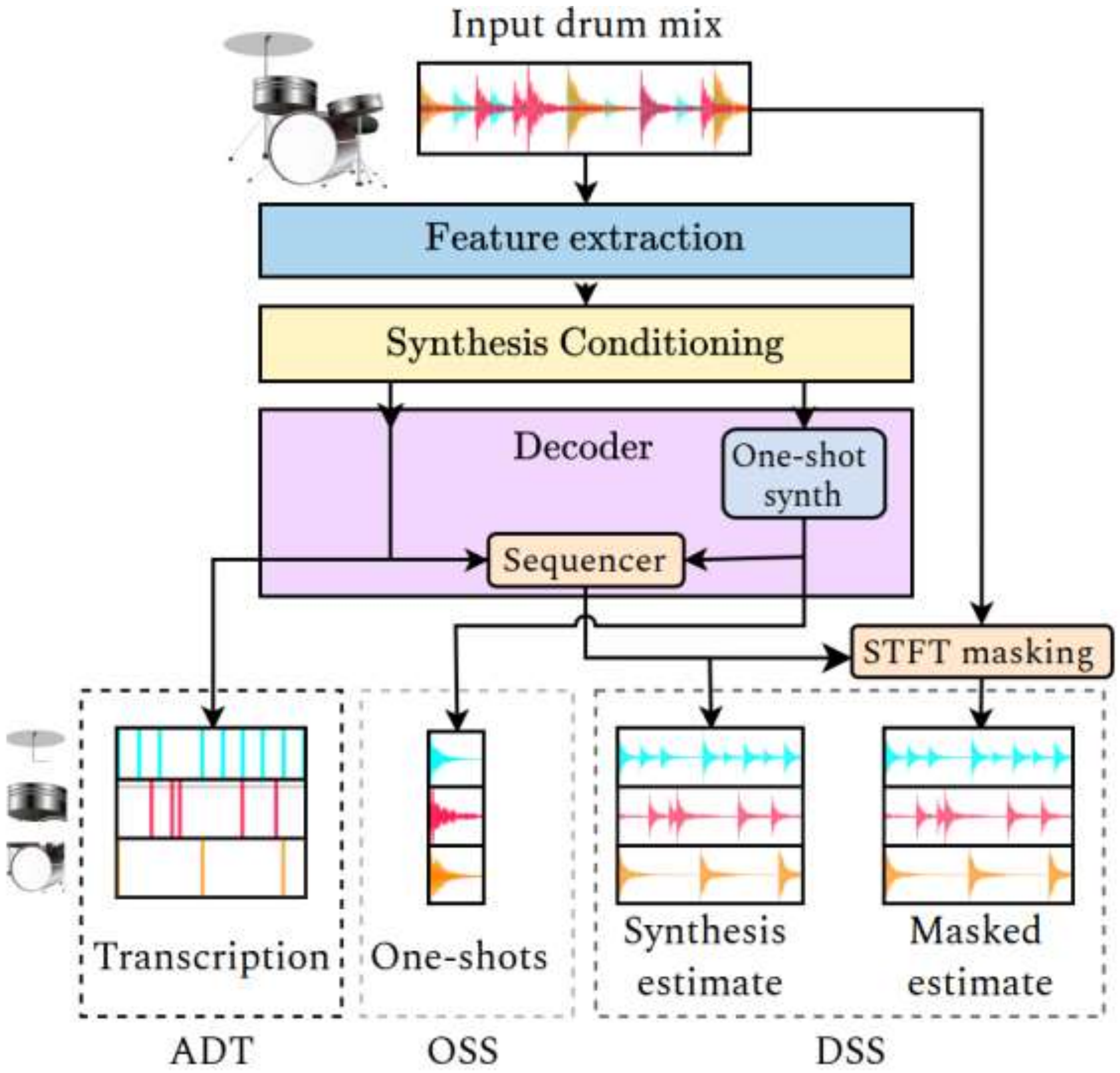
A short audio demo and some take aways

- **A short demo at**
- https://schufo.github.io/umss/
  - Ou local link
- **And for the fully differentiable model at:**
- https://pierrechouteau.github.io/umss_icassp/audio

Another example with Drum Source Separation

Inverse Drum machine

- A novel analysis-by-synthesis framework for Drum Source Separation (DSS)
  - works without isolated stems, relying only on transcription data for training.
- A jointly trained model that unifies Automatic Drum Transcription (ADT) and One-shot drum Sample Synthesis (OSS) in a single end-to-end system.
- A modular separation model that achieves separation quality comparable to supervised, state-of-the-art methods while using ≈ 100 times fewer parameters.



