

Explainable and interpretable deep audio processing based on hybrid deep learning

Gaël RICHARD*

Professor, Telecom Paris, Institut polytechnique de Paris

*work with collaborators and in particular

K. Schulze-Forster, C. Doire, L. Kelley, B. Torres, P. Chouteau, R. Badeau, G. Peeters, M. Agarwal, C. Wang, T. Baoueb, J. Leroux, M. Fontaine, H. Liu

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 deep learning**
 - Some examples in other domains
 - Hybrid deep learning in audio
 - Several application examples:
 - *Audio synthesis*
 - *Unsupervised music source separation*
- Discussion and conclusion

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 my ERC project : 

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

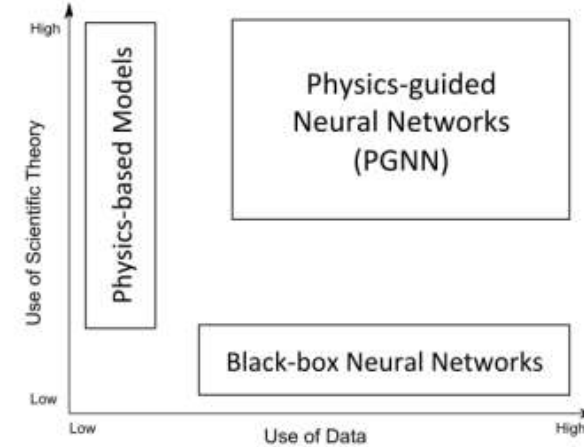
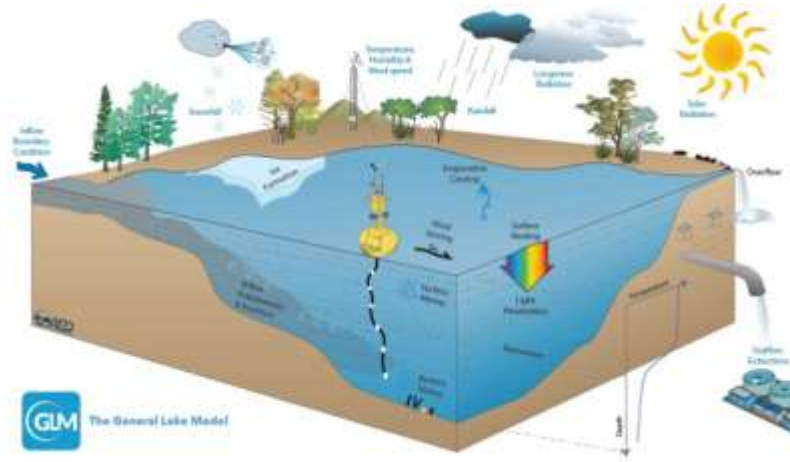
My approach: to build *Hybrid deep learning models*, by **integrating our prior knowledge** about the nature of the processed data.



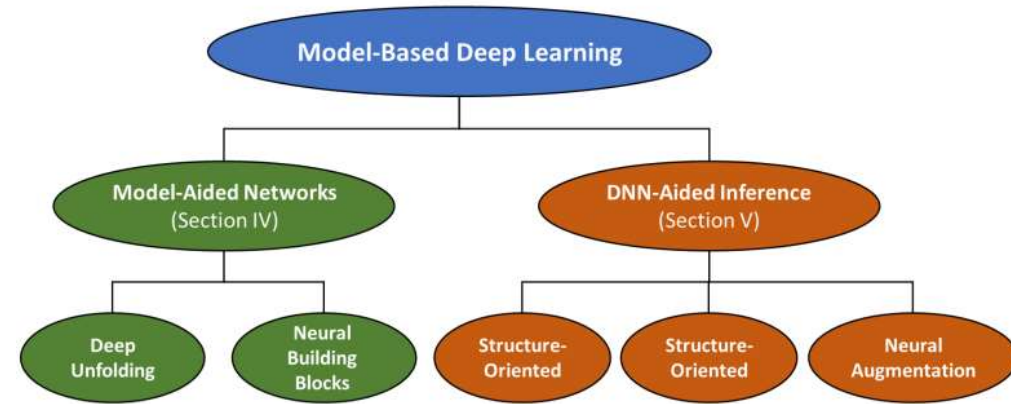
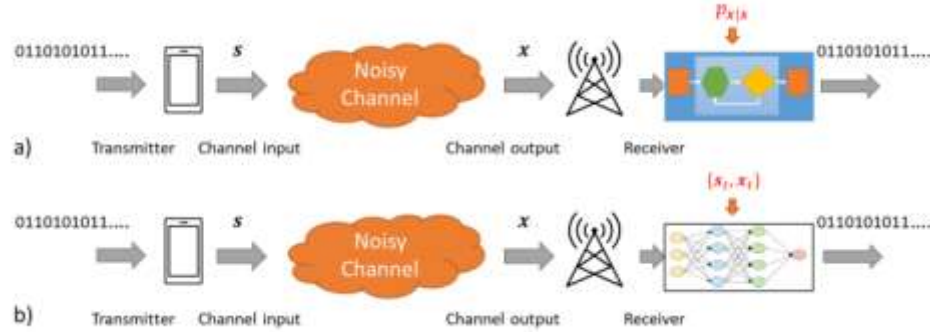
Towards Hybrid deep learning

... some prior works.

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



- Digital communication and Image restoration [2,3]

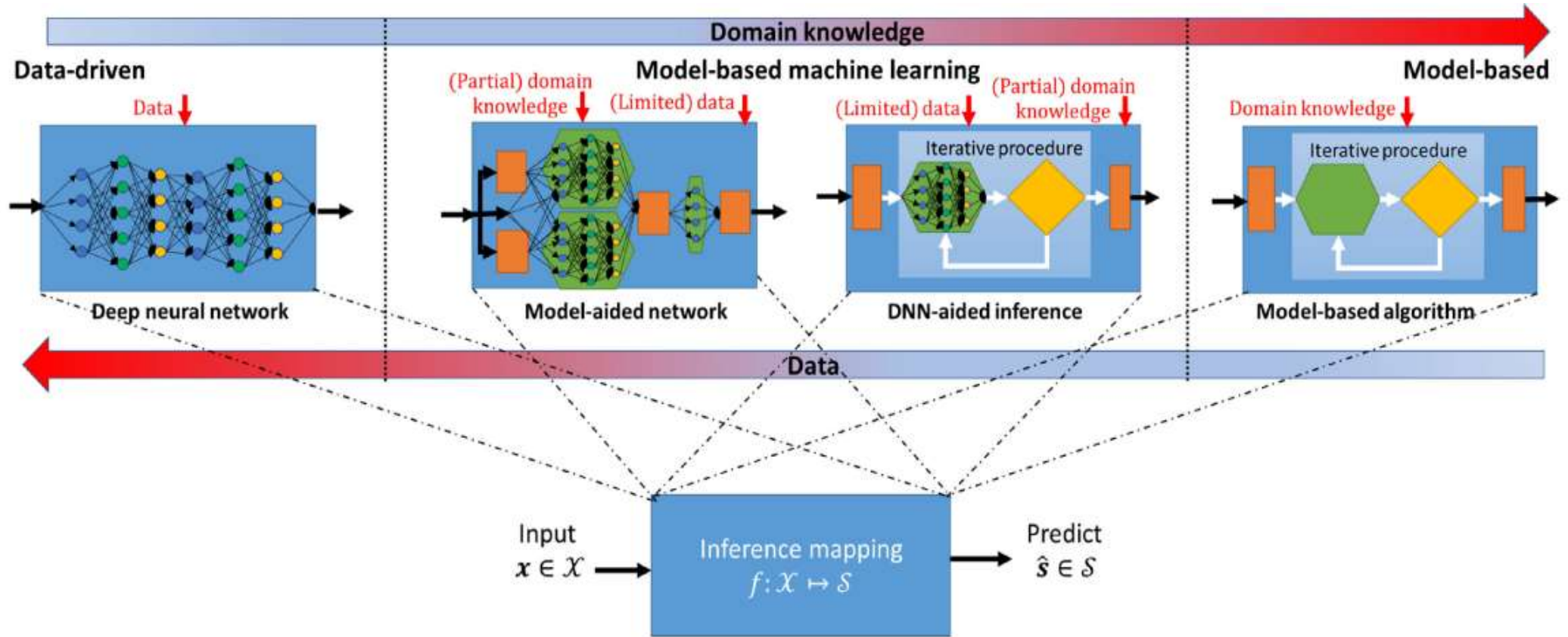


[1] A. Karpatne & al. "Physics-guided Neural Networks (PGNN): An Application in Lake Temperature Modeling," arXiv, 1710.11431, 2017.
 [2] B. Lecouat & al., "Fully Trainable and Interpretable Non-Local Sparse Models for Image Restoration.," 2020. (hal-02414291v2).
 [3] N. Shlezinger, & al., "Model-Based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

Towards Hybrid deep learning

... some prior works.

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



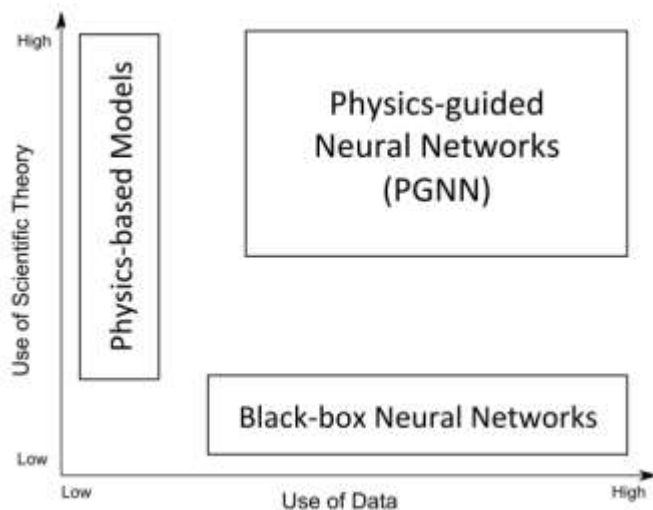
Content

- Context and motivation
- Towards hybrid deep learning
 - Some examples in other domains
 - **Hybrid deep learning in audio**
 - Several application examples:
 - *Audio synthesis*
 - *Unsupervised music source separation*
 - *Discrete neural representation*
- Discussion and conclusion

