Exploiting knowledge for model-based deep music generation

Gaël RICHARD*

Professor, Telecom Paris, Institut polytechnique de Paris

MBZUAI Workshop 2025

*work with collaborators and in particular **K. Schulze-Forster, M. Agarwal, T. Baoueb,** X. Bie, C. Wang, C. Doire, L. Kelley, B. Torres, P. Chouteau, R. Badeau

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

- **Context and motivation**
- **Towards hybrid (or model-based) deep learning**
  - Some examples in other domains
  - Hybrid deep learning in audio
- **Specific examples in**
  - Unsupervised music source separation
  - Symbolic music generation with transformers
  - Music timbre transfer
- **Conclusion**

G. Richard

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 (>> 100 millions parameters)*
  - Overconsumption regime
  - Non-interpretable/non-controllable

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 (> 100 millions parameters)*
  - Overconsumption regime
  - Non-interpretable/non-controllable
- **The main goal of the project :** Hi-AUDiO

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

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

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

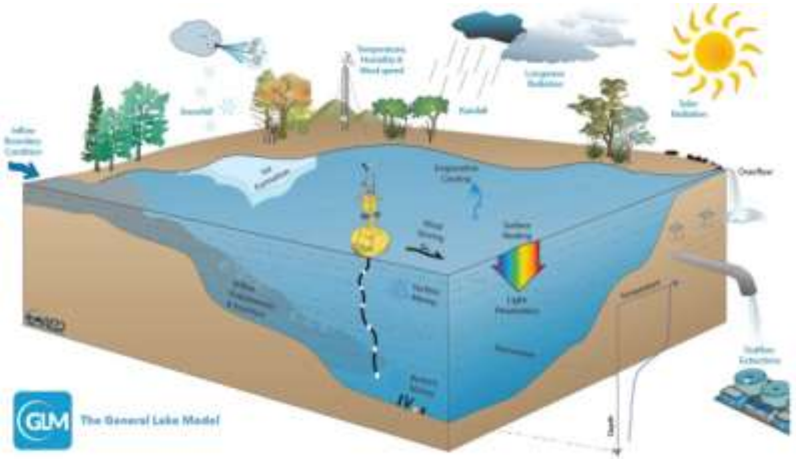
G. Richard

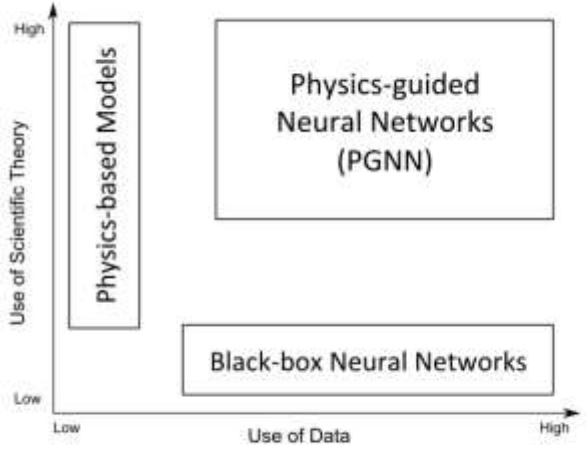
Exploiting knowledge for model-based deep music generation

Towards Hybrid *(or model-based)* deep learning

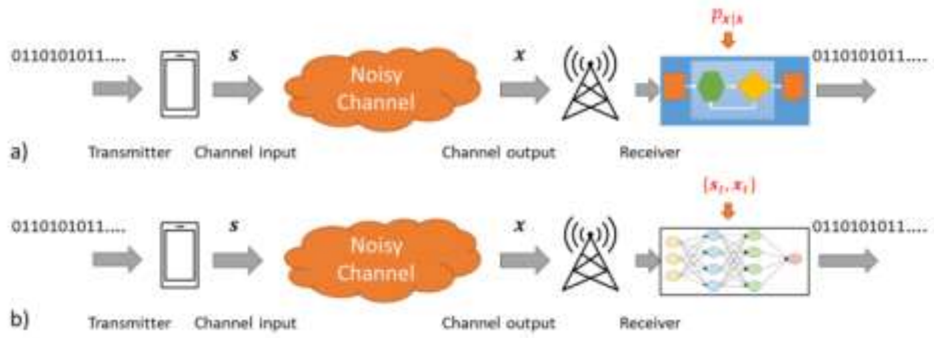
… some prior works.

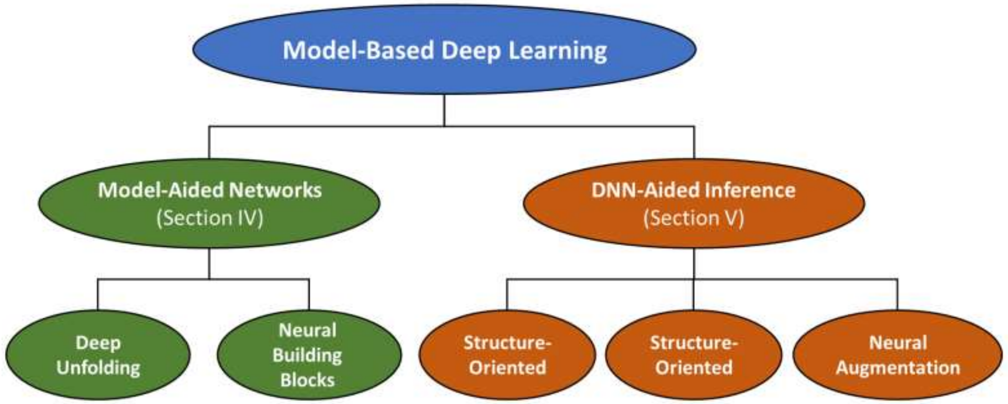
- Physics-guided neural networks in remote sensing [1],





- Digital communication and Image restoration [2,3]





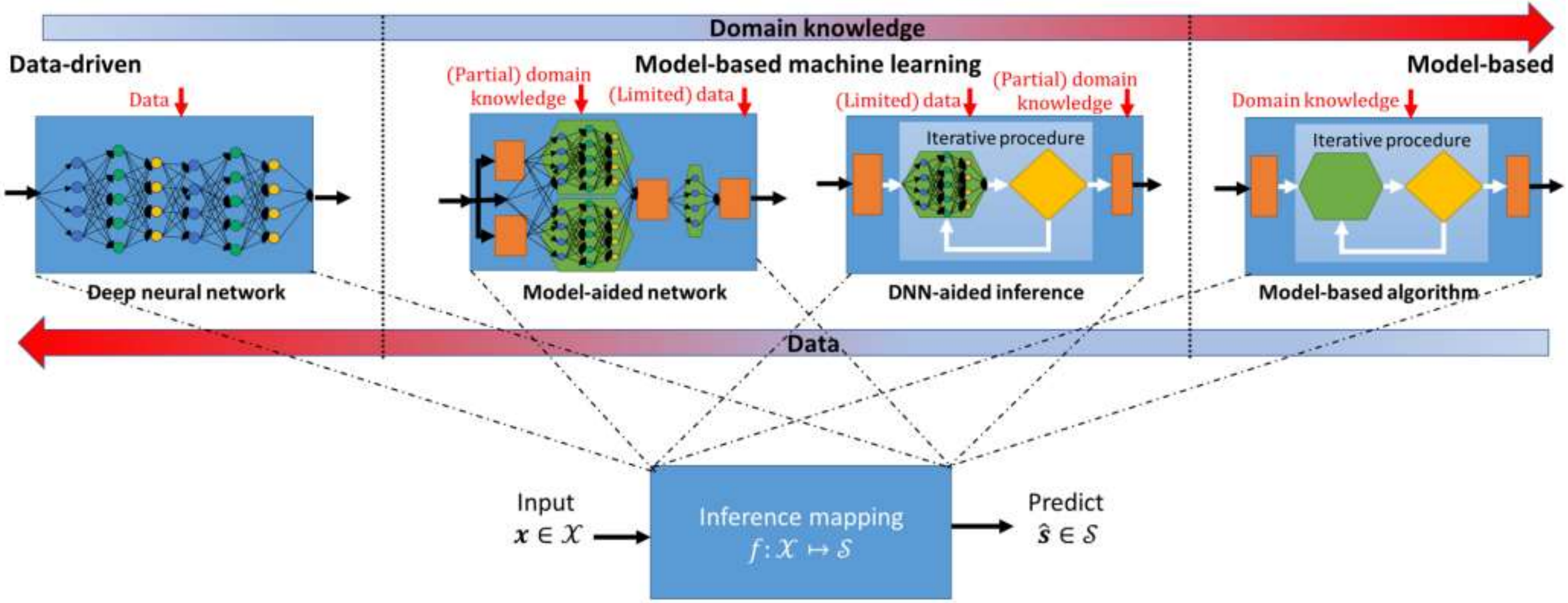
[1] A. Karpatne & al. "Physics-guided Neural Networks (PGNN): An Application in Lake Temperature Modeling," arXiv, 1710.11431, 2017.
[2] B. Lecouat & al., "Fully Trainable and Interpretable Non-Local Sparse Models for Image Restoration.," 2020. ⟨hal-02414291v2⟩.
[3] N. Shlezinger, & al., "Model-Based Deep Learning," in *Proceedings of the IEEE*, vol. 111, no. 5, pp. 465-499, May 2023,

G. Richard

Exploiting knowledge for model-based deep music generation

Towards Hybrid *(or model-based)* deep learning
… some prior works.

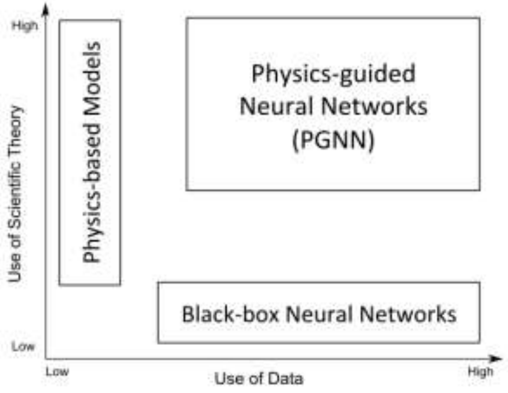
- Illustration of model-based versus data-driven inference (*from [3]*)



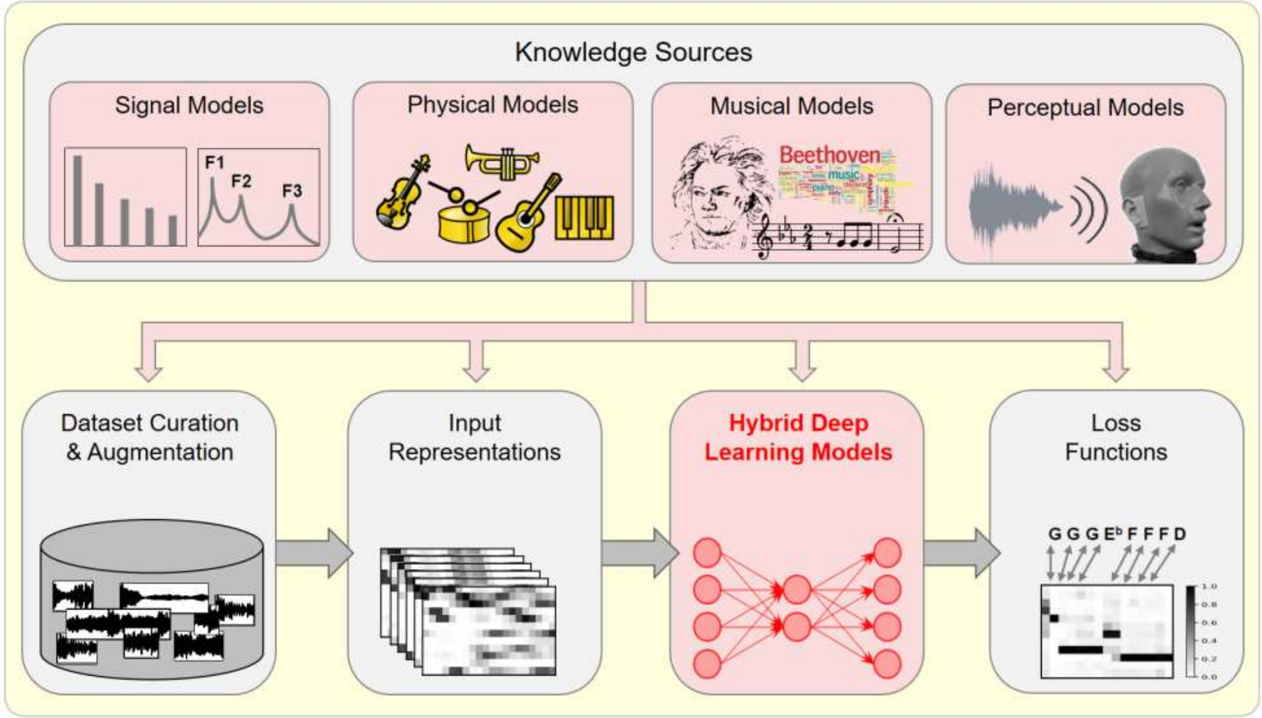
[3] N. Shlezinger, & al., "Model-Based Deep Learning," in *Proceedings of the IEEE*, vol. 111, no. 5, pp. 465-499, May 2023,