B. Torres, G. Peeters and G. Richard, "The Inverse Drum Machine: Source Separation Through Joint Transcription and Analysis-by-Synthesis," in *IEEE Transactions on Audio, Speech and Language Processing*, vol. 34, pp. 84-95, 2026, Preprint accessible at:https://hal.science/hal-05056592/document

G. Richard

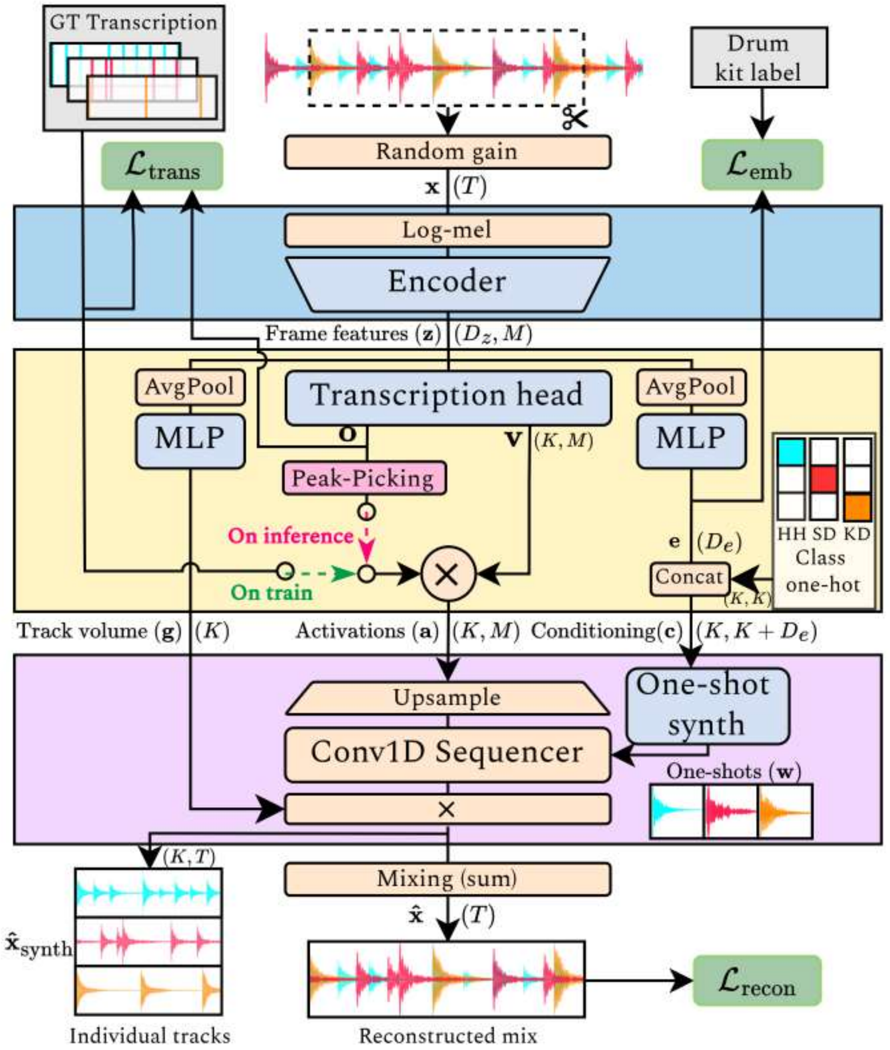
Model-based audio deep learning

Inverse Drum machine : a Multitask learning for Drum Source Separation

1. **Automatic Drum Transcription (ADT):** The precise estimation of the onset times of each drum instrument is achieved by training a transcription head to predict onset activations.