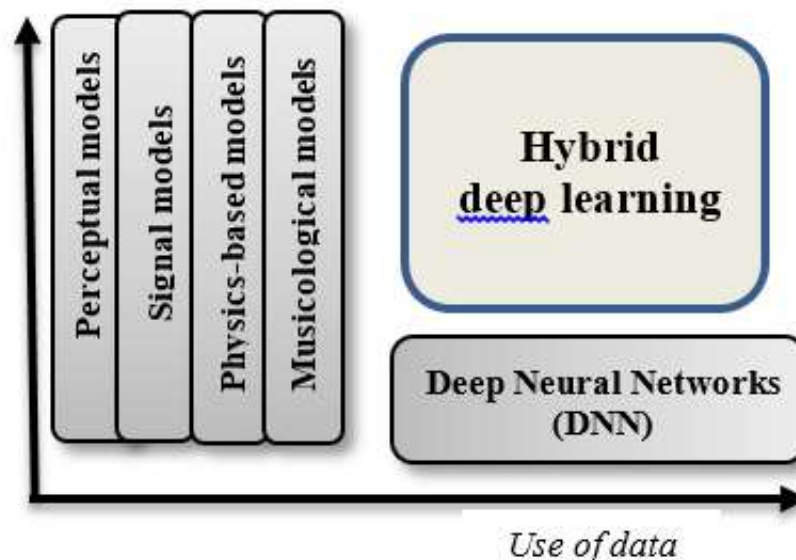
Towards Hybrid deep learning

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

- Towards novel models for hybrid deep learning combining parameter-efficient and interpretable audio models with deep neural architectures



Use of parameter-efficient and interpretable models

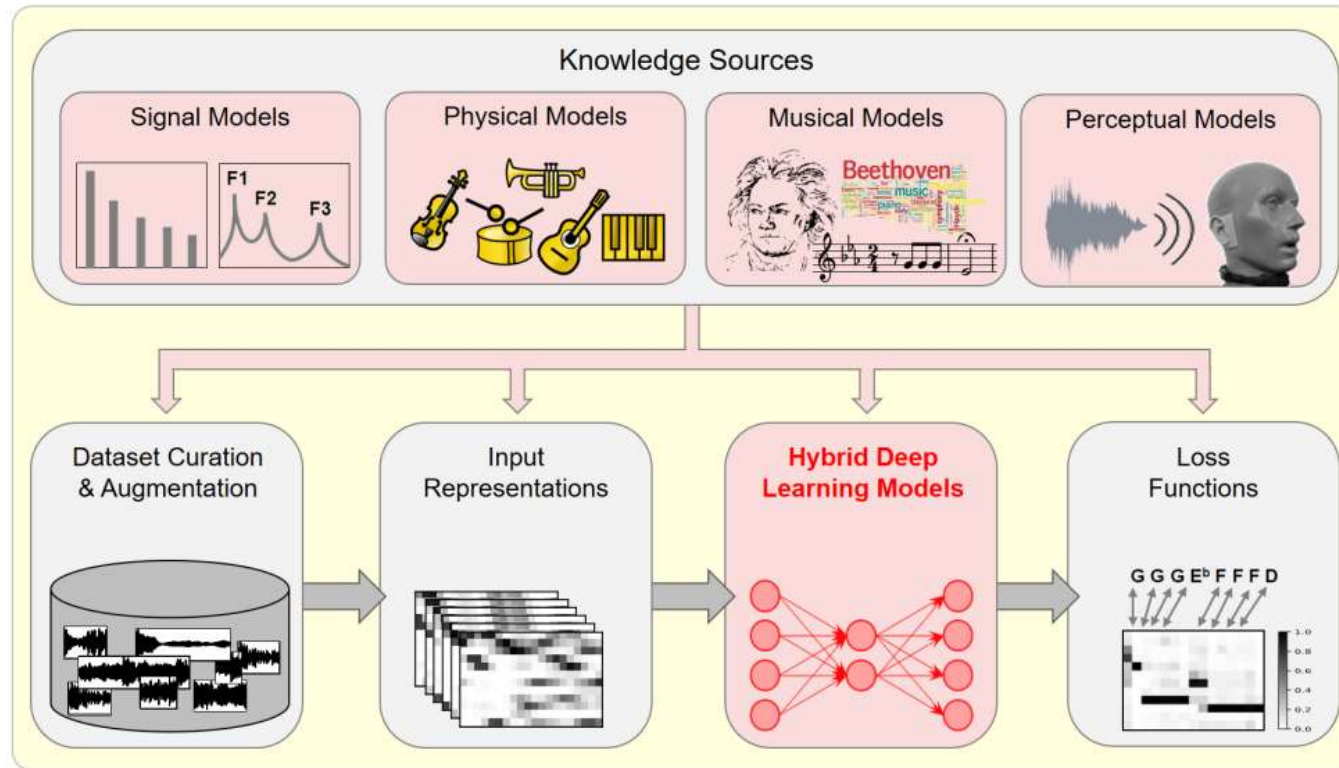


Towards Hybrid deep learning approaches

Examples with Hybrid deep model for Music signals

- Coupling model-based and deep learning:

Example with Hybrid deep model for Music signals



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, 2024, to appear.*

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

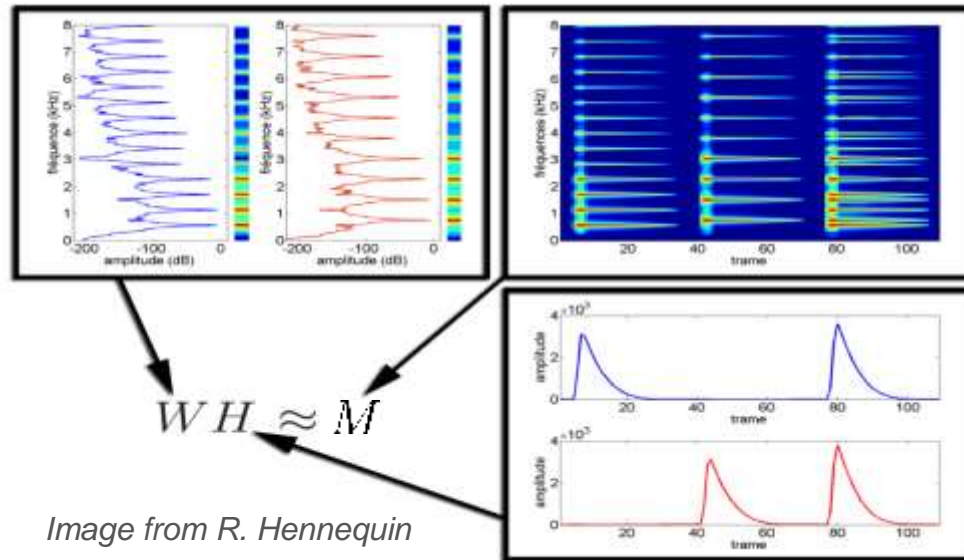


Towards Hybrid deep learning

... some prior works in Audio.

- **Signal models can be used as an advanced representation:**
 - An example: non-negative factorization models with CNNs for audio scene classification

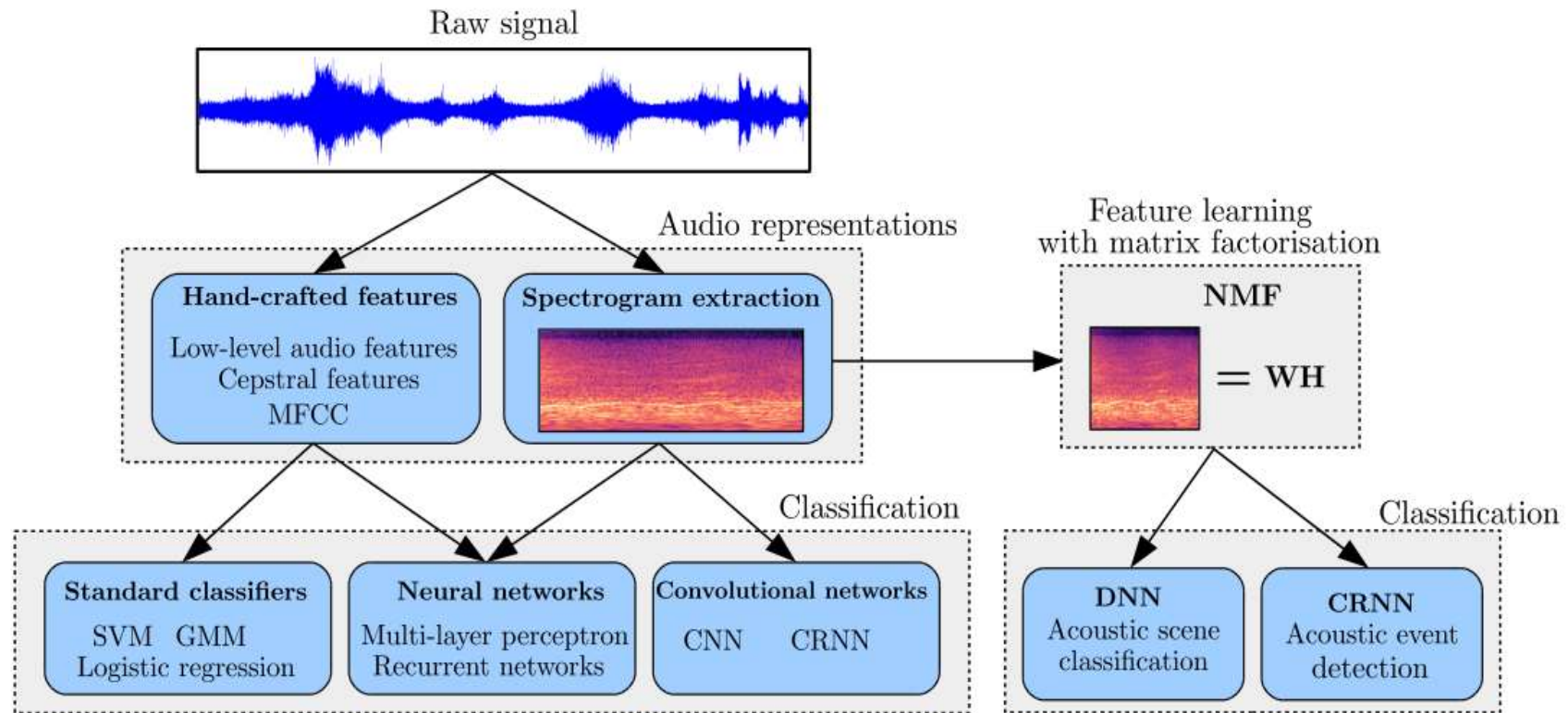
Principle of Non-Negative Matrix Factorization on Audio spectrograms



Towards Hybrid deep learning

... some prior works in Audio.