Towards model-based deep learning approaches

- 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

Exploiting knowledge for model-based deep music generation

Towards model-based deep learning

… some prior works in audio

- **Use of a model-based feature representation**
  - Non-Negative Matrix Factorization (NMF) models with CNNs for audio scene classification [1, 2]
- **Exploit the concept of deep unrolling**
  - **Deep NMF :** Converting one iteration of NMF (iterative algorithm) into one layer of a DNN [3]
- **Use DNN as noise estimator**
  - **Deep Griffin-Lim:** Each iteration of an iterative phase retrieval algorithm is « denoised » by DNN [4]
- ... Many other examples

[1] V. Bisot & al., "Feature Learning with Matrix Factorization Applied to Acoustic Scene Classification", ACM/IEEE Trans. on ASLP, vol. 25, no. 6, 2017
[2] V. Bisot & al., Leveraging deep neural networks with nonnegative representations for improved environmental sound classification IEEE International Workshop on Machine Learning for Signal Processing MLSP, Sep 2017, Tokyo,
[3] J. L. Roux & al., "Deep NMF for speech separation,," in IEEE Int. Conf. on Acous., Speech and Signal Proc. (ICASSP), 2015
[4] Y. Masuyama, K. Yatabe, Y. Koizumi, Y. Oikawa and N. Harada, "Deep Griffin–Lim Iteration: Trainable Iterative Phase Reconstruction Using Neural Network," in *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 1, pp. 37-50, Jan. 2021,

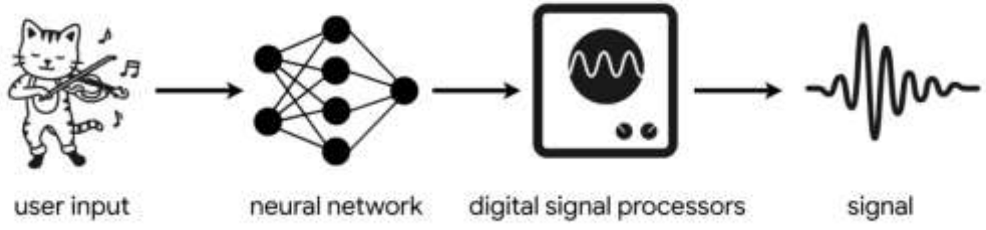
G. Richard

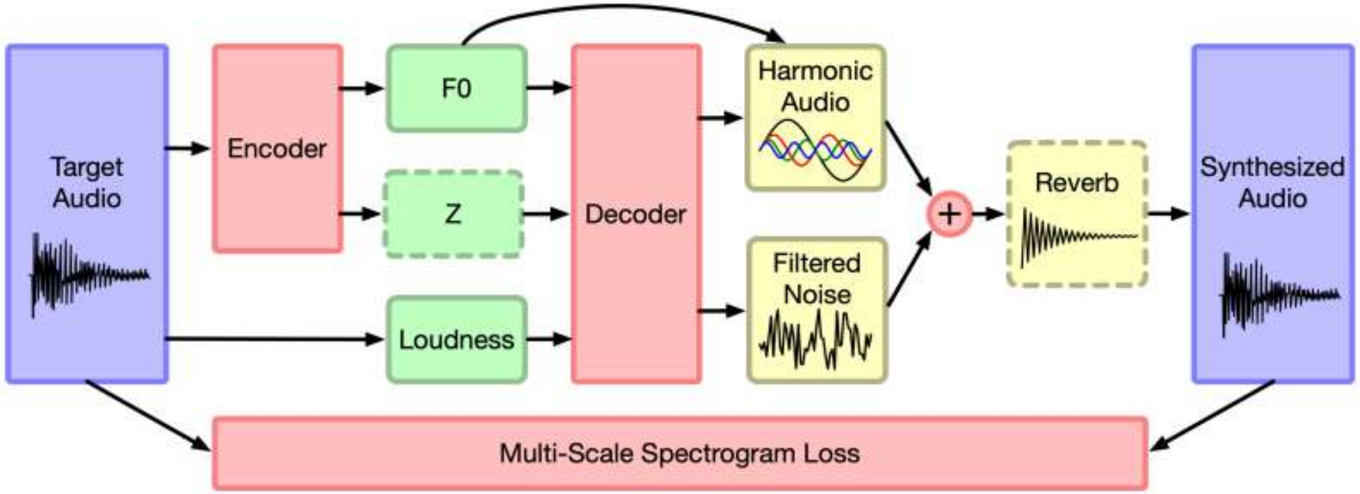
Exploiting knowledge for model-based deep music generation

Towards model-based deep learning

… some prior works in audio

- Coupling signal processing modules with deep learning for audio synthesis
- The example of DDSP (Engel & al.)



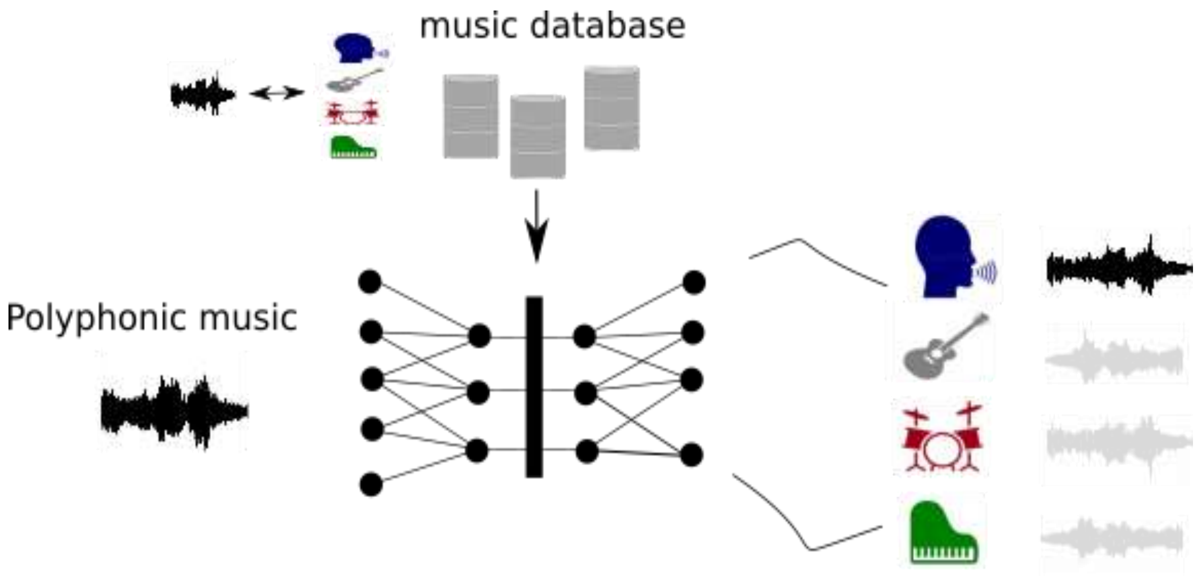


X. Wang & al. "Neural Source-Filter Waveform Models for Statistical Parametric Speech Synthesis," in IEEE/ACM Trans. on ASLP Proc., vol. 28, 2020.
J. Engel & al., "DDSP: Differentiable Digital Signal Processing," in Int. Conf. on Learning Representations (ICLR), 2020.

Towards model-based deep learning

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

- For example in music source separation



***Main limitations**:*

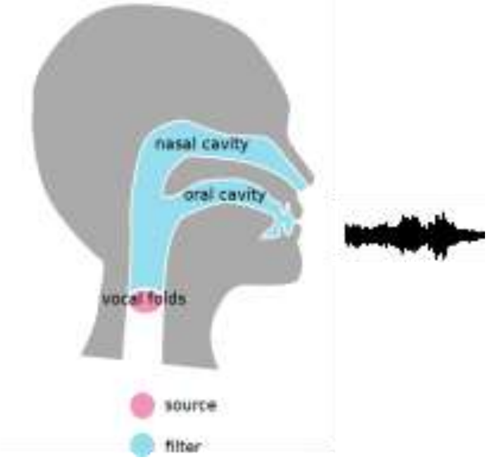
- *Difficulty to obtain « aligned » data*
- *Knowledge learned (only) from data*
- *Complexity: overparametrized models*
- Overconsumption regime
- **Non-interpretable/non-controllable**

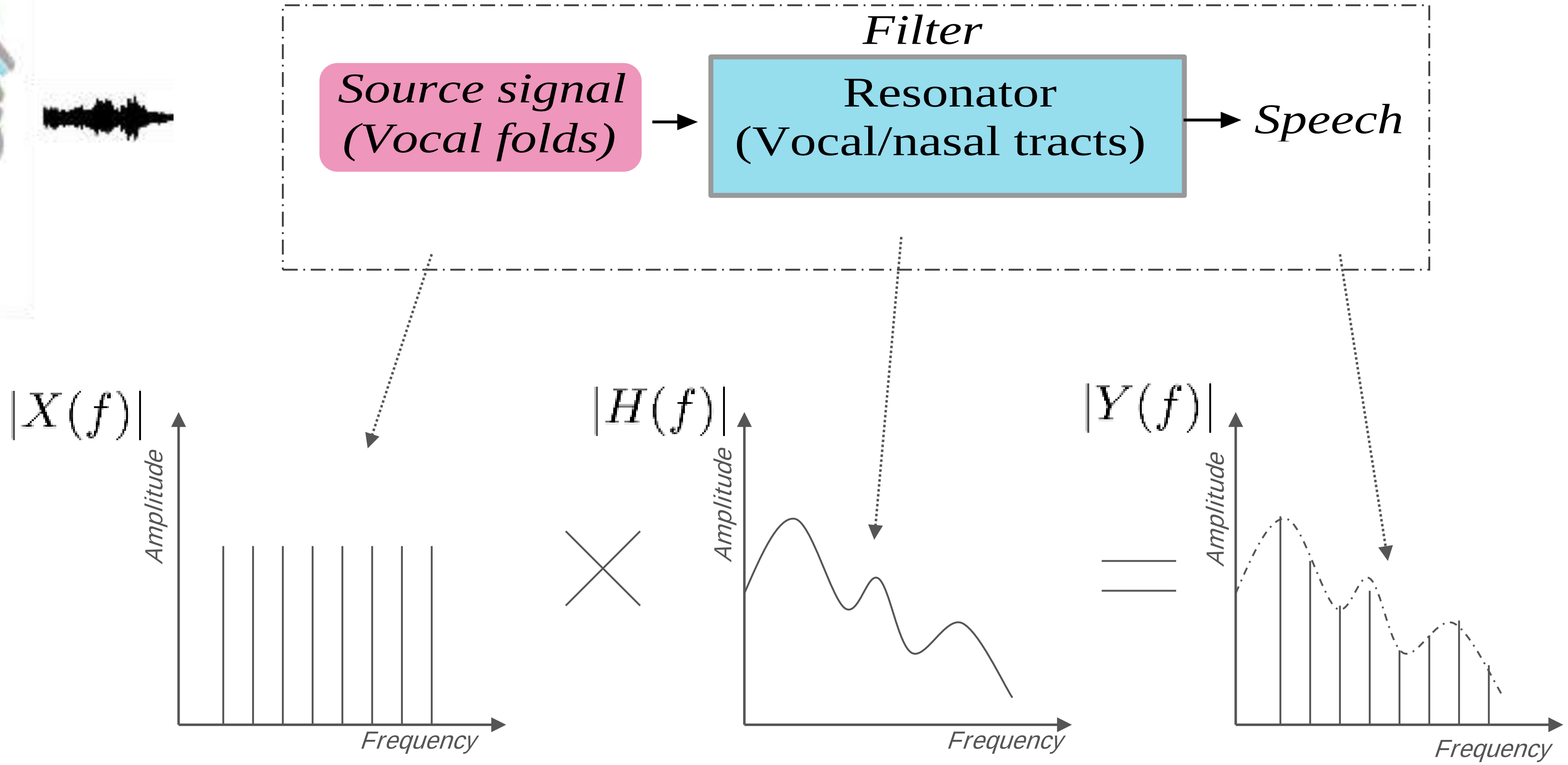
G. Richard

Exploiting knowledge for model-based deep music generation

The source filter model

an efficient speech production model





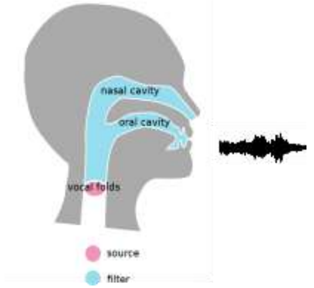
Fant, G. Acoustic theory of speech production, 1960, The Hague, The Netherlands, Mouton.