2. **One-shot drum Sample Synthesis (OSS):** High-quality one-shot samples for each drum instrument are generated by a Temporal Convolutional Network (TCN) conditioned on instrument type and mixture embedding.

3. **Drum Source Separation (DSS):** Individual drum tracks are extracted from the mixture by sequencing the synthesized one-shot samples with the estimated transcription.



B. Torres, G. Peeters and G. Richard, "The Inverse Drum Machine: Source Separation Through Joint Transcription and Analysis-by-Synthesis," in *IEEE Transactions on Audio, Speech and Language Processing*, vol. 34, pp. 84-95, 2026, Preprint accessible at:https://hal.science/hal-05056592/document

Inverse Drum machine : Training

- Training: end-to-end training using 3 combined losses

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{trans}} + \mathcal{L}_{\text{emb}}$$

- **Reconstruction loss :** The input mixture x is modelled by recomposing the individual drum tracks by sequencing onset activations with generated one-shot samples. Individual tracks are mixed together to obtain a reconstructed mixture xˆsynth.

$$\mathcal{L}_{\text{recon}}(\mathbf{x}, \hat{\mathbf{x}}_{\text{synth}}) = \sum_{\gamma \in \Gamma} \left\| |\mathbf{X}^{(\gamma)}| - |\hat{\mathbf{X}}^{(\gamma)}| \right\|_1 + \left\| \log(|\mathbf{X}^{(\gamma)}|) - \log(|\hat{\mathbf{X}}^{(\gamma)}|) \right\|_1$$

- **Transcription loss:** is the Binary Cross-Entropy loss between the estimated onsets and the ground-truth onsets for all drum instruments.

- **Mixture Embedding loss:** is essentially a drum kit classification loss, implemented as the Cross-Entropy between the estimated mixture embedding.

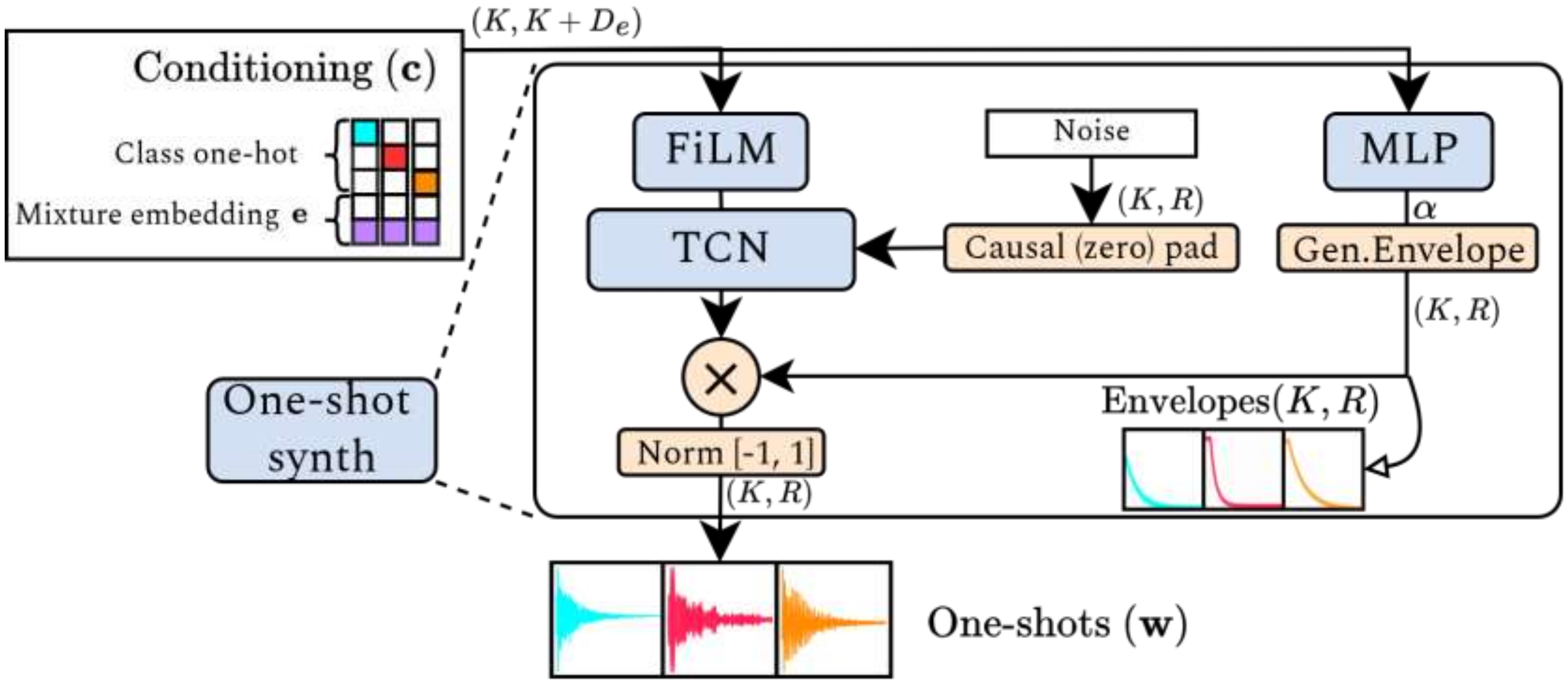
B. Torres, G. Peeters and G. Richard, "The Inverse Drum Machine: Source Separation Through Joint Transcription and Analysis-by-Synthesis," in *IEEE Transactions on Audio, Speech and Language Processing*, vol. 34, pp. 84-95, 2026, Preprint accessible at:https://hal.science/hal-05056592/document