- Feature learning with NMF for audio scene classification



V. Bisot & al., "Feature Learning with Matrix Factorization Applied to Acoustic Scene Classification", ACM/IEEE Trans. on ASLP, vol. 25, no. 6, 2017

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,

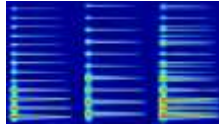


Towards Hybrid deep learning

... some prior works in Audio.

- Deep NMF : the concept of deep unrolling

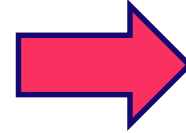
- Classic NMF



$$\mathbf{M} \approx \mathbf{W}\mathbf{H} = \hat{\mathbf{M}}$$

- Minimizing a distance

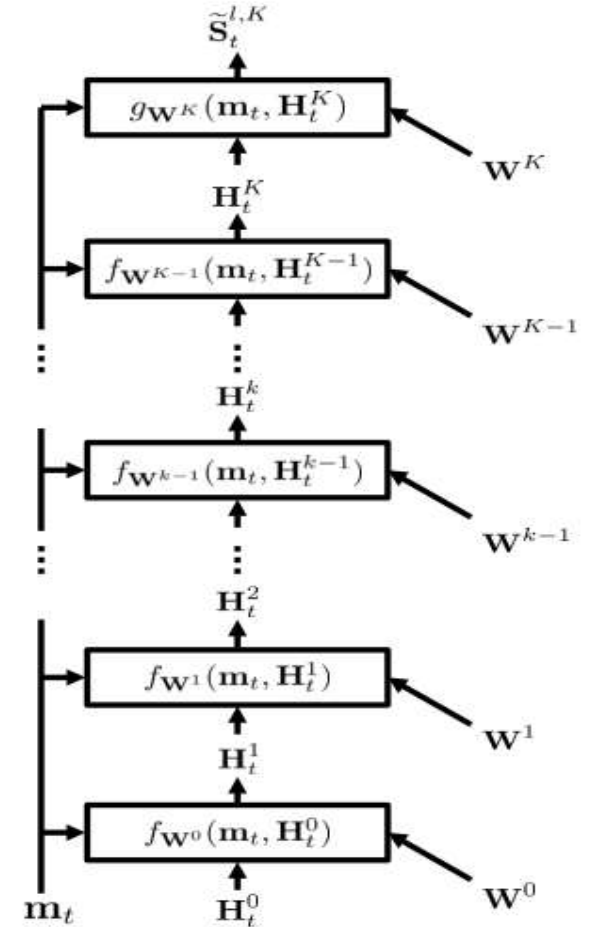
$$D(\mathbf{M}, \hat{\mathbf{M}}) = \sum_{f=1}^F \sum_{n=1}^N d(v_{fn} | \hat{v}_{fn})$$



- ..towards iterative update rules

$$\mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\mathbf{W}^T \mathbf{M}}{\mathbf{W}^T (\mathbf{W}\mathbf{H})}$$

$$\mathbf{W} \leftarrow \mathbf{W} \otimes \frac{\mathbf{M}\mathbf{H}^T}{(\mathbf{W}\mathbf{H})\mathbf{H}^T}$$



Towards Hybrid deep learning

... some prior works in Audio

- Phase retrieval from the magnitude spectrogram

$$\text{Find } \mathbf{X} \text{ s.t. } |X[\omega, \tau]| = A[\omega, \tau]$$

- The classic Griffin-Lim Algorithm (GLA)**

- Exploits spectrogram consistency
(\mathbf{X} should correspond to the complex spectrogram of a time domain signal x)

$$\text{Find } \mathbf{X} \text{ s.t. } \begin{cases} |X[\omega, \tau]| = A[\omega, \tau] \\ \mathbf{X} \in \text{Im}(\mathcal{G}) \end{cases}$$

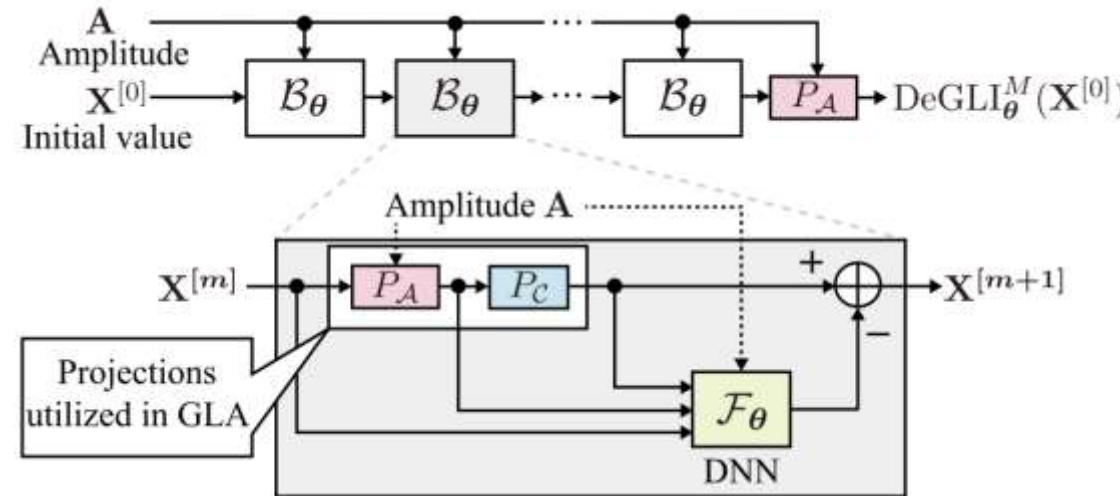
- Implemented as an iterative algorithm

$$\mathbf{X}^{[m+1]} = P_C(P_A(\mathbf{X}^{[m]}))$$

$$P_A(\mathbf{X})[\omega, \tau] = A[\omega, \tau] \frac{X[\omega, \tau]}{|X[\omega, \tau]|} \quad P_C(\mathbf{X}) = \mathcal{G}(\mathcal{G}^\dagger(\mathbf{X}))$$

- $\mathcal{G}, \mathcal{G}^\dagger$ are respectively the STFT and ISTFT operators

Deep griffin-Lim



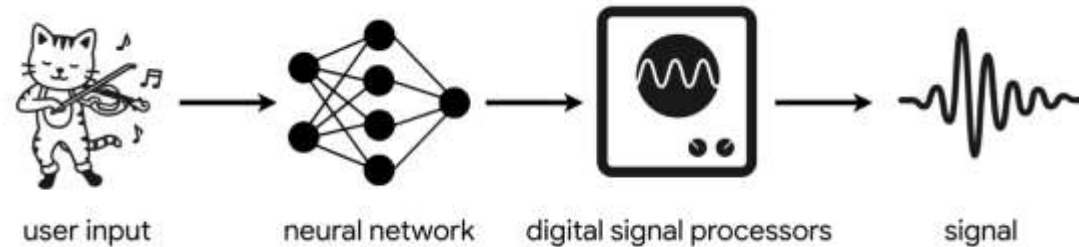
Content

- Context and motivation
- Towards hybrid deep learning
 - Some examples in other domains
 - Hybrid deep learning in audio
 - **Several application examples:**
 - ***Audio synthesis***
 - *Unsupervised music source separation*
- Discussion and conclusion

Towards Hybrid deep learning approaches

Audio synthesis

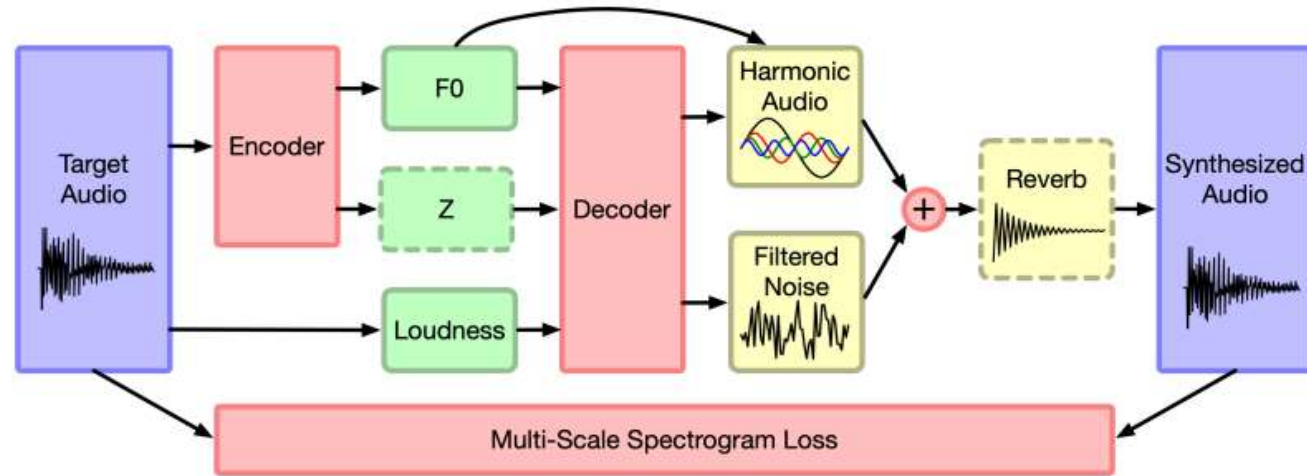
- Coupling model-based and deep learning
- For example, using deep learning for learning the parameters of a signal processing model



Towards Hybrid deep learning approaches

Audio synthesis

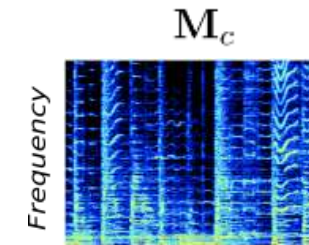
- The example of DDSP



- A multi-scale spectral loss $\mathcal{L}_{rec} = \sum_c \mathcal{L}_c$

With $\mathcal{L}_c = \|\mathbf{M}_c - \tilde{\mathbf{M}}_c\|_1 + \|\log(\mathbf{M}_c) - \log(\tilde{\mathbf{M}}_c)\|_1$

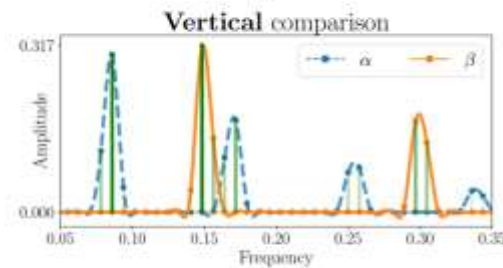
and with $c = [2048, 1024, 512, 256, 128, 64]$ indicates the FFT size used to compute the STFT.