G. Richard

Towards model-based deep learning

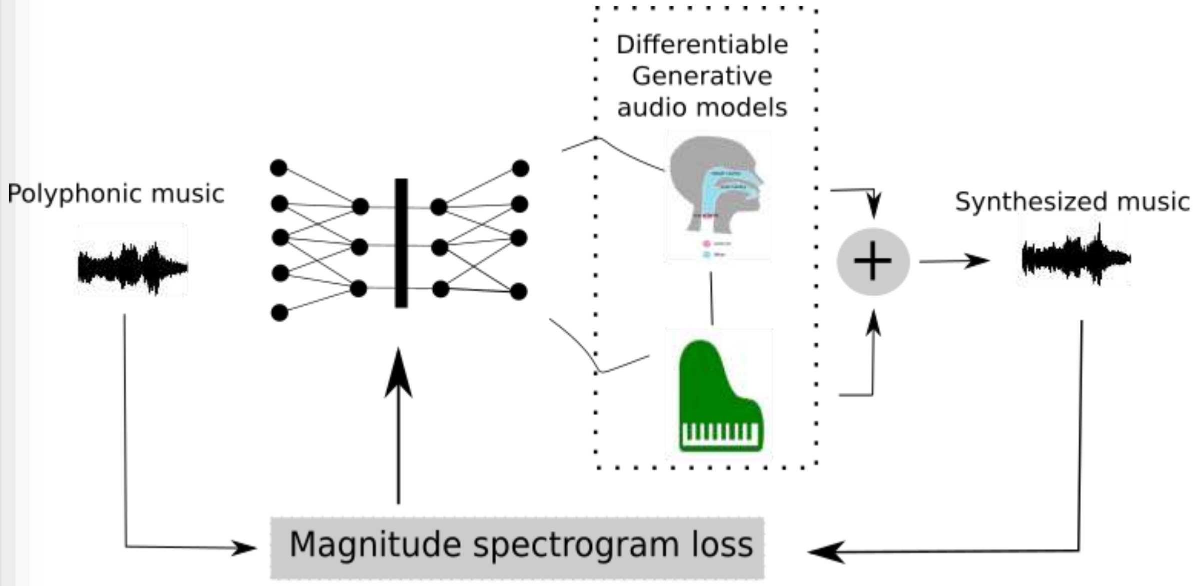
… by **integrating our prior knowledge** about the nature of the processed data.

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



Singing voice as a source / filter model :

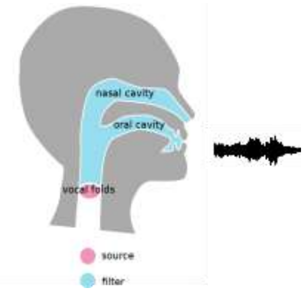
- source = vibration of vocal folds
- Filter = resonances of vocal/nasal cavities



Towards model-based deep learning

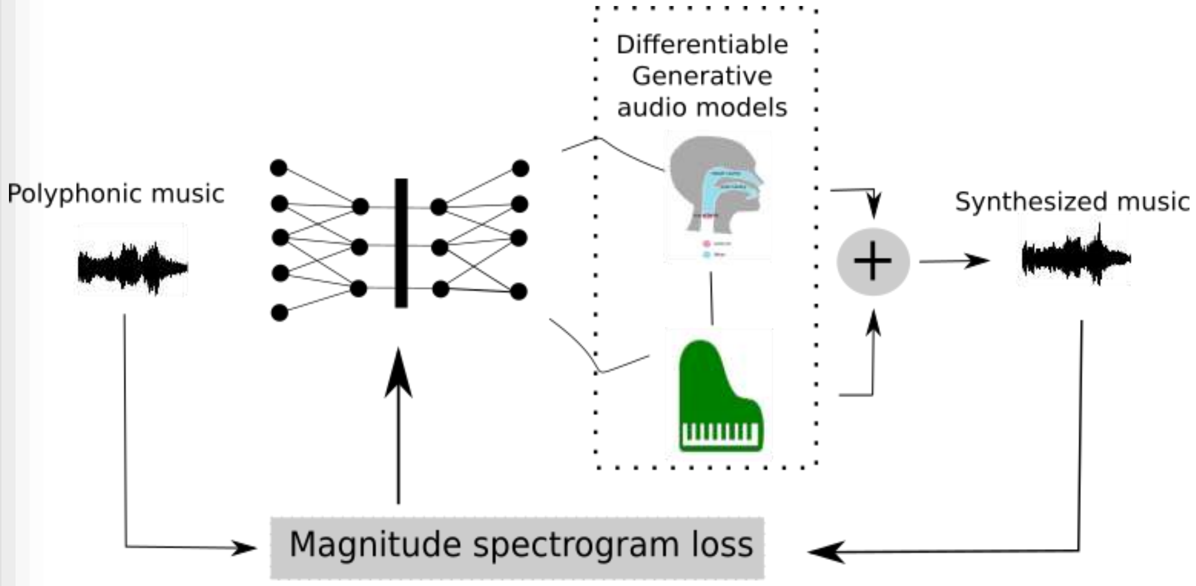
… by **integrating our prior knowledge** about the nature of the processed data.

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



A new paradigm

- Model is at the « core » of neural architecture
- Source separation **by synthesis** (*no interference from other sources*)
- Learning only from the polyphonic recording (*no need of the true individual tracks*)

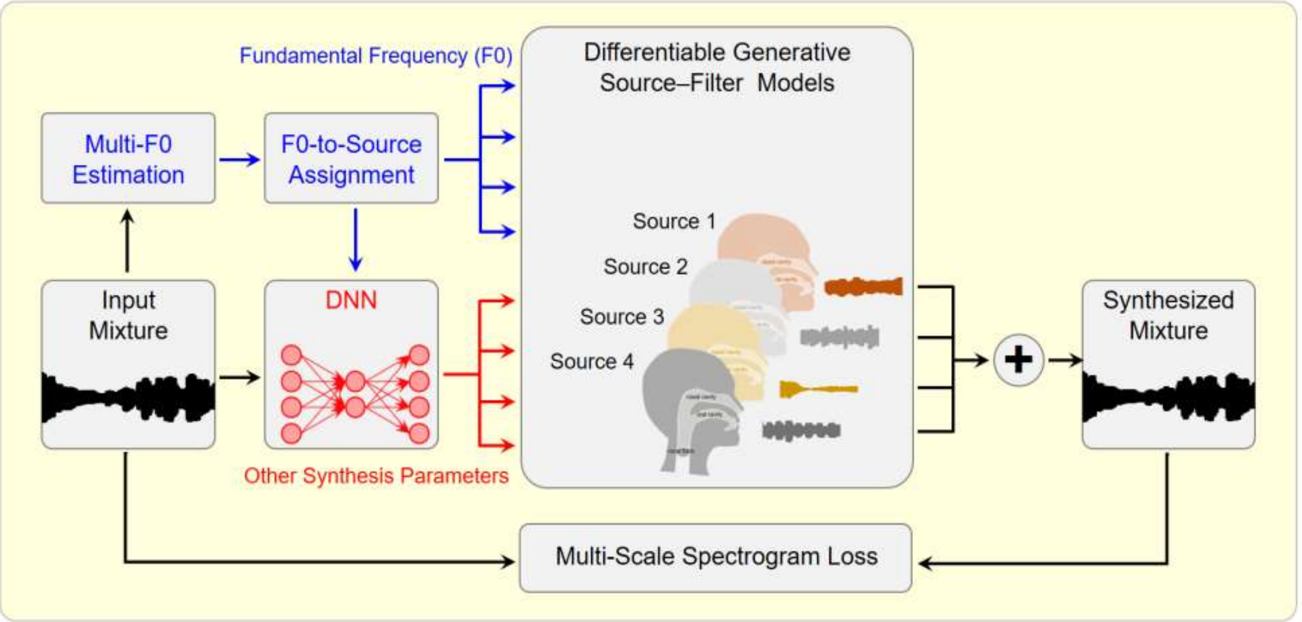
Novel sound transformation capabilities:

- Timbre/melody of the voice,
- Lyrics, translation
- Re-harmonization

Towards model-based deep learning

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

- An example for unsupervised singing voice separation



Highlights

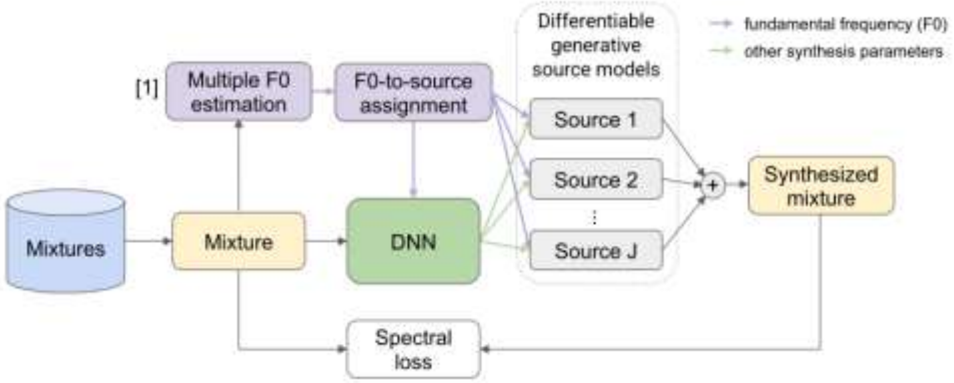
- Unsupervised :
  - Learning only from the polyphonic recording (*no need of the true individual tracks*)
- Homogeneous sources :
  - All sources have similar acoustic properties

K Schulze-Forster, G. Richard, L. Kelley, C. Doire, R Badeau Unsupervised Music Source Separation Using Differentiable Parametric Source Models, IEEE Trans. On AASP, 2023
G. Richard, V. Lostanlen, Y.-H. Yang, M. Müller, "Model-based Deep Learning for Music Information Research", IEEE Signal Processing Magazine - Special Issue on Model-based and Data-Driven Audio Signal Processing, 2025 (preprint accessible at: https://arxiv.org/abs/2406.11540)
Multi-F0 estimation from *'H. Cuesta, B. McFee, and E. Gómez. Multiple f0 estimation in vocal ensembles using convolutional neural networks. ISMIR, 2020.'*

G. Richard

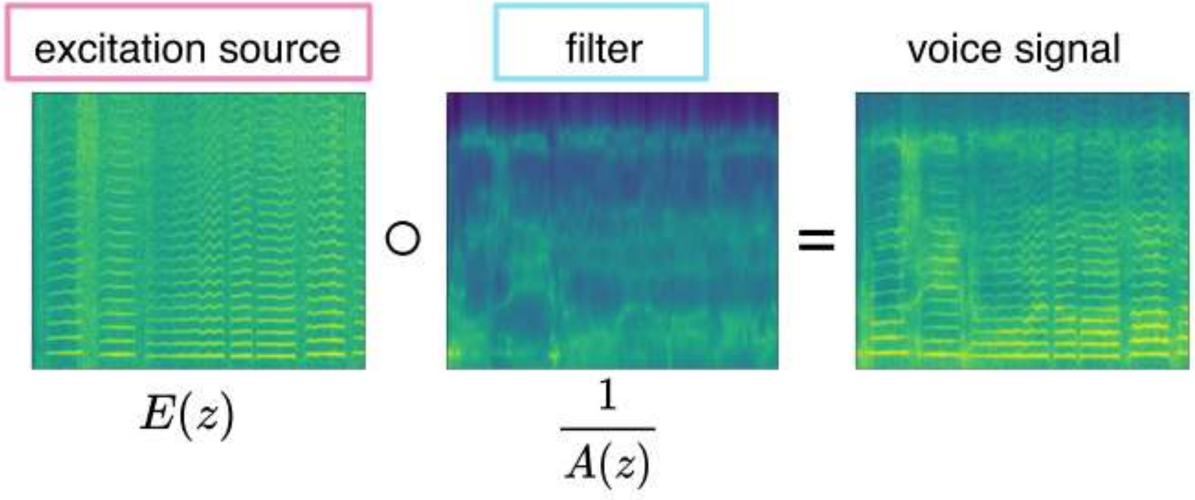
Exploiting knowledge for model-based deep music generation