G. Richard

Model-based audio deep learning

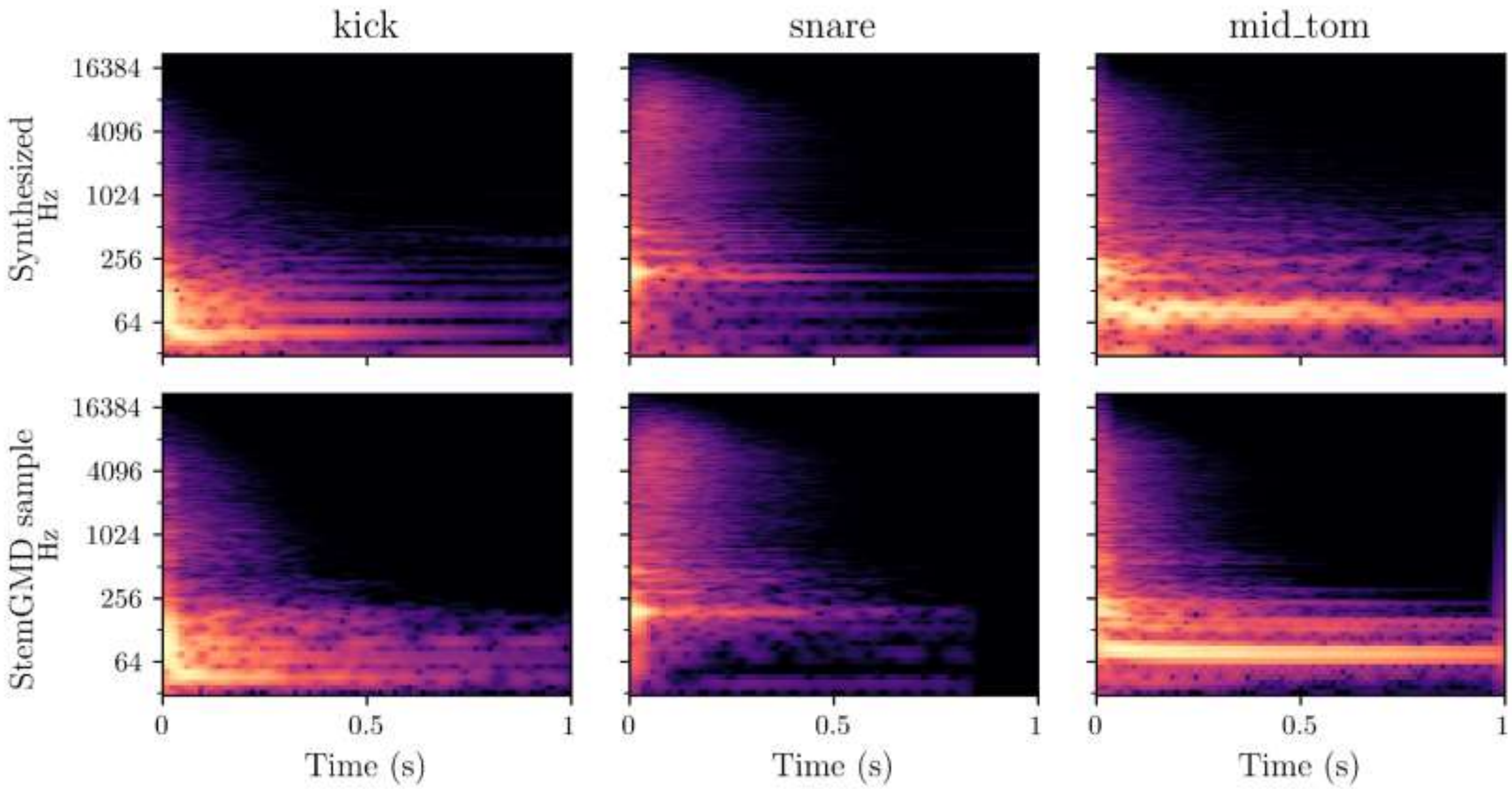
Inverse Drum machine

- A focus on the one-shot synthesis model
  - White noise is fed to a Temporal Convolutional Network (TCN) conditioned via Feature-wise Linear Modulation (FiLM) on a conditioning vector **c**, which has disentangled instrument class/timbre dimensions.



B. Torres, G. Peeters and G. Richard, "The Inverse Drum Machine: Source