Exploiting “physical” principles in the loss in the context of DDSP

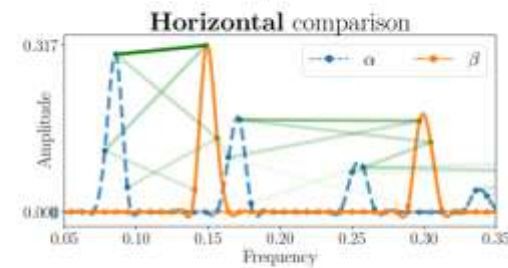
- **The issue:** traditional reconstruction losses do not work very well when trying to estimate frequency parameters (e.g. fundamental frequency)
- **Contribution:** loss function that compares audio measuring frequency displacement of spectral frames

Usual reconstruction losses
(e.g. Multi-Scale Spectral loss)



Bad gradient orientation w.r.t. frequency parameters, many local minima

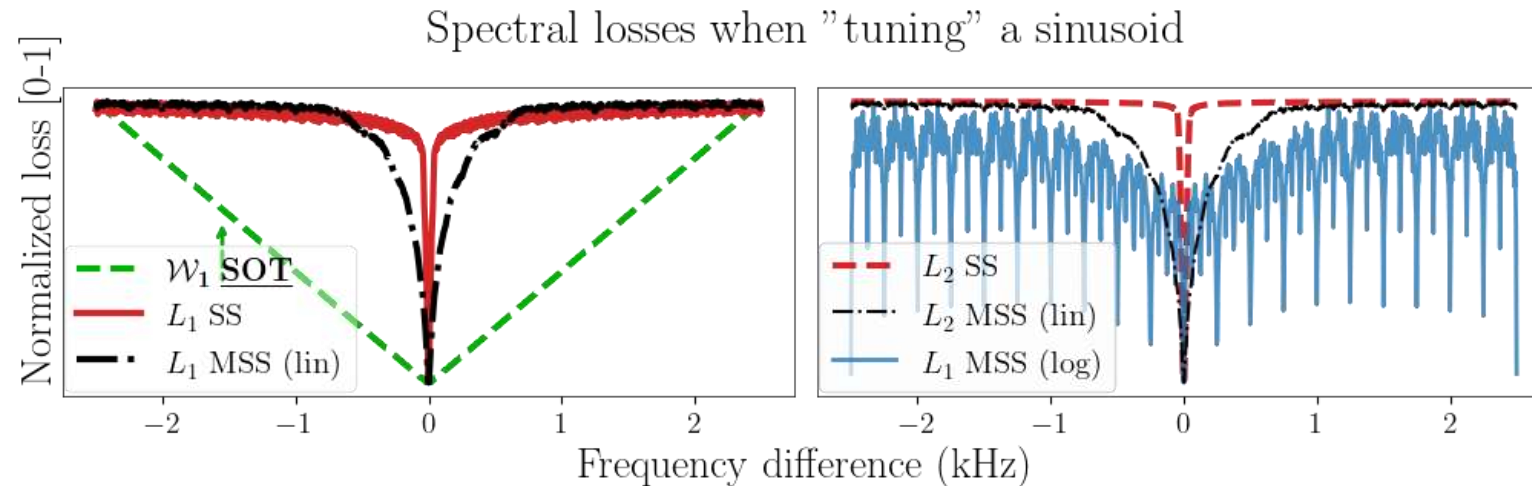
This work
Spectral Optimal Transport (SOT)



Better gradients for oscillator frequency estimation

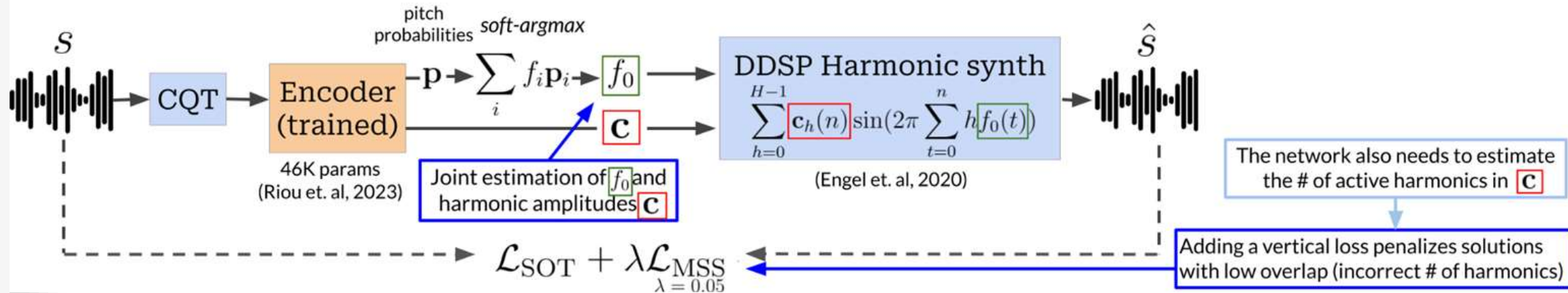
Exploiting “physical” principles in the loss in the context of DDSP

- **Motivating example:** different reconstruction losses when tuning a sinusoid
- Vertical (L_1 , L_2) converge smoothly **only** when close to global min.
- **Horizontal SOT has good gradient orientation**



Exploiting “physical” principles in the loss in the context of DDSP

- Example use case: f_0 and harmonic amplitudes estimation task

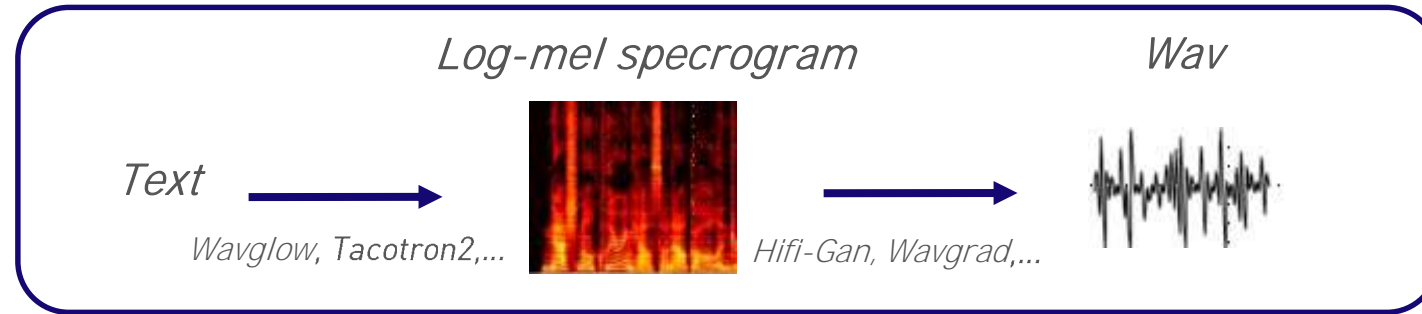


- Tested on synthetic data and compared with Multi-Scale Spectral loss
- **Improvement on pitch estimation and audio reconstruction metrics**

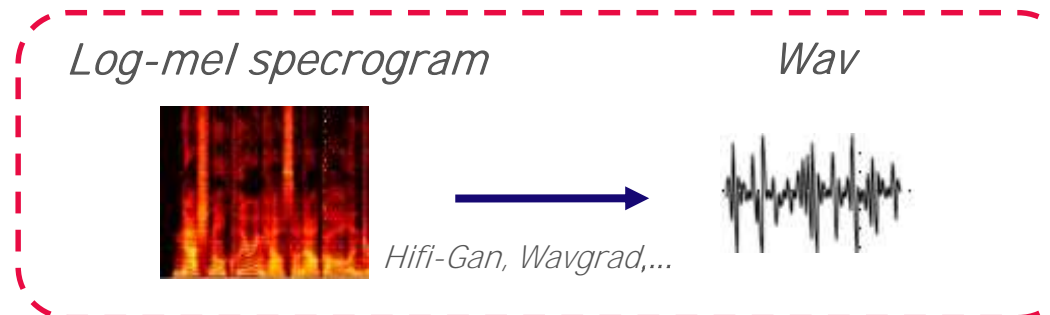


Exploiting music/audio knowledge in sound generation

- For example, a classic pipeline in recent Text-to-speech



- ...the task of sound generation from (mel) spectrogram



Exploiting music/audio knowledge in sound generation

From Mel-Spectrogram to waveform

Background

- Denoising Diffusion Probabilistic Models (DDPM)¹

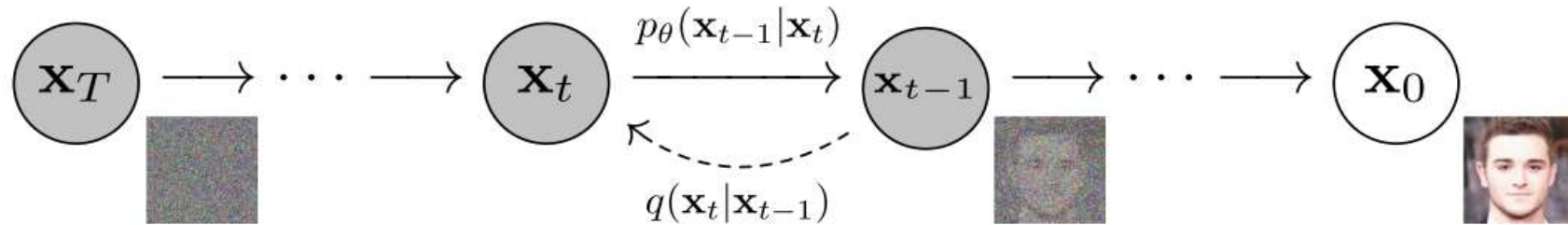


Figure: Overview of DDPMs

- **Framework:** Speech generation conditioned on mel spectrograms
- **Examples:** WaveGrad², SpecGrad³

¹Jonathan Ho/Ajay Jain/Pieter Abbeel: Denoising diffusion probabilistic models, in: *Proc. NeurIPS 33* (2020), pp. 6840–6851.

²Nanxin Chen et al.: WaveGrad: Estimating Gradients for Waveform Generation, in: *Proc. ICLR, 2020*.

³Yuma Koizumi et al.: SpecGrad: Diffusion Probabilistic Model based Neural Vocoder with Adaptive Noise Spectral Shaping, in: *Proc. Interspeech, 2022*.