Parametric source models



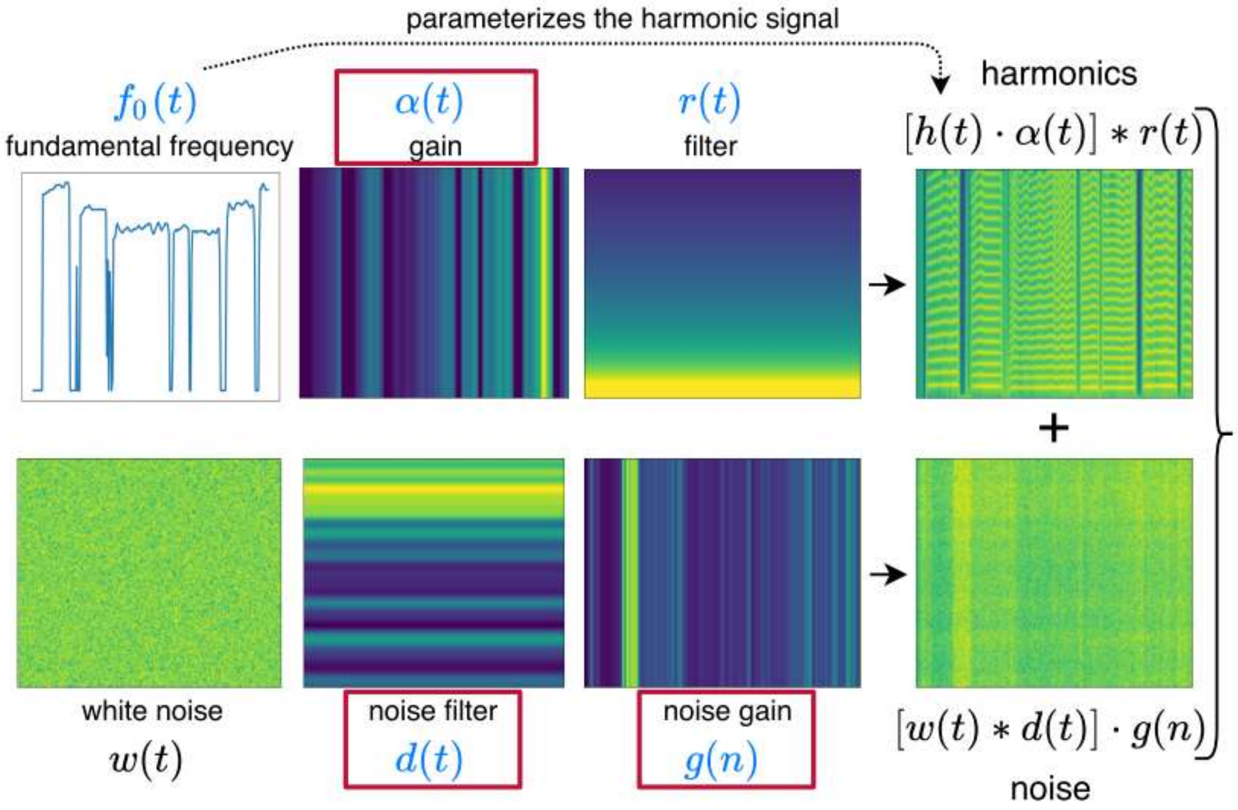
Singing voice as a source / filter model :

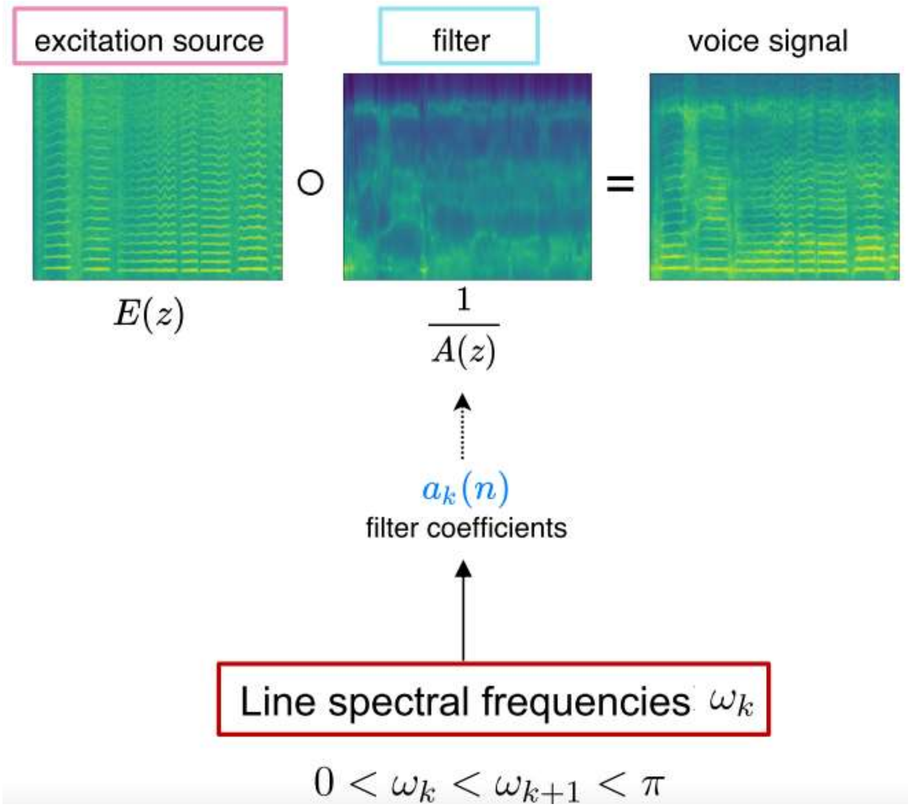
- source = vibration of vocal folds
- Filter = resonances of vocal/nasal cavities



$$E(z) \circ \frac{1}{A(z)} =$$

Parametric source models



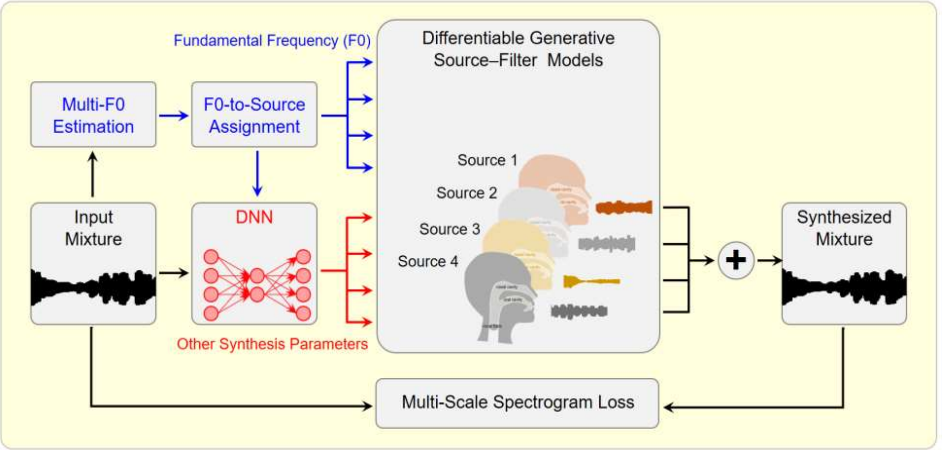


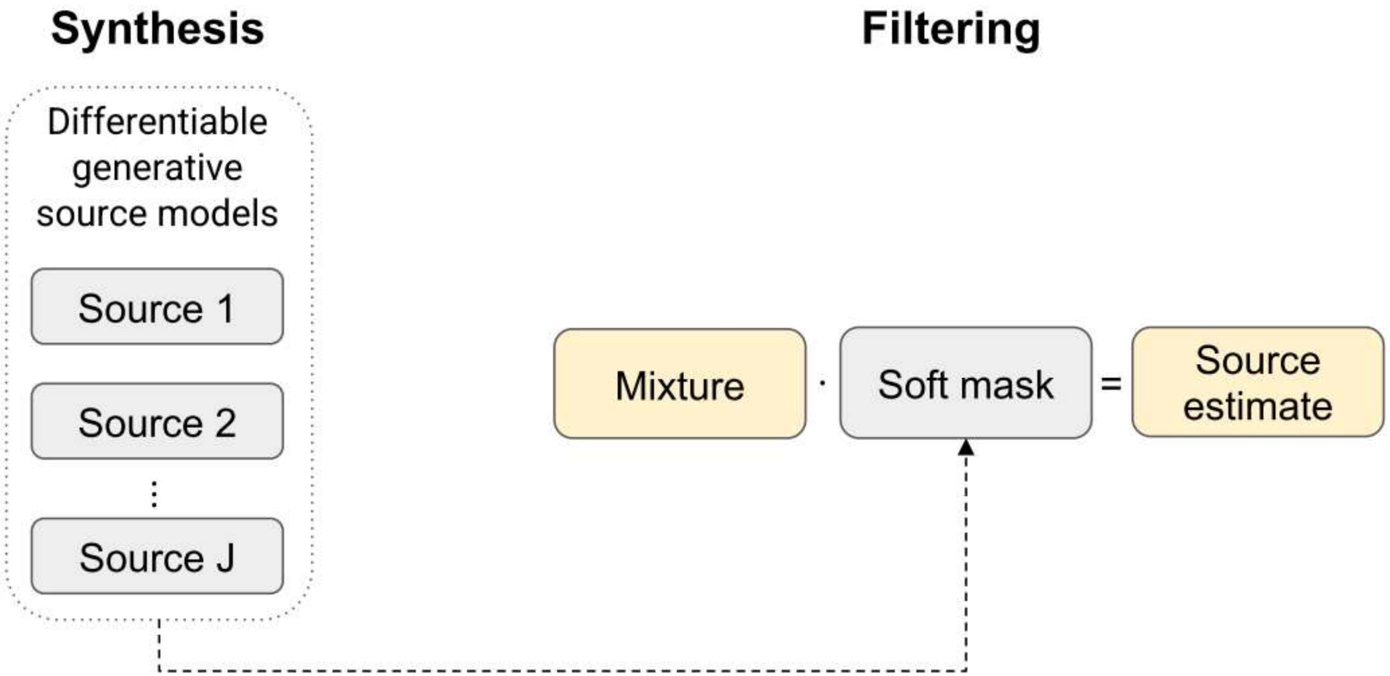
G. Richard

Exploiting knowledge for model-based deep music generation

G. Richard

Synthesis or filtering



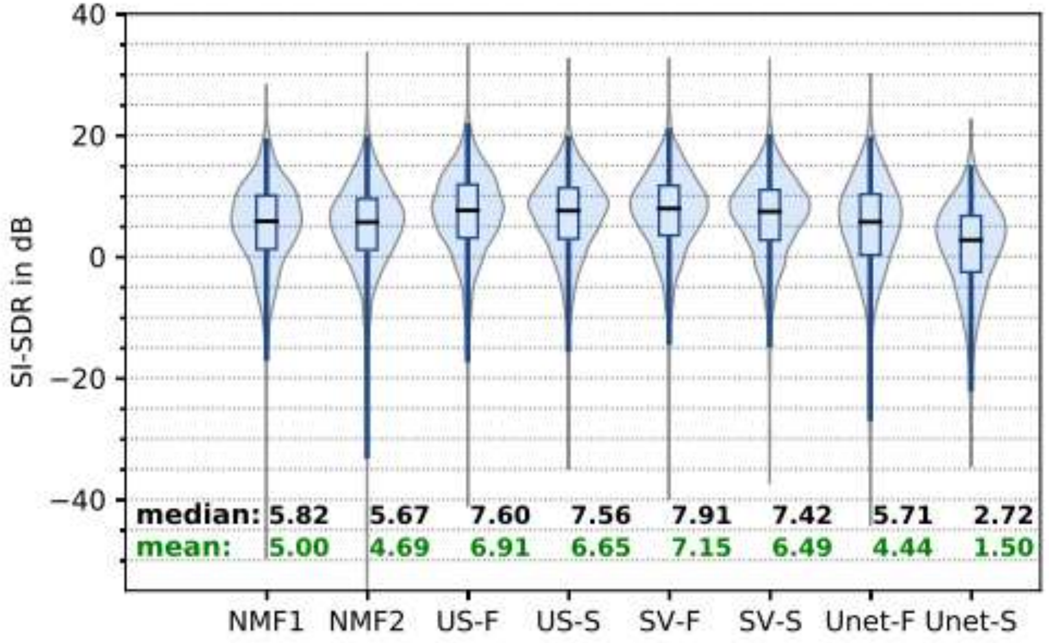


G. Richard

Exploiting knowledge for model-based deep music generation

Some results

- Unsupervised (US) ≈ supervised (SV)



(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.

TELECOM Paris

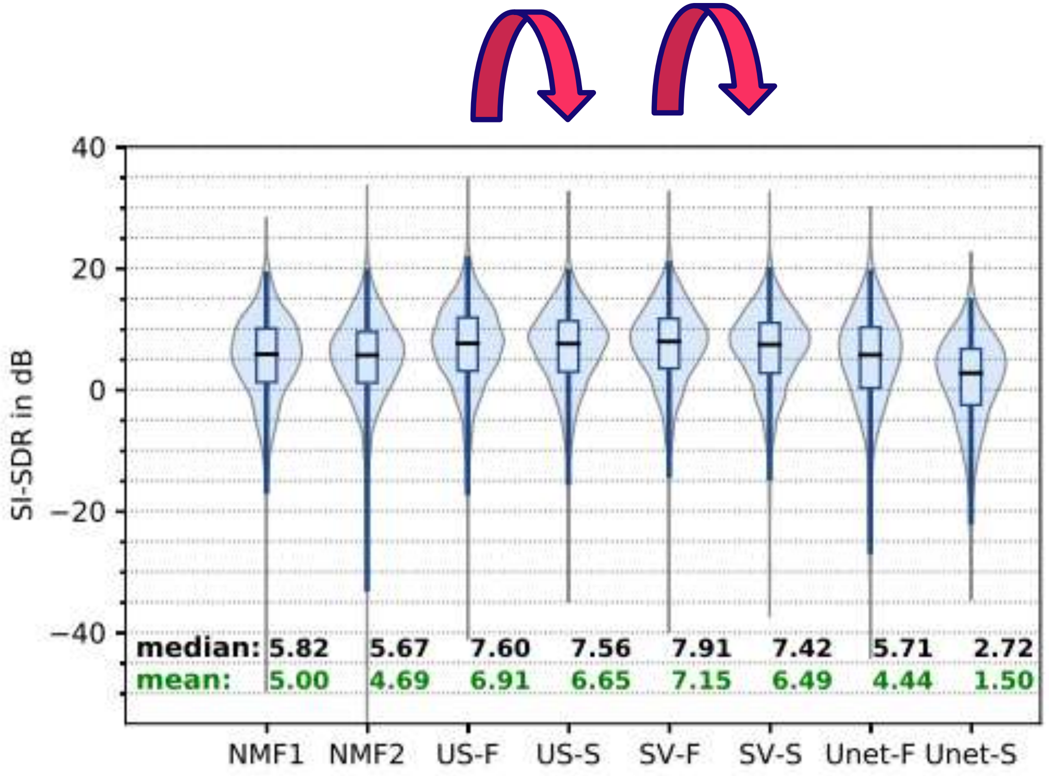
IP PARIS

G. Richard

Exploiting knowledge for model-based deep music generation

Some results

- Unsupervised (US) ≈ supervised (SV)
- 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.