Separation Through Joint Transcription and Analysis-by-Synthesis," in *IEEE Transactions on Audio, Speech and Language Processing*, vol. 34, pp. 84-95, 2026, Preprint accessible at:https://hal.science/hal-05056592/document

Inverse Drum machine: some results



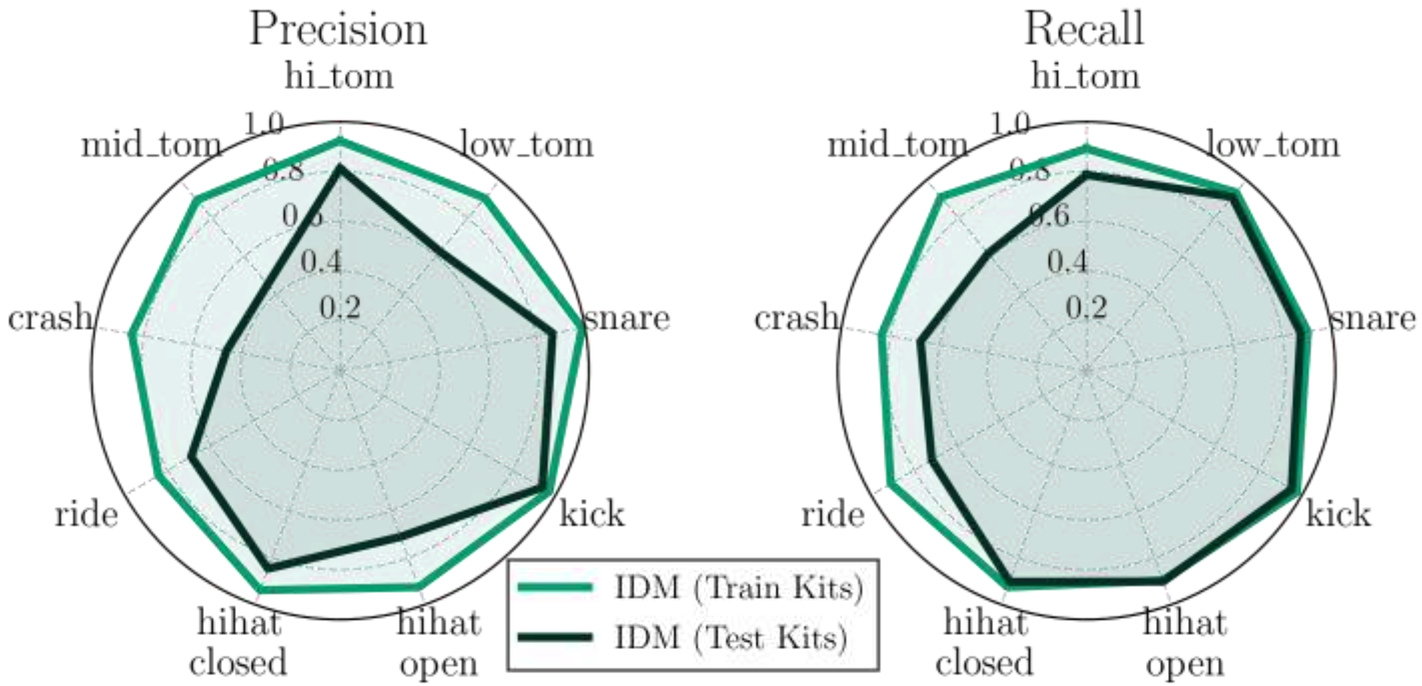
Log-magnitude spectrograms of synthesized, one-second-long one-shot synthesized (top) and real (bottom) samples for three instruments.