Exploiting music/audio knowledge in sound generation

From Mel-Spectrogram to waveform

- Example: wavgrad, specgrad conditioned on mel-spectrogram

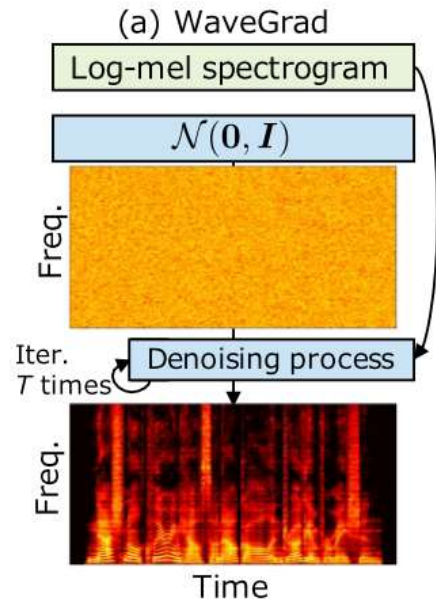
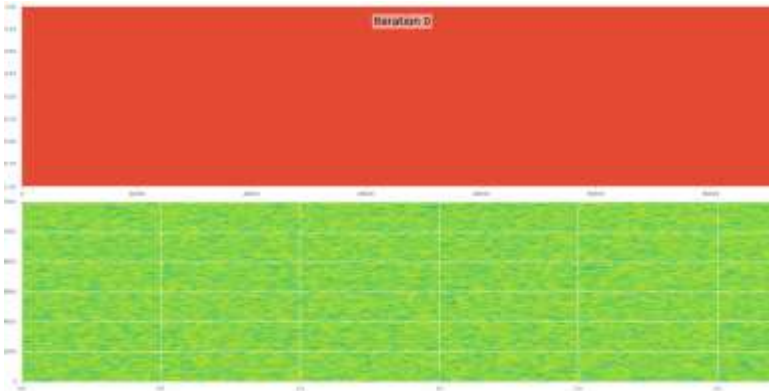


Illustration of the diffusion process (50 iterations)



Sound examples

reference



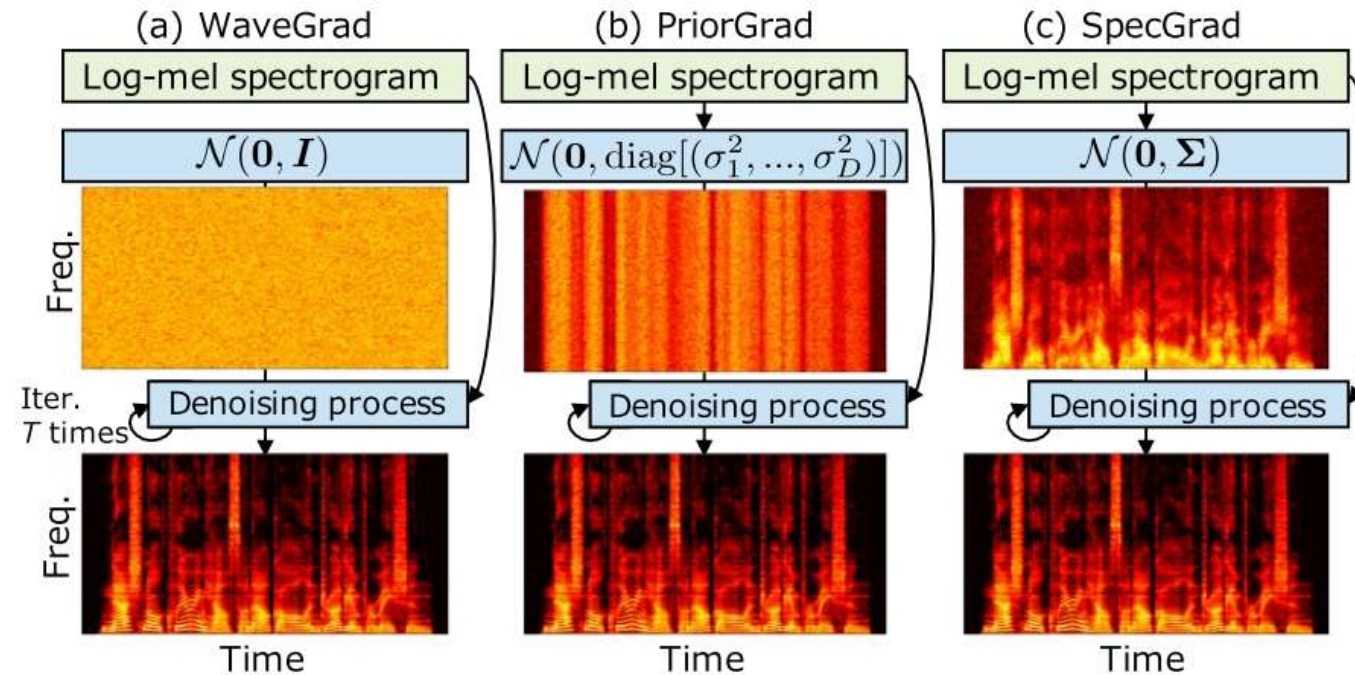
wavgrad



Exploiting music/audio knowledge in sound generation

From Mel-Spectrogram to waveform

- The example of priorgrad, specgrad, ...



Exploiting music/audio knowledge in sound generation

From Mel-Spectrogram to waveform

GLA-Grad: Method Overview

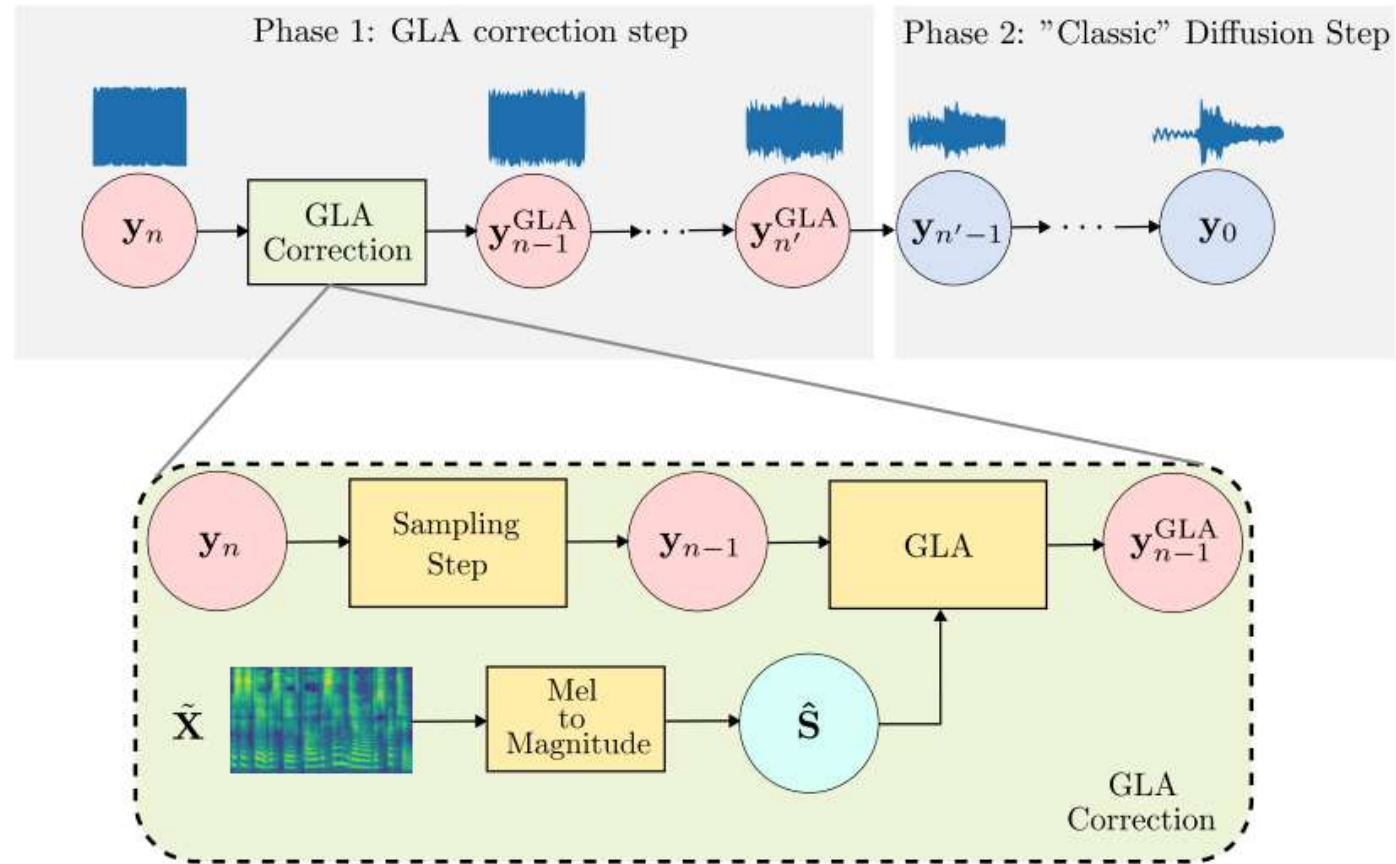
- **Griffin-Lim Algorithm (GLA)**⁴: magnitude spectrogram \rightarrow time-domain signal by phase estimation
- In DDPMs, the iteration process may lead to a signal at time $n - 1$ that is out of distribution of the training data if not carefully considered
 \rightarrow Introduce GLA into the inference process of WaveGrad to fix the bias between the generated signal and our expected signal
- Refining the spectrogram phase for the first steps of the diffusion process using a Griffin-Lim algorithm
- No retraining required

⁴Daniel Griffin/Jae Lim: Signal estimation from modified short-time Fourier transform, in: *IEEE Transactions on acoustics, speech, and signal processing* 32.2 (1984), pp. 236–243.