Hi-AUDiO

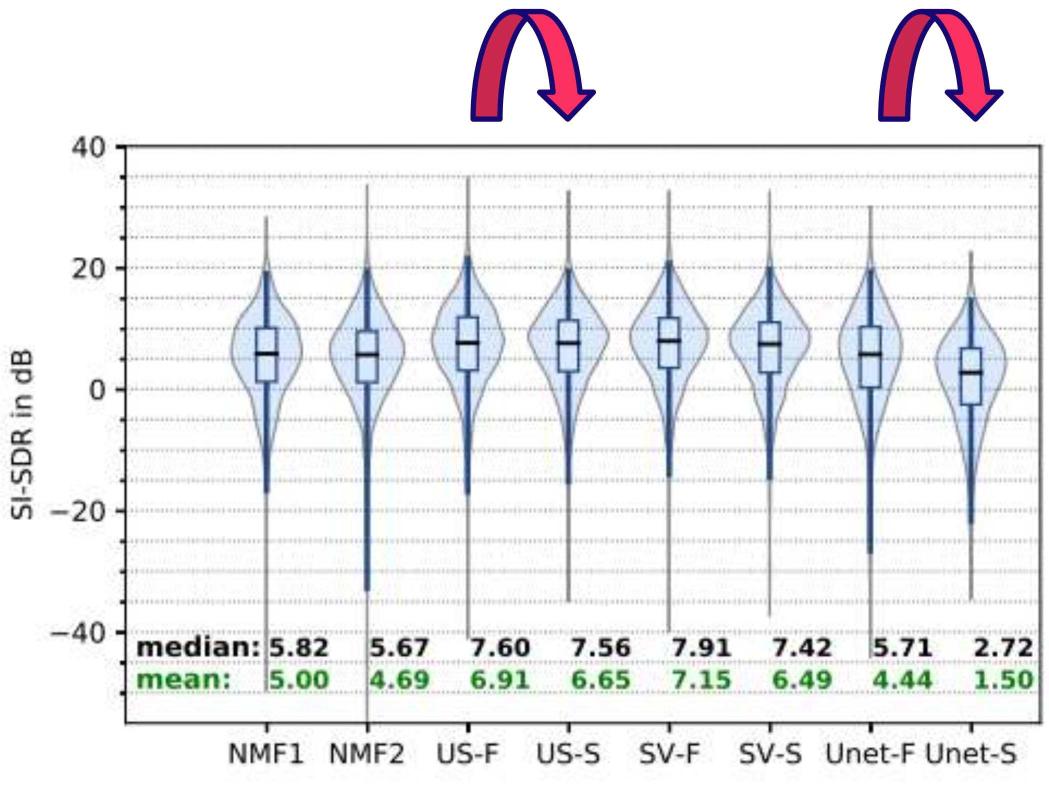
erc

G. Richard

Exploiting knowledge for model-based deep music generation

Some results

- Unsupervised (US) ≈ supervised (SV)
- 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.

A short audio demo and some take aways

G. Richard

Exploiting knowledge for model-based deep music generation

- **A short demo at**
- https://schufo.github.io/umss/
  - Or local link
- **Some take aways**
  - Only a small amount of data needed
  - Filtering the mixture better than synthesis
  - Differentiable stable all-pole filter
  - Parameterization of the mixture is provided
  - Extension possible to a fully end-to-end approach [1]

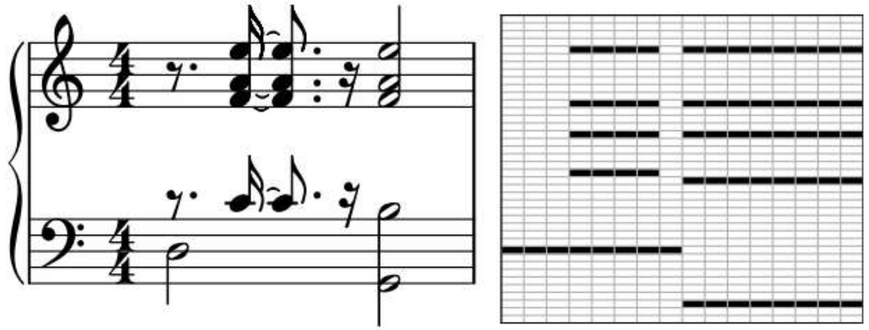
[1] G. Richard, P. Chouteau, B. Torres, A fully differentiable model for unsupervised singing voice separation, ICASSP 2024, with demo at https://pierrechouteau.github.io/umss_icassp/audio

Symbolic music generation with transformers

Symbolic music generation with transformers

- **Symbolic music**
  - **Input: Tokens (text) of pianoroll**

Music score



MIDI representation (or piano roll)

```
NoteOn(50) TimeShift(9) NoteOn(60) NoteOn(65)
NoteOn(69) NoteOn(76) TimeShift(12) NoteOff(60)
NoteOff(65) NoteOff(69) NoteOff(76) TimeShift(3)
NoteOff(50) NoteOn(43) NoteOn(59) NoteOn(65)
NoteOn(69) NoteOn(76) TimeShift(24) NoteOff(All)
```

Representation as sequence of tokens

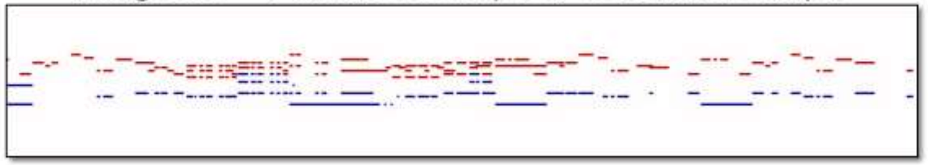
G. Richard

Exploiting knowledge for model-based deep music generation

Symbolic music generation with transformers

- **Data-driven Symbolic Music Generation is hard!**
  - ✓ Inconsistency in melody and rhythm, absence of multi-scale structures found in real music [1]
  - ✓ Practical limitations: limited dataset sizes (compared to, e.g., language, vision) feeds into limited model sizes

ImageNet - 14 million samples, 65k dims/sample



LakhMIDI - 600k samples, 65k dims/sample

- **Possible solution to do more with less: hybrid deep models**
  - ✓ Add knowledge about musical structure to data-driven models

[1] Wu & Yang, "Compose & Embellish: Well-structures piano performance generation via a two-stage approach", arXiV, 2022.

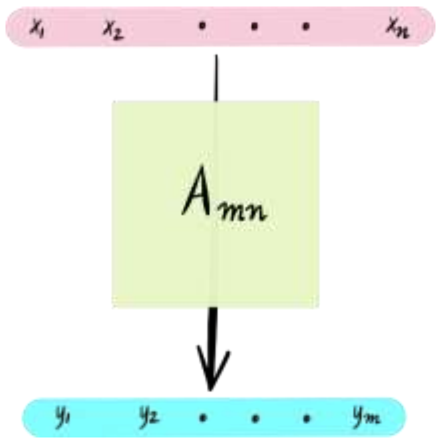
Symbolic music generation with transformers

- **Already many possibilities to exploit musical structure …**
  - Long line of research to include structure in music generation systems [1]
  - Particularly for Transformers:
    - Tokens for musical structures [2-4]
    - Positional Encoding using musical structure [5,6]
- **… But can we improve how Transformers represent and use structural information?**

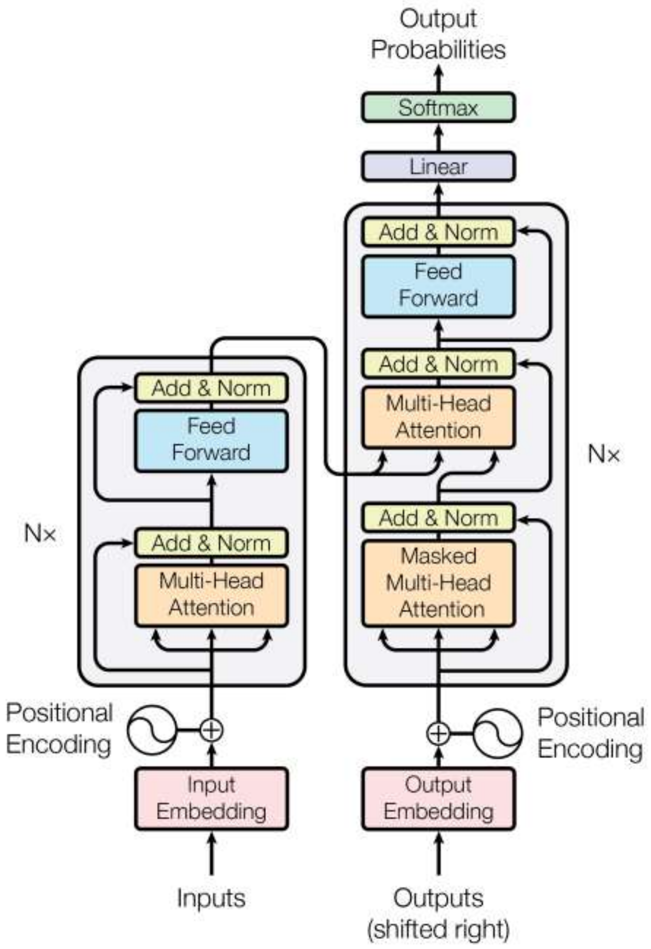
[1] Bhandari & Colton, "Motifs, Phrases, and Beyond: The Modelling of Structure in Symbolic Music Generation", EvoMUSART, 2024.
[2] Ren, et al., "PopMAG: Pop music accompaniment generation", ACM MM, 2020.
[3] Huang & Yang, "Pop Music Transformer: Beat-based modelling and generation of expressive pop piano compositions", ACM MM, 2020.
[4] Hsiao, et al., "Compound word transformer: Learning to compose full-song music over dynamic directed hypergraphs", AAAI, 2021.
[5] Liu, et al., "Symphony Generation with permutation invariant language model", arXiV, 2022.
[6] Guo, Kang & Herremans, "A domain-knowledge-inspired music embedding space and a novel attention mechanism for symbolic music modeling", AAAI, 2022.

Symbolic Music Generation

- For example with transformers :
  - Attention: Invariance to temporal order of inputs



  - **Role of the PE:** to provide the information about which element of the input sequence comes in which order.