B. Torres, G. Peeters and G. Richard, "The Inverse Drum Machine: Source Separation Through Joint Transcription and Analysis-by-Synthesis," in *IEEE Transactions on Audio, Speech and Language Processing*, vol. 34, pp. 84-95, 2026, Preprint accessible at:https://hal.science/hal-05056592/document

G. Richard

Model-based audio deep learning

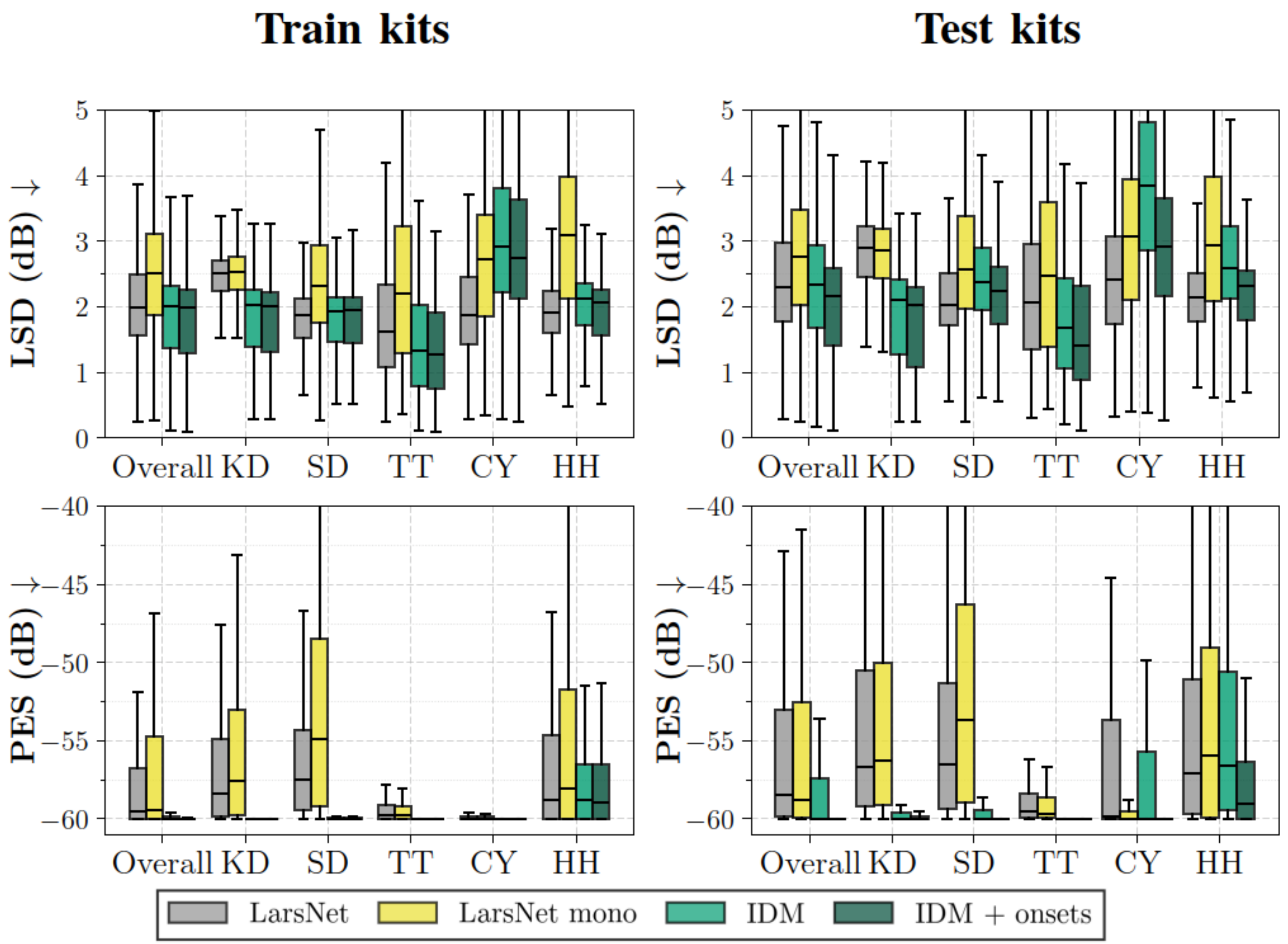
Inverse Drum machine: some results



Performance of the transcription module

B. Torres, G. Peeters and G. Richard, "The Inverse Drum Machine: Source Separation Through Joint Transcription and Analysis-by-Synthesis," in *IEEE Transactions on Audio, Speech and Language Processing*, vol. 34, pp. 84-95, 2026, Preprint accessible at:https://hal.science/hal-05056592/document

Inverse Drum machine: some results



Comparison of synthesis-based separation metrics

B. Torres, G. Peeters and G. Richard, "The Inverse Drum Machine: Source Separation Through Joint Transcription and Analysis-by-Synthesis," in *IEEE Transactions on Audio, Speech and Language Processing*, vol. 34, pp. 84-95, 2026, Preprint accessible at:https://hal.science/hal-05056592/document

Inverse Drum machine: demo

- **A full demo page at : https://bernardo-torres.github.io/projects/inverse-drum-machine/**

- **.. + code ..**

B. Torres, G. Peeters and G. Richard, "The Inverse Drum Machine: Source Separation Through Joint Transcription and Analysis-by-Synthesis," in *IEEE Transactions on Audio, Speech and Language Processing*, vol. 34, pp. 84-95, 2026, Preprint accessible at:https://hal.science/hal-05056592/document