Exploiting music/audio knowledge in sound generation

From Mel-Spectrogram to waveform

- The GLA-Grad approach



Exploiting music/audio knowledge in sound generation

From Mel-Spectrogram to waveform

- **Results:**

- GLA-Grad: novel scheme for diffusion-based generation of speech from mel-spectrogram
- Stronger generalization performance to unseen speakers
 - → Advantage of incorporating a phase retrieval module
- Slight decrease in generation speed, but lower cost compared to a longer reverse process
 - → Good trade-off between quality and inference speed

Table: Results when training and evaluating on LJ Speech

Model	PESQ (↑)	STOI (↑)	WARP-Q (↓)
GLA-Grad	3.46 ± 0.11	0.963 ± 0.005	1.677 ± 0.076
WaveGrad	3.59 ± 0.13	0.970 ± 0.004	1.654 ± 0.075
WaveGrad-50	3.72 ± 0.11	0.978 ± 0.004	1.363 ± 0.054
SpecGrad	3.62 ± 0.14	0.963 ± 0.005	1.408 ± 0.054
Griffin-lim	1.02 ± 0.00	0.565 ± 0.042	3.234 ± 0.118

Table: Results when training on LJ Speech and evaluating on 19 speakers of VCTK

Model	PESQ (↑)	STOI (↑)	WARP-Q (↓)
GLA-Grad	2.73 ± 0.28	0.944 ± 0.017	1.722 ± 0.132
WaveGrad	2.08 ± 0.31	0.873 ± 0.035	1.913 ± 0.128
WaveGrad-50	2.00 ± 0.29	0.670 ± 0.111	2.122 ± 0.411
SpecGrad	2.48 ± 0.38	0.812 ± 0.066	1.593 ± 0.103
Griffin-lim	1.04 ± 0.01	0.522 ± 0.098	3.411 ± 0.164

Table: Results when training on 19 speakers of VCTK and evaluating on 90 other speakers of VCTK

Model	PESQ (↑)	STOI (↑)	WARP-Q (↓)
GLA-Grad	2.88 ± 0.44	0.856 ± 0.081	1.520 ± 1.102
WaveGrad	2.09 ± 0.48	0.803 ± 0.076	1.801 ± 0.133
WaveGrad-50	1.99 ± 0.38	0.706 ± 0.093	2.024 ± 0.157
SpecGrad	2.56 ± 0.35	0.814 ± 0.080	1.492 ± 0.127
Griffin-lim	1.04 ± 0.02	0.542 ± 0.112	3.410 ± 0.169



Exploiting music/audio knowledge in sound generation

From Mel-Spectrogram to waveform

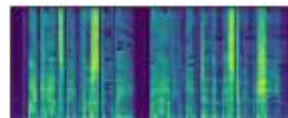
- GANS are difficult to train (GANs)
- Diffusion models have a slow inference speed

SpecDiff-Gan:

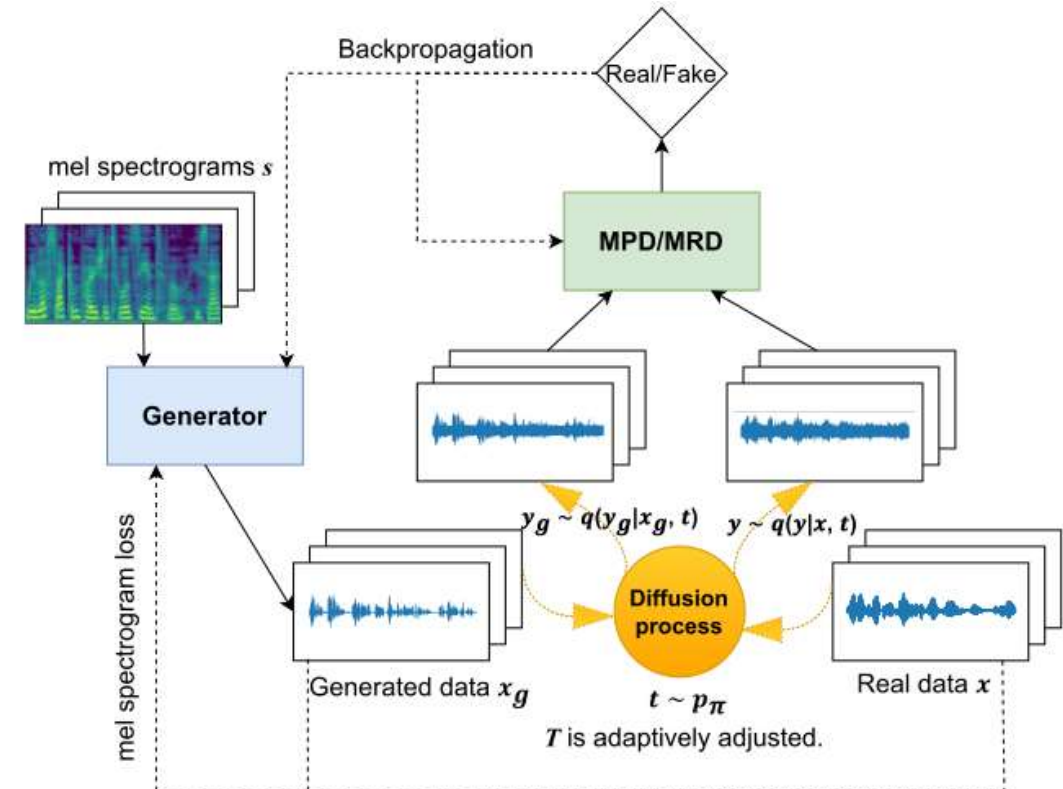
Combines principles of

- Diffusion-gans, Hifi-Gan and specgrad
- ...for *speech and music*

Spectrally shaped noise:



- to increase noise in low-energy regions, thereby challenging the discriminator.



Exploiting music/audio knowledge in sound generation

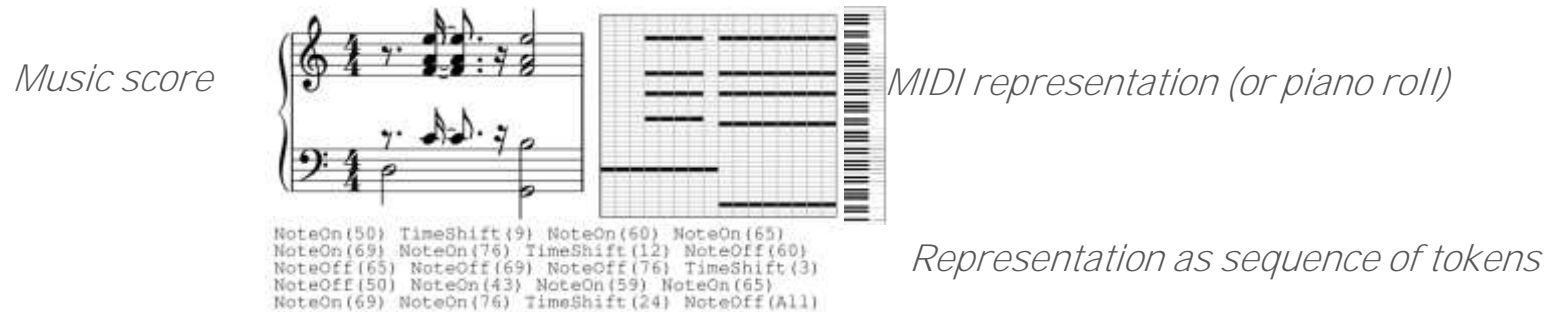
From Mel-Spectrogram to waveform

- SpecdiffGan:
 - Merged elements from diffusion models with GANs via the forward diffusion process
 - Enhancing stability and quality
 - Swift inference speed: ~ 200 times faster than real-time
 - Adaptability to various GAN-based audio synthesis models

	Ground truth	SpecDiff-Gan
speech		
piano		
drums		

Exploiting music knowledge in “Symbolic Music Generation”

- Symbolic music :



- Data-driven Symbolic Music Generation is difficult!
 - Inconsistency in melody and rhythm
 - Absence of multi-scale structures found in real music
 - ...

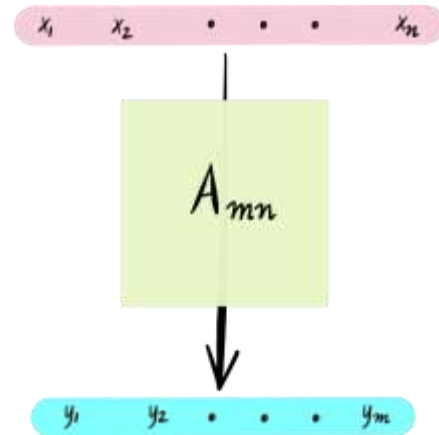


Add knowledge about musical structure to data-driven models



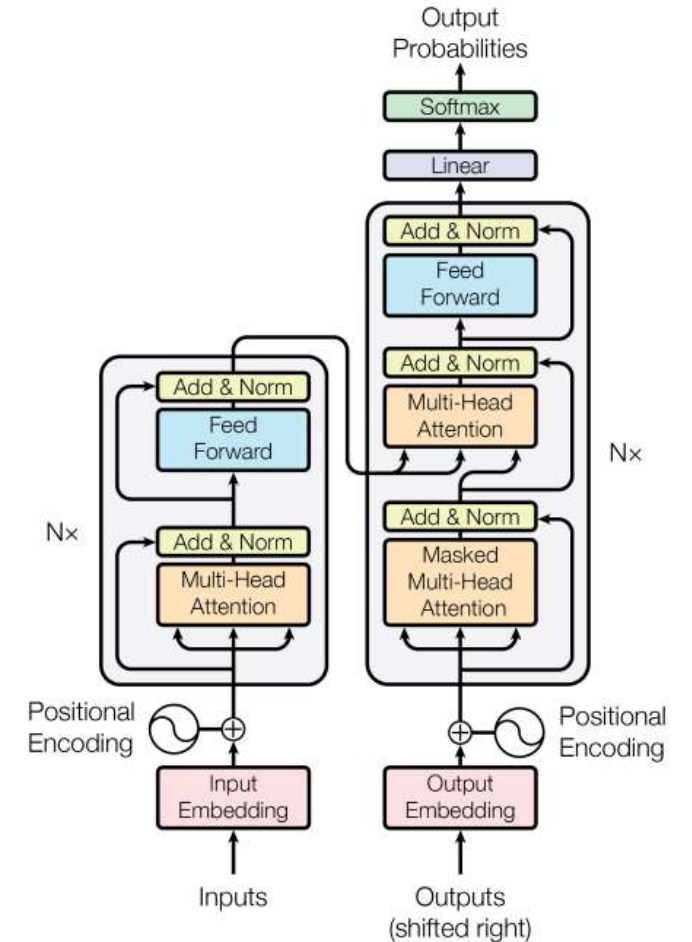
Symbolic Music Generation

- 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.