<u>Towards exploiting « musical structured informed » Position Encoding (PE)</u>

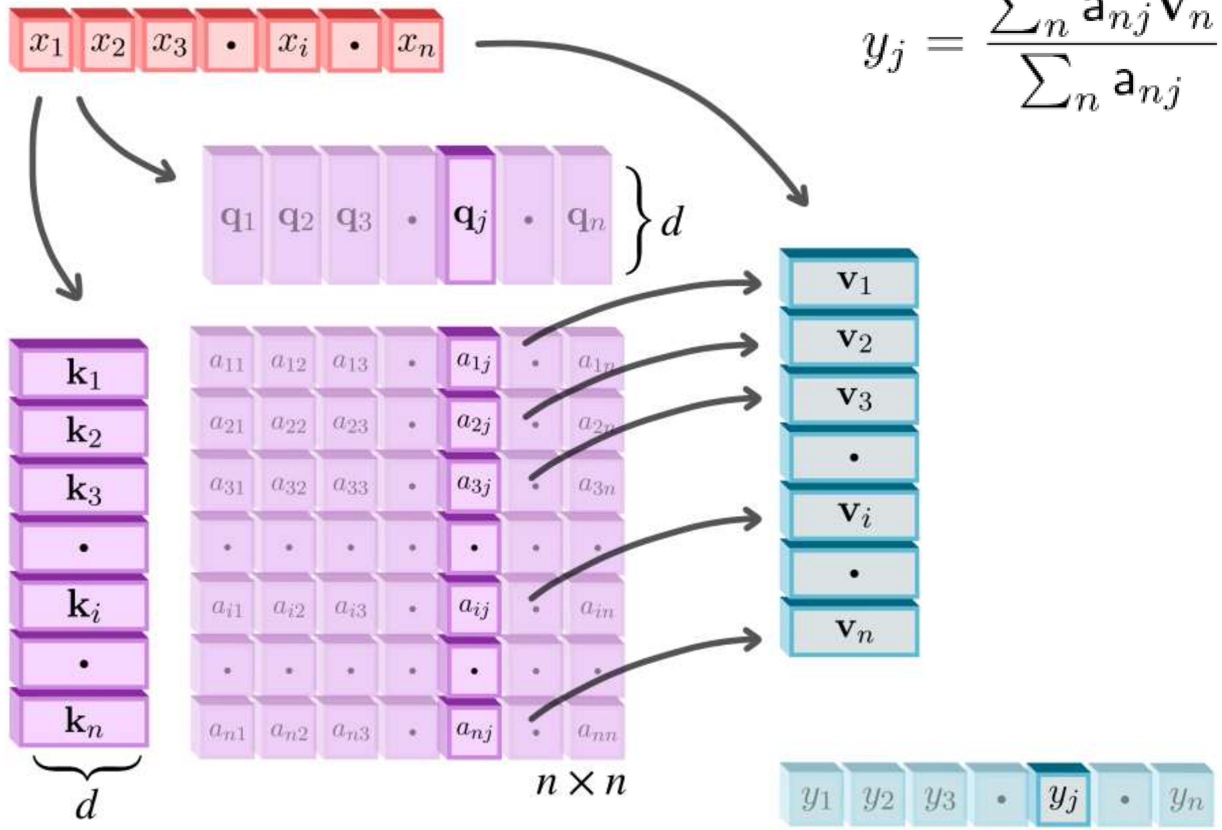
M. Agarwal, C. Wang, G. Richard. Structure-informed Positional Encoding for Music Generation. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr 2024, Seoul, South Korea.

G. Richard

Exploiting knowledge for model-based deep music generation

Symbolic Music Generation

- Attention



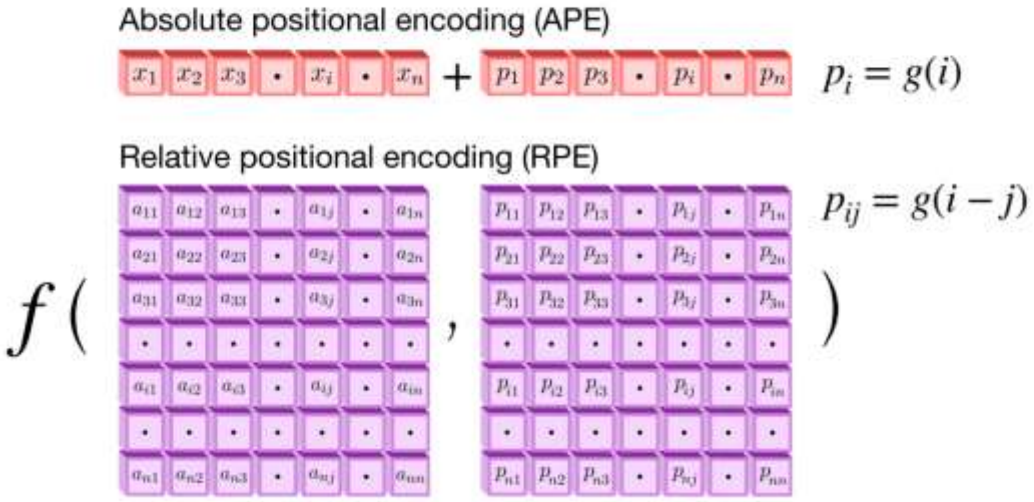
$$y_j = \frac{\sum_n \mathsf{a}_{nj} \mathbf{v}_n}{\sum_n \mathsf{a}_{nj}} \text{ with } \mathsf{a}_{nj} = \exp\left(\frac{a_{nj}}{\sqrt{D}}\right)$$

$$a_{nj} = \mathbf{q}_j \mathbf{k}_n^\top$$

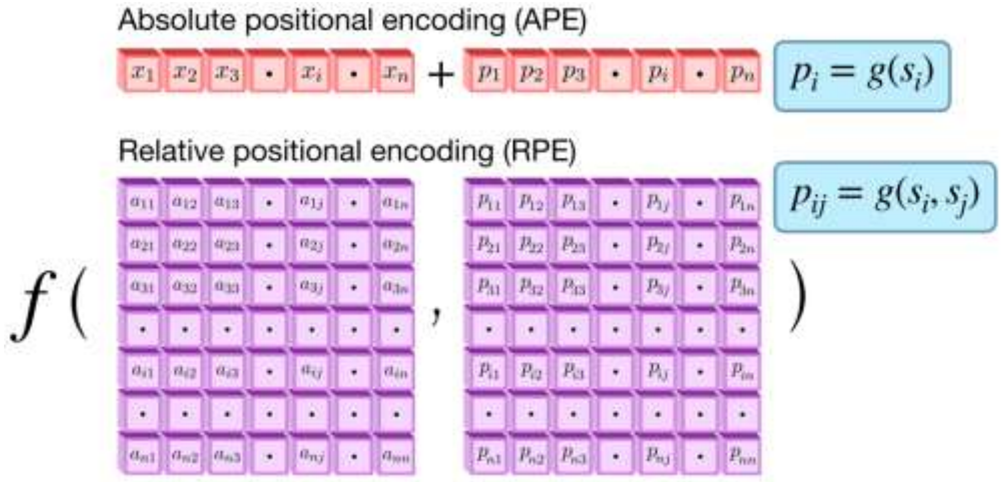
M. Agarwal, C. Wang, G. Richard. Structure-informed Positional Encoding for Music Generation. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr 2024, Seoul, South Korea.

Symbolic Music Generation

- Classic Positional Encoding



- Structure Positional Encoding

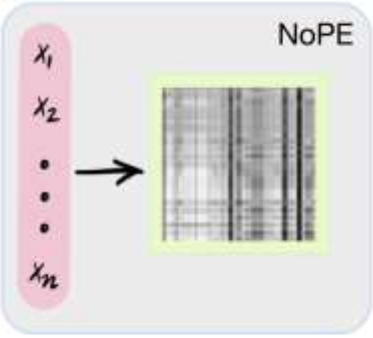


M. Agarwal, C. Wang, G. Richard. Structure-informed Positional Encoding for Music Generation. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr 2024, Seoul, South Korea.

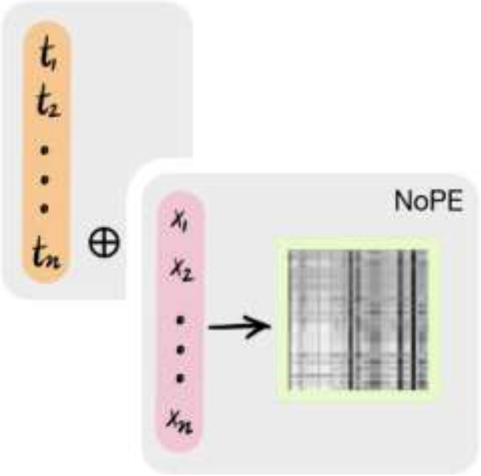
Symbolic Music Generation

« musical structure-informed » Position Encoding (PE)

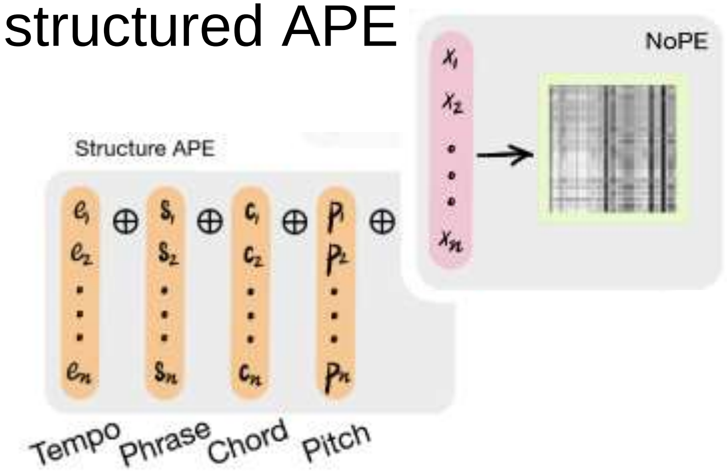
- From No Positional Encoding



…to Absolute PE



…to structured APE



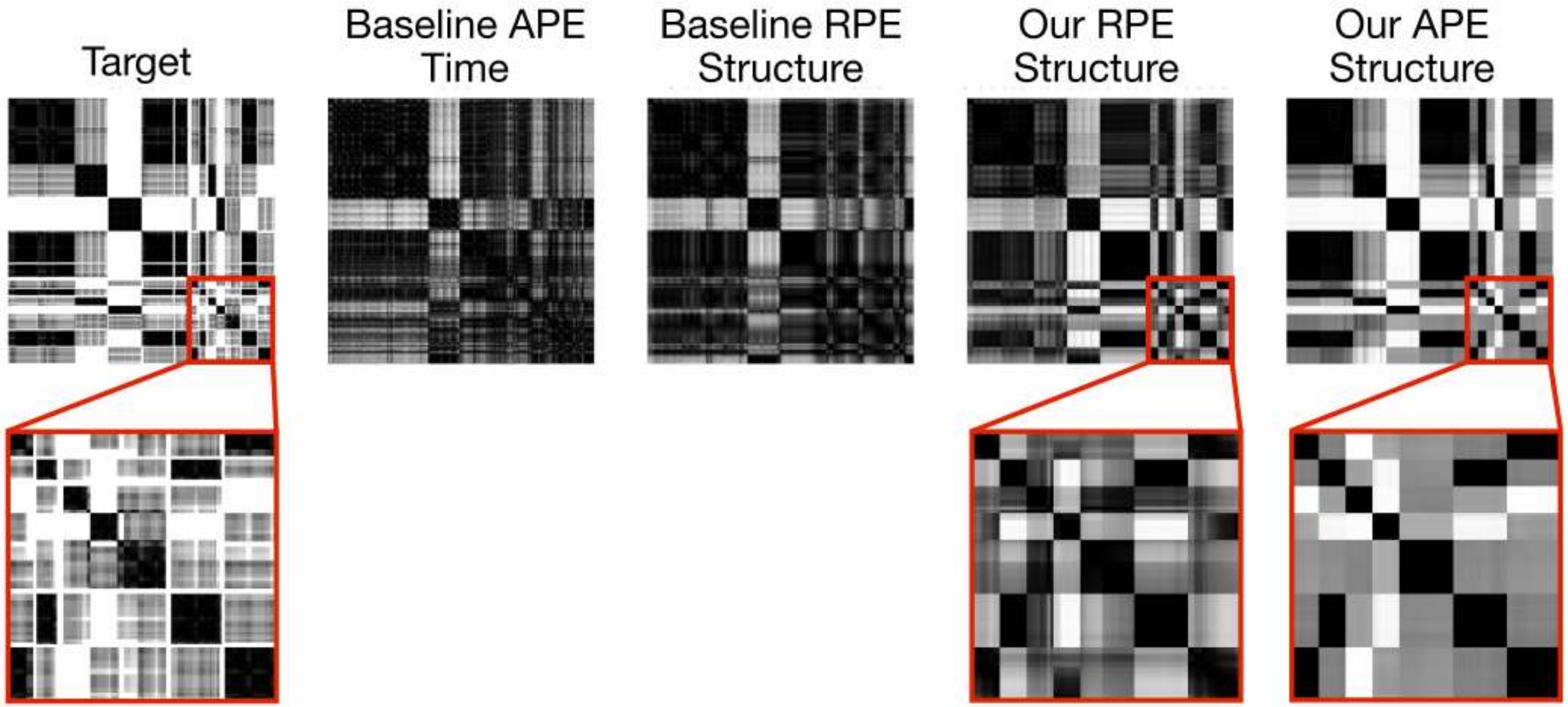
Results show that better music generation can be achieved by using knowledge about musical structure in data-driven Transformers through Positional Encoding

M. Agarwal, C. Wang, G. Richard. Structure-informed Positional Encoding for Music Generation. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr 2024, Seoul, South Korea.

Accompaniment generation from melody tracks

Illustration

- Our structure-informed positional encoding captures large-scale and small-scale structures :
  - ✓ Self-similarity matrices of chroma profiles *(chroma is a feature representation capturing chords information)*



M. Agarwal, C. Wang, G. Richard. Structure-informed Positional Encoding for Music Generation. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr 2024, Seoul, South Korea.

G. Richard

Exploiting knowledge for model-based deep music generation

Extension for linear complexity structure-informed PE

- **Exploiting a kernelized form of attention [1,2]**

$$a_{mn} = \mathcal{K}(\mathbf{q}_m, \mathbf{k}_n) = \mathbb{E}\left[\phi(\mathbf{q}_m)\phi(\mathbf{k}_n)^\top\right]$$

  - With multiple instantiations, ϕ captures, on average, the relationship between $\mathbf{q}_m$ and $\mathbf{k}_n$

    ⇨ leads to linear-complexity Transformers.