To conclude

- As in many domains, the prominence of deep learning solutions is progressing …
- … but I believe in hybrid methods, hybrid deep learning … which bring
  - **Interpretability, Controllability, Explainability**
    - Hybrid model becomes controllable by human-understandable parameters
    - Hybrid model can lead to unsupervised methods
  - **Frugality: gain of several orders of magnitude** in the need of data and model complexity
  - **Can be applied to many audio processing problems**
    - Exploiting room acoustics for Audio dereverberation [1],
    - Exploiting physical/signal models for music synthesis [2],
    - Exploiting "audio class specific" codebooks for audio compression and separation [3]
    - Exploiting key speech attributes for controlled speech synthesis and transformation [4]
    - …

[1] Louis Bahrman, Mathieu Fontaine, Gael Richard. A Hybrid Model for Weakly-Supervised Speech Dereverberation. *IEEE ICASSP 2025*, ⟨hal-04931672⟩
[2] Lenny Renault, Rémi Mignot, Axel Roebel. Differentiable Piano Model for MIDI-to-Audio Performance Synthesis. Int. Conf.on Digital Audio Effects (DAFx20in22), Sep 2022, Vienna,
[3] Xiaoyu Bie, Xubo Liu, Gaël Richard. Learning Source Disentanglement in Neural Audio Codec. *IEEE ICASSP 2025* , ⟨hal-04902131⟩
[4] Samir Sadok, Simon Leglaive, Laurent Girin, Gaël Richard, Xavier Alameda-Pineda. AnCoGen: Analysis, Control and Generation of Speech with a Masked Autoencoder. *IEEE ICASSP 2025* , ⟨hal-04891286⟩

G. Richard

Model-based audio deep learning

Thank you !!

Hi-AUDiO