    - Applicable for Absolute Position Encoding

- **Stochastic Position Encoding [3]** => Applicable to Relative PE with linear complexity
  - <u>Key ideas:</u>
    - Express the Attention matrix with position kernels $\mathbf{A} = \exp\left(\left[\sum_{d=1}^{D} q_{md}\mathcal{P}_d(m,n)k_{nd}\right]_{mn} \Big/ \sqrt{D}\right)$
    - Express the position kernel as a covariance matrix $(\forall \mathcal{M}, \mathcal{N})\ (\forall m, n)\ \mathcal{P}_d(m,n) = \mathbb{E}\left[\overline{Q}_d(m)\overline{K}_d(n)\right]$

- **Extension to structure-informed stochastic Position Encoding** [4]

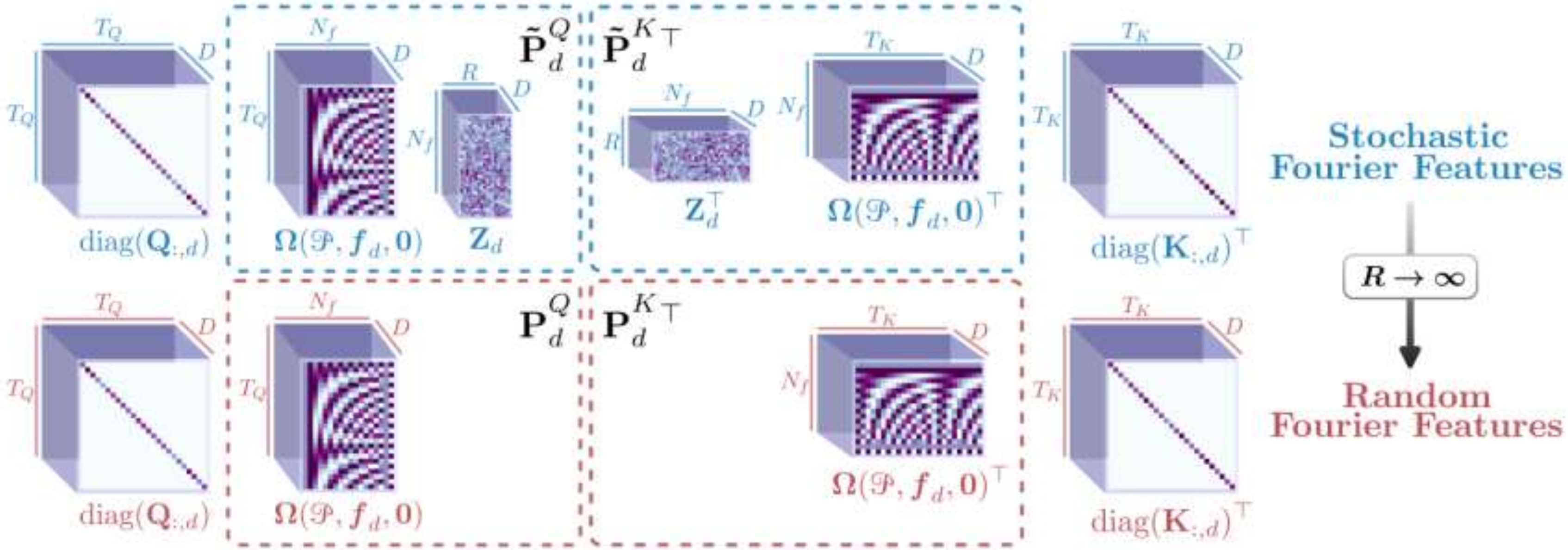
[1] Y.-H. H. Tsai & al. Transformer Dissection: An Unified Understanding for Transformer's Attention via the Lens of Kernel," EMNLP, 2019
[2] K. M. Choromanski, & al. Rethinking Attention with Performers," ICML,2021
[3] A. Liutkus & al. Relative Positional Encoding for Transformers with Linear Complexity," ICML, 2021
[4] M. Agarwal & al, F-StrIPE: Fast Structure-Informed Positional Encoding for Symbolic Music Generation, ICASSP 2025.

Extension for linear complexity structure-informed PE

- **F-StrIPE: Structure informed stochastic Position Encoding** [4]

The positional matrix $\mathbf{P}_d$ captures the relationship between pairs (m, n) of timesteps from the positional index sequences $\mathcal{P}_Q = \{1, ..., m, ..., T_Q\}$ and $\mathcal{P}_K = \{1, ..., n, ..., T_K\}$.

F-StrIPE: exploiting structure-aware positional indices $p_i = \mathbf{s}(i)$ instead of classic time indices $p_i = i$



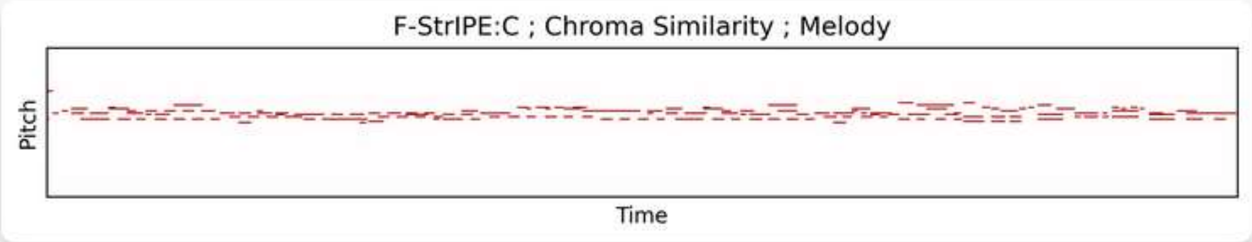
[4] M. Agarwal & al, F-StrIPE: Fast Structure-Informed Positional Encoding for Symbolic Music Generation, ICASSP 2025.

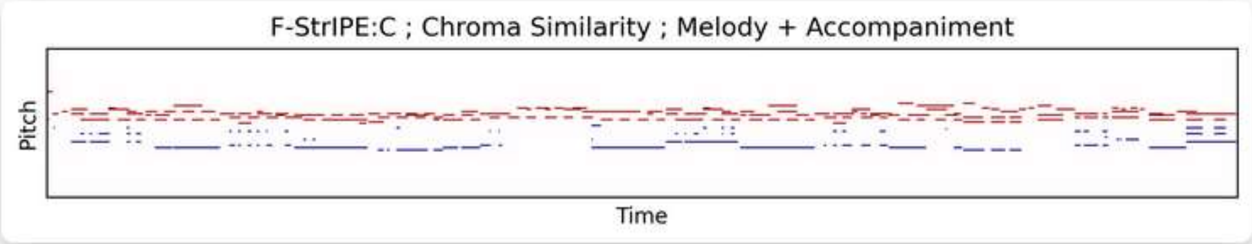
G. Richard

Exploiting knowledge for model-based deep music generation

F-StrIPE: Structure informed stochastic Position Encoding [4]

- Demo page at : [*bit.ly/faststructurepe*](bit.ly/faststructurepe)
  - Best example for « Chroma similarity » metric *(training 16 bars of melody – generation 16 bars of accompaniement)*





[4] M. Agarwal & al, F-StrIPE: Fast Structure-Informed Positional Encoding for Symbolic Music Generation, ICASSP 2025.

Musical Timbre transfer

Timbre transfer : a specific application of style transfer to music

- Image style transfer

Content

Generated image

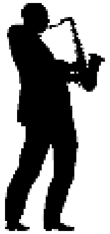
Style

- Timbre transfer in music

Content

Trumpet

Timbre



Saxophone

Generated audio:
Saxophone playing the original trumpet content

WavTransfer: A Flexible End-to-end Multi-instrument Timbre Transfer with Diffusion

- **Timbre Transfer:**
  - Essential for distinguishing sounds with the same pitch and loudness
  - Modifies the tonal quality while preserving pitch and structure

- *Common models :* Need for separate models for each pair of instrument for timbre transfer

- **WaveTransfer[1]:**
  - Works for audio mixtures and individual instruments
  - Generates audio waveforms directly
  - Operates at multiple sampling frequencies 16 kHz and 44.1 kHz

[1] T. Baoueb, X. Bie, G. Richard, WaveTransfer: A Flexible End-to-end Multi-instrument Timbre Transfer with Diffusion, ICASSP 2025

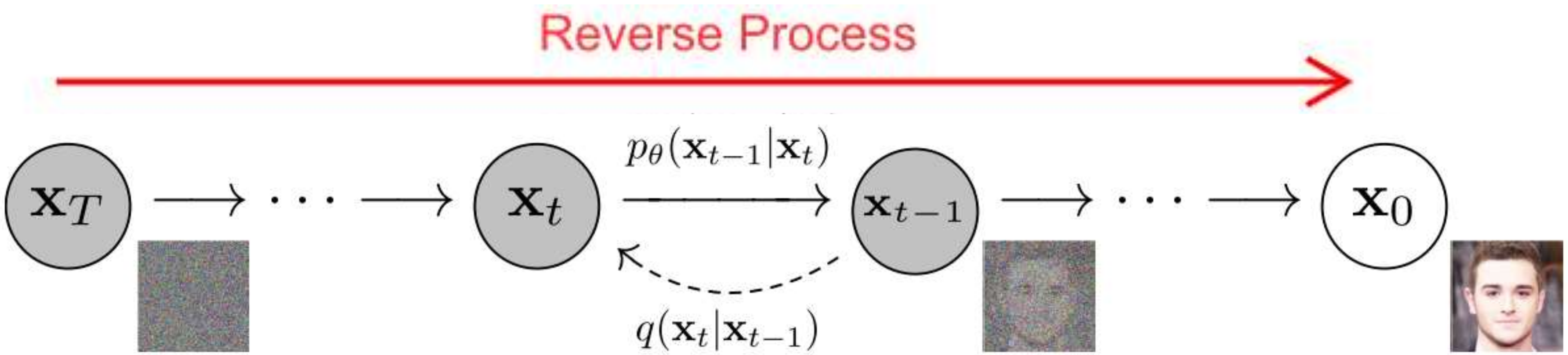
G. Richard

Exploiting knowledge for model-based deep music generation

Background

Denoising diffusion probabilistic models (DDPMs)

- Characterising a data distribution by gradually introducing noise into samples for *T* steps and then learning the process of reversing it

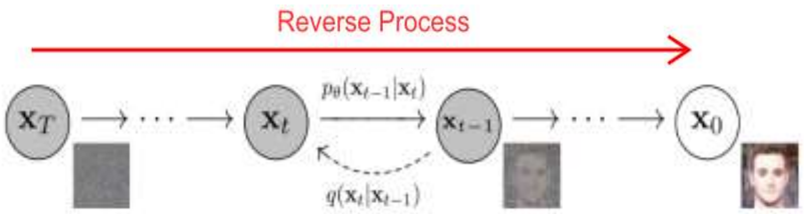


$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$$

J. Ho & al. Denoising diffusion probabilistic models, in NeurIPS, 2020

Background

Denoising diffusion probabilistic models (DDPMs)



$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$$

- $\mathbf{x}_0 \sim q(\mathbf{x}_0)$: an initial sample, $\{\beta_t\}_{t=1}^{T}$: a noise schedule
- Let $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{u=1}^{t} \alpha_u$. $\mathbf{x}_t$ can be sampled at any arbitrary time step t:

Forward Process:
$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, \ \varepsilon \sim \mathcal{N}(\varepsilon; \mathbf{0}, \mathbf{I}) \tag{1}$$

- The training loss is given by:

$$\mathscr{L}_\theta = \min_\theta \mathbb{E}\left[\|\varepsilon_\theta(\mathbf{x}_t, t) - \varepsilon\|_2^2\right], \tag{2}$$

- During inference, we can iteratively sample the data from $\mathbf{x}_T \sim \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$ to $\mathbf{x}_0$ via:

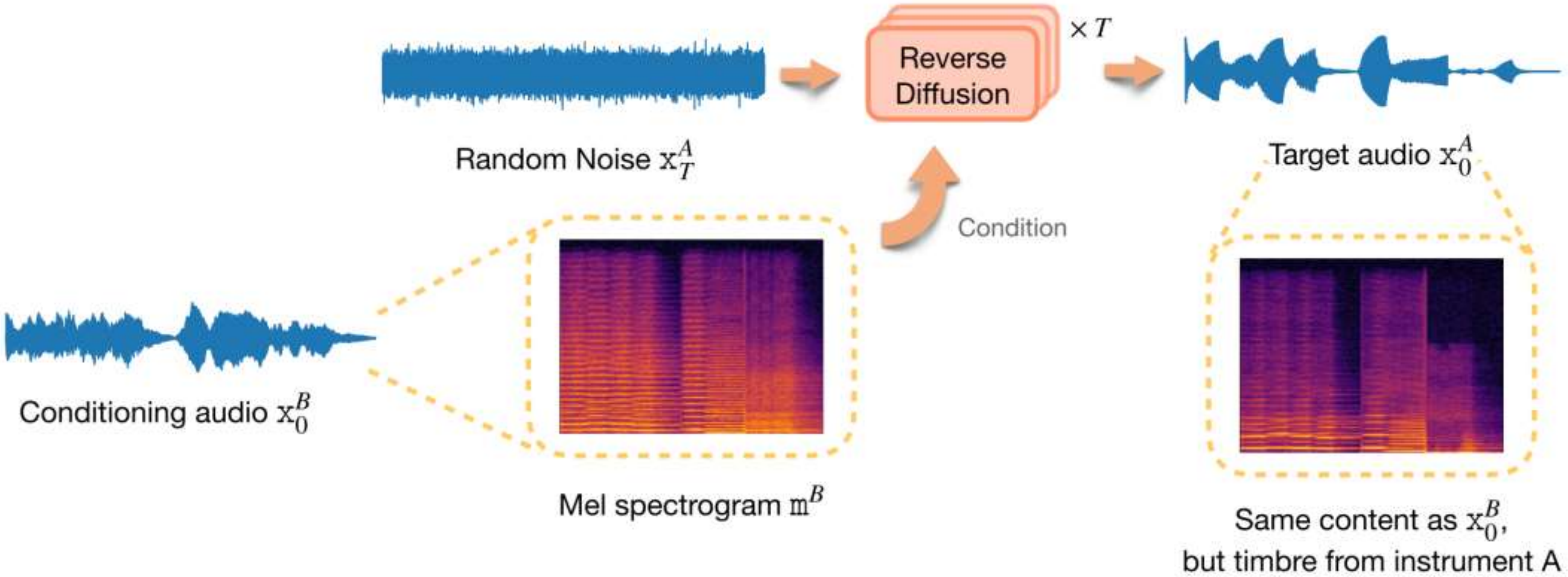
$$\mathbf{x}_{t-1} = \mathcal{N}(\mathbf{x}_{t-1}; \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta(\mathbf{x}_t, t)), \sigma_t^2 I), \tag{3}$$

where σ_t is a time dependent constant.

J. Ho & al. Denoising diffusion probabilistic models, in NeurIPS, 2020

Timbre transfer : principle of WaveTransfer [1]

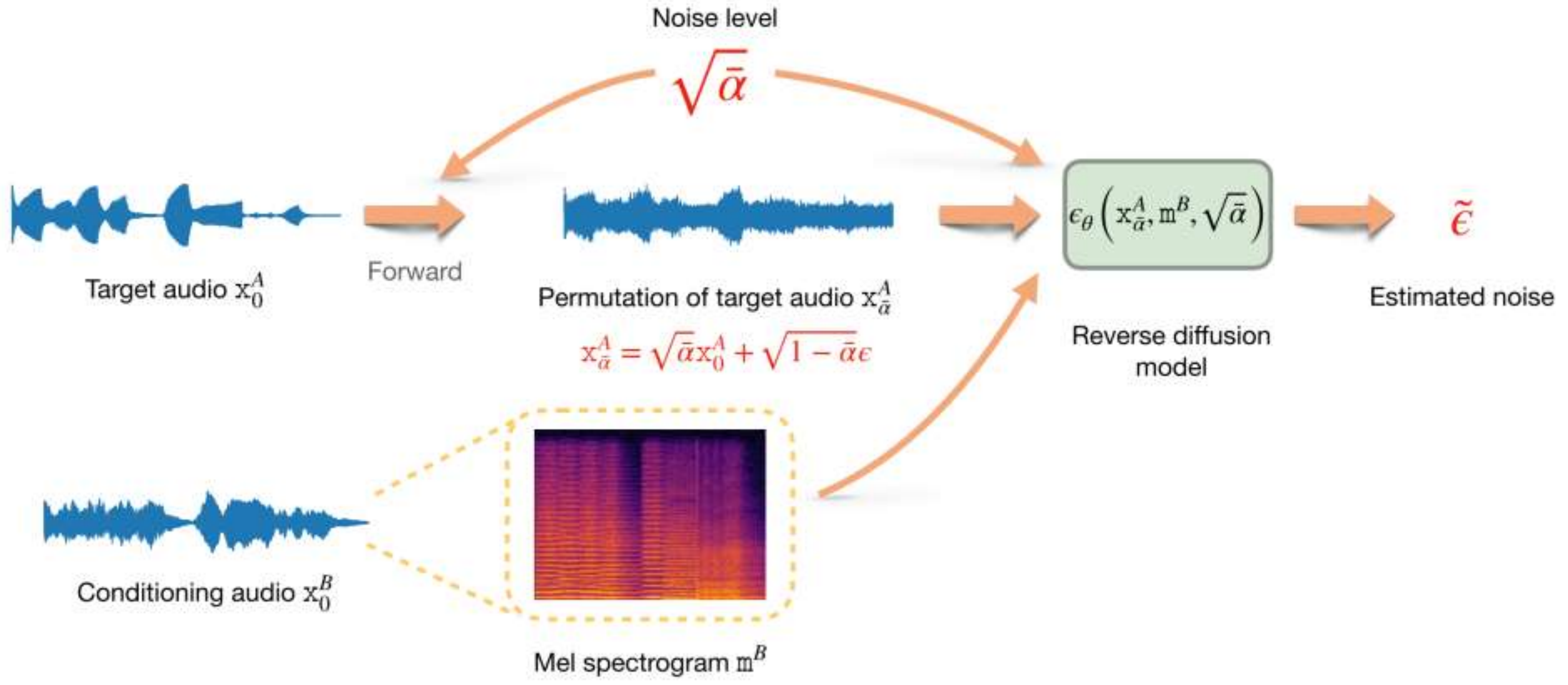
- Extending Wavgrad [2] for timbre transfer.
- Timbre transfer objective: generate a target audio x_0^A from a random noise x_T^A and conditioning audio x_0^B



[1] T. Baoueb, X. Bie, G. Richard, WaveTransfer: A Flexible End-to-end Multi-instrument Timbre Transfer with Diffusion, ICASSP 2025
[2] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan, "Wavegrad: Estimating gradients for waveform generation," in Proc. ICLR, 2021

WaveTransfer: Training process

- Supervised Training: Aligned dataset
- $\mathbf{x}_0^A$ and $\mathbf{x}_0^B$: same content, $\neq$ instruments



$$\mathcal{L}_{\boldsymbol{\theta}} = \min_{\boldsymbol{\theta}} \mathbb{E}\left[\left\|\boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}}\mathbf{x}_0^A + \sqrt{1-\bar{\alpha}}\boldsymbol{\epsilon}, \mathbf{m}^B, \sqrt{\bar{\alpha}}\right) - \boldsymbol{\epsilon}\right\|_1\right]$$

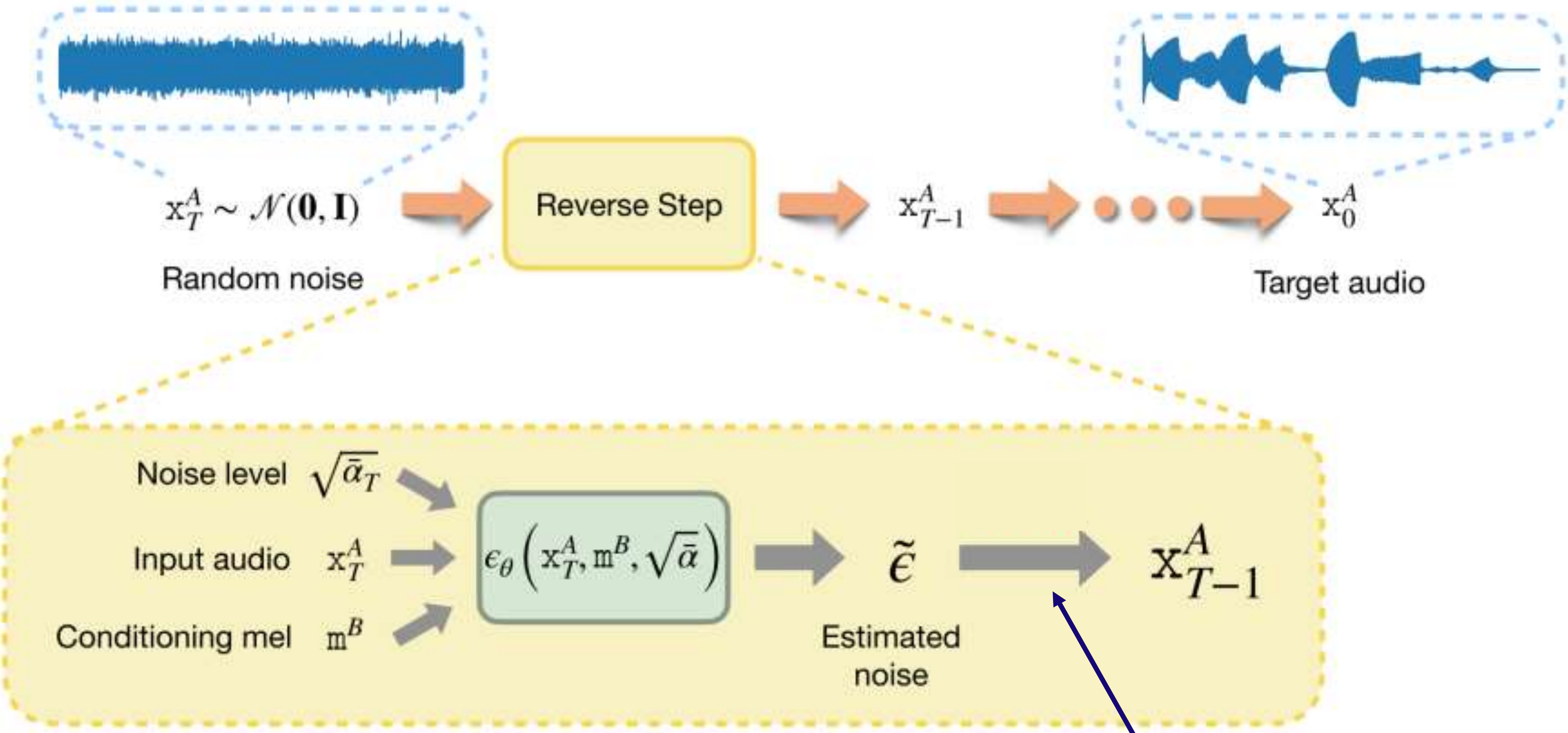
G. Richard

Exploiting knowledge for model-based deep music generation

G. Richard

Exploiting knowledge for model-based deep music generation

WaveTransfer: Inference



$$\mathbf{x}_{n-1} = \frac{1}{\sqrt{\alpha_n}}\left(\mathbf{x}_n - \frac{1-\alpha_n}{\sqrt{1-\bar{\alpha}_n}}\boldsymbol{\epsilon}_\theta\left(\mathbf{x}_n, \mathbf{m}^B, \sqrt{\bar{\alpha}_n}\right)\right) + \sigma_n \boldsymbol{z},$$

T. Baoueb, X. Bie, G. Richard, WaveTransfer: A Flexible End-to-end Multi-instrument Timbre Transfer with Diffusion, ICASSP 2025

G. Richard

Exploiting knowledge for model-based deep music generation

Wavetransfer: Timbre transfer demo

https://wavetransfer.github.io/

- Timbre transfer : piano to vibraphone (16 kHz)

Name	Input (ground truth)	Target (ground truth)	Music-STAR	DiffTransfer	WT^{16}_{global} with WG-6	WT^{16}_{global} with BDDM-20
Pirates of Caribbean						

- Mixture of Timbre transfer : piano+strings -> vibraphone + clarinet

Name	Input (ground truth)	Target (ground truth)	Music-STAR	DiffTransfer	WT^{16}_{global} with WG-6	WT^{16}_{global} with BDDM-20	WT^{16}_{mix} with WG-6	WT^{16}_{mix} with BDDM-20
Beethoven								

Only trained on the specific mixtures

T. Baoueb, X. Bie, G. Richard, WaveTransfer: A Flexible End-to-end Multi-instrument Timbre Transfer with Diffusion, ICASSP 2025

Wavetransfer

- **Capabilities of the model**
  - Handles timbre transfer for both audio mixtures and individual instruments in one model
  - Eliminates the requirement for separate model training for each timbre transfer
- **Current Limitations**
  - Relies on an aligned dataset
  - Limited instrument diversity in timbre transfer

T. Baoueb, X. Bie, G. Richard, WaveTransfer: A Flexible End-to-end Multi-instrument Timbre Transfer with Diffusion, ICASSP 2025

To conclude

- The potential for hybrid deep learning …
  - **Interpretability, Controllability, Explainability**
    - Hybrid model becomes controllable by human-understandable parameters
    - New audio capabilities: perceptually meaningful sound transformation
  - **Frugality: gain of several orders of magnitude** in the need of data and model complexity
  - **Towards a more resource efficient and sustainable AI**
  - **Applicable 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⟩
…