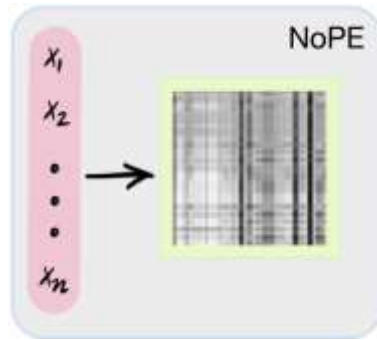
Towards exploiting « musical structured informed » Position Encoding (PE)



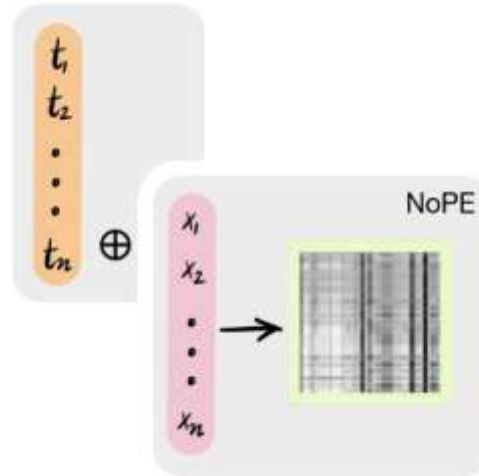
Symbolic Music Generation

« musical structured informed » Position Encoding (PE)

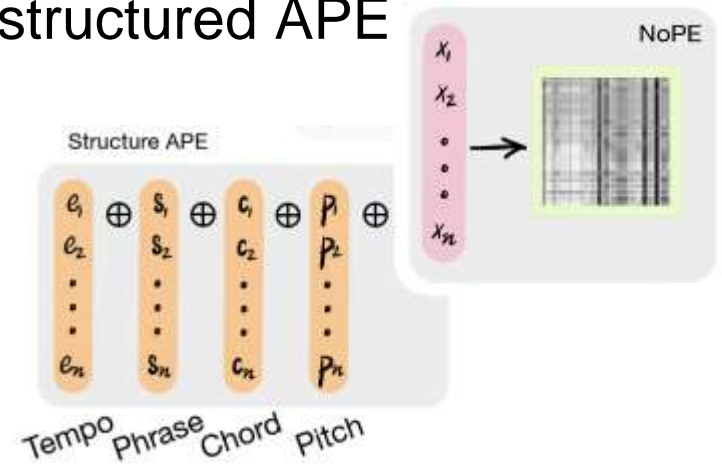
- From No Positional Encoding



...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

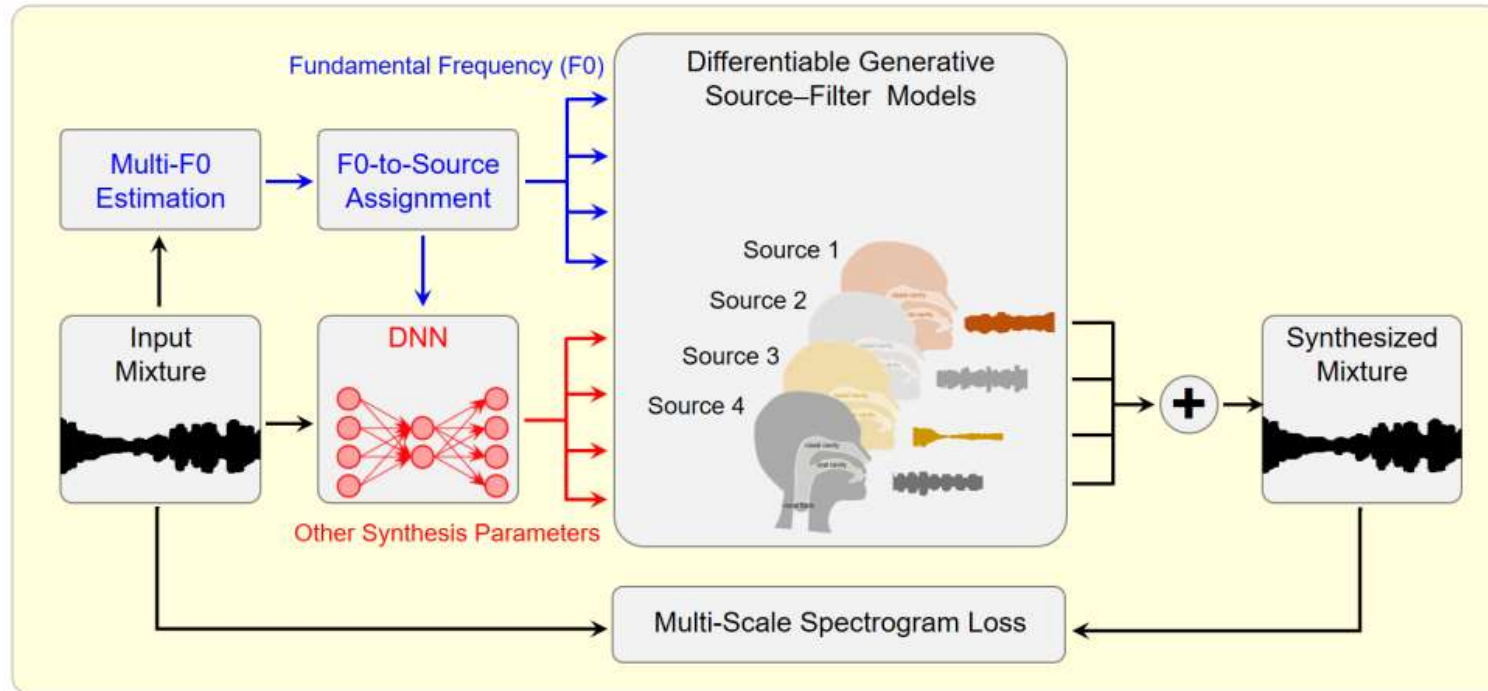


Content

- Context and motivation
- Towards hybrid deep learning
 - Some examples in other domains
 - Hybrid deep learning in audio
 - **Several application examples:**
 - *Audio synthesis*
 - ***Unsupervised music source separation***
- Discussion and conclusion

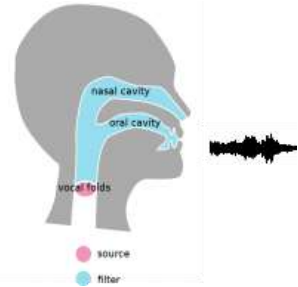
Exploiting speech production models for source separation

- An example for unsupervised singing voice separation



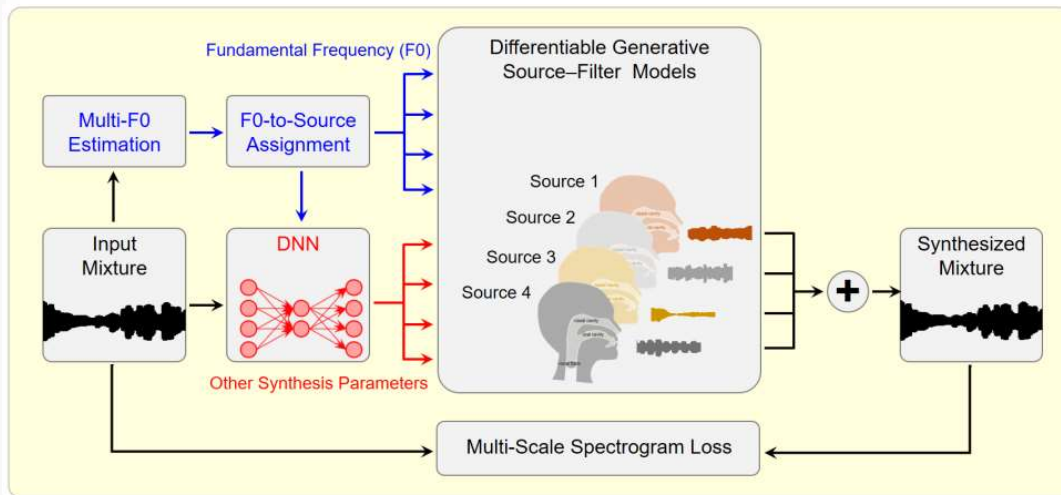
Exploiting speech production models for source separation

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*)

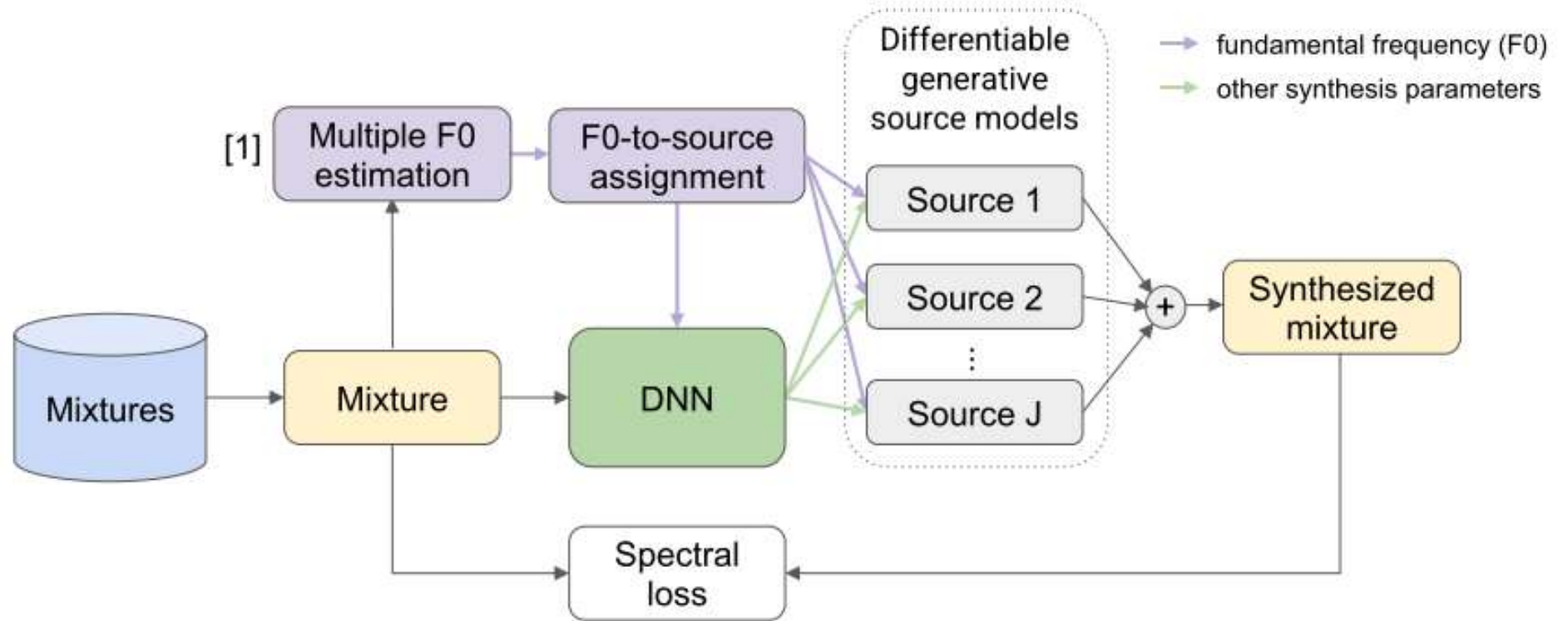
Novel sound transformation capabilities:

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

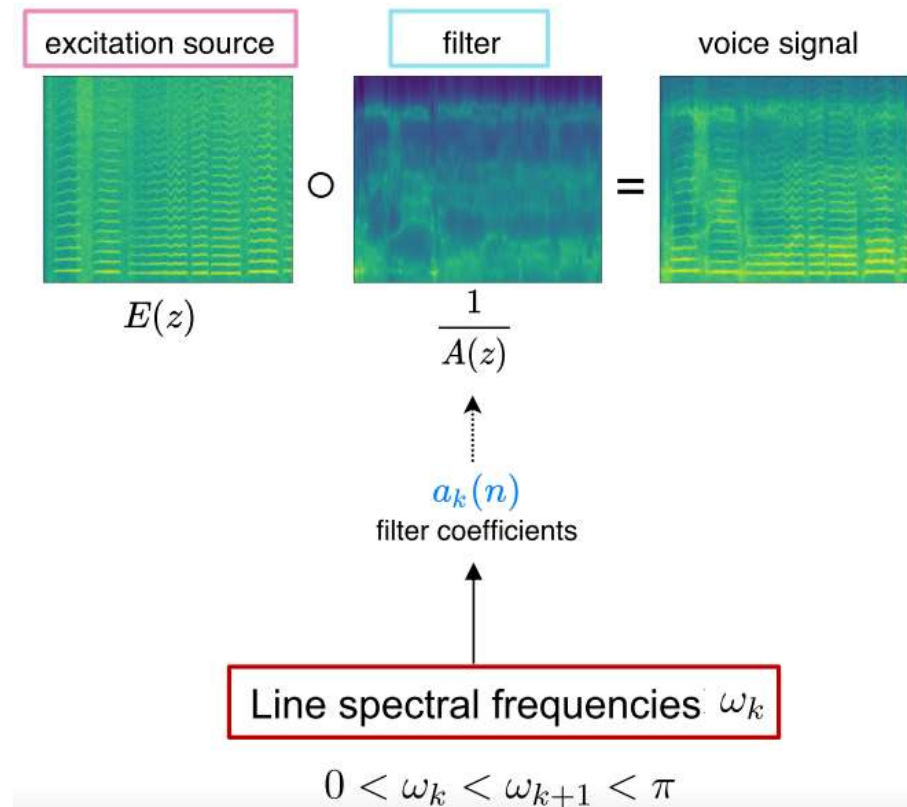
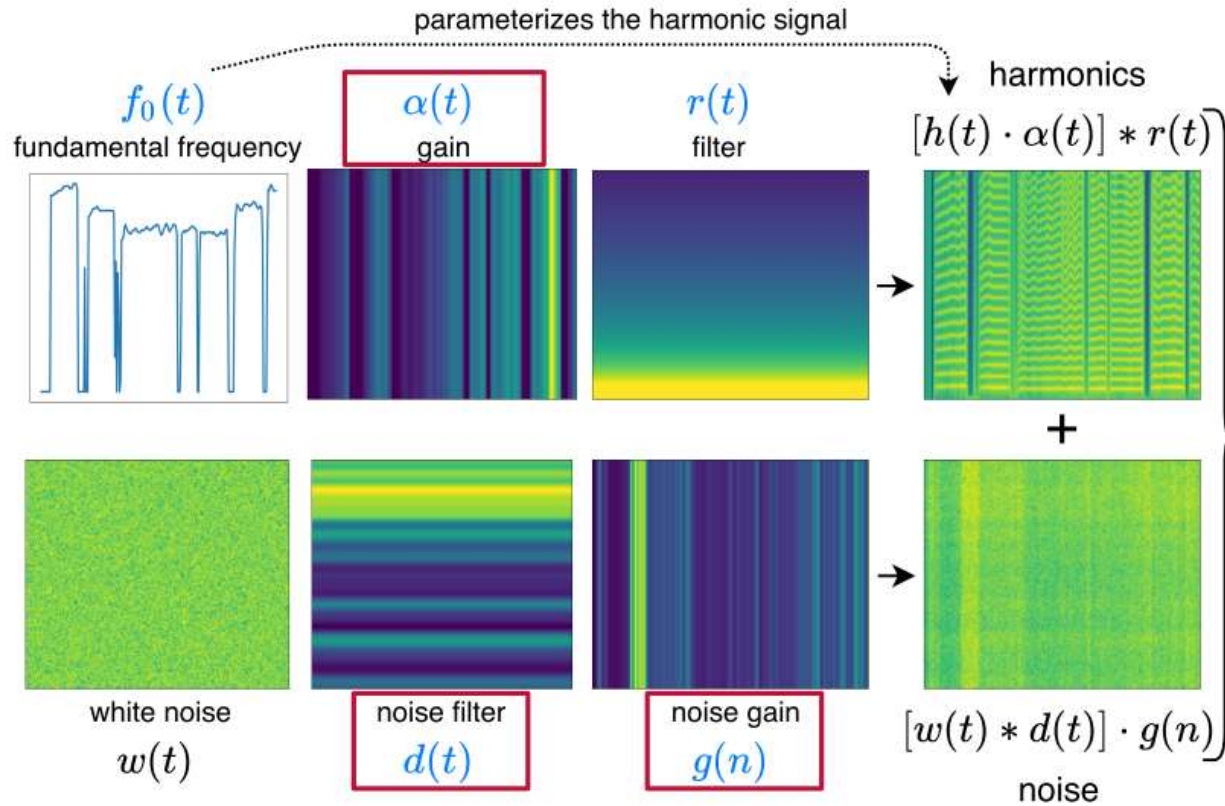


Unsupervised learning strategy

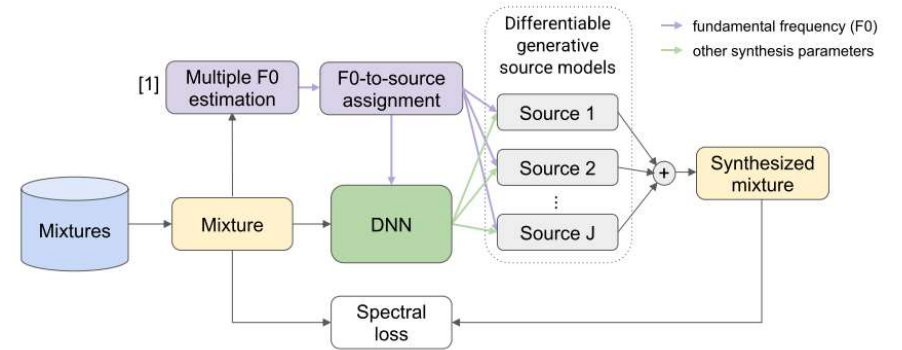
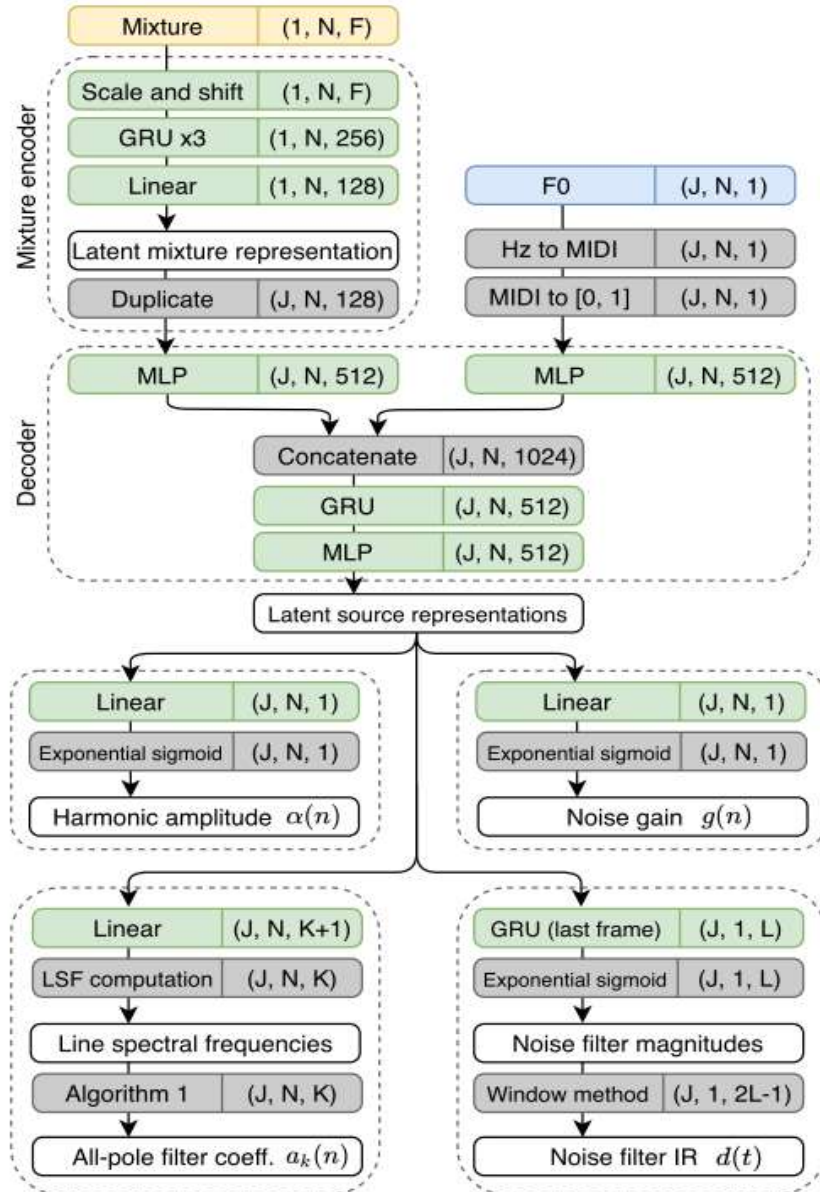
(e.g. no need of the individual source signals)



Parametric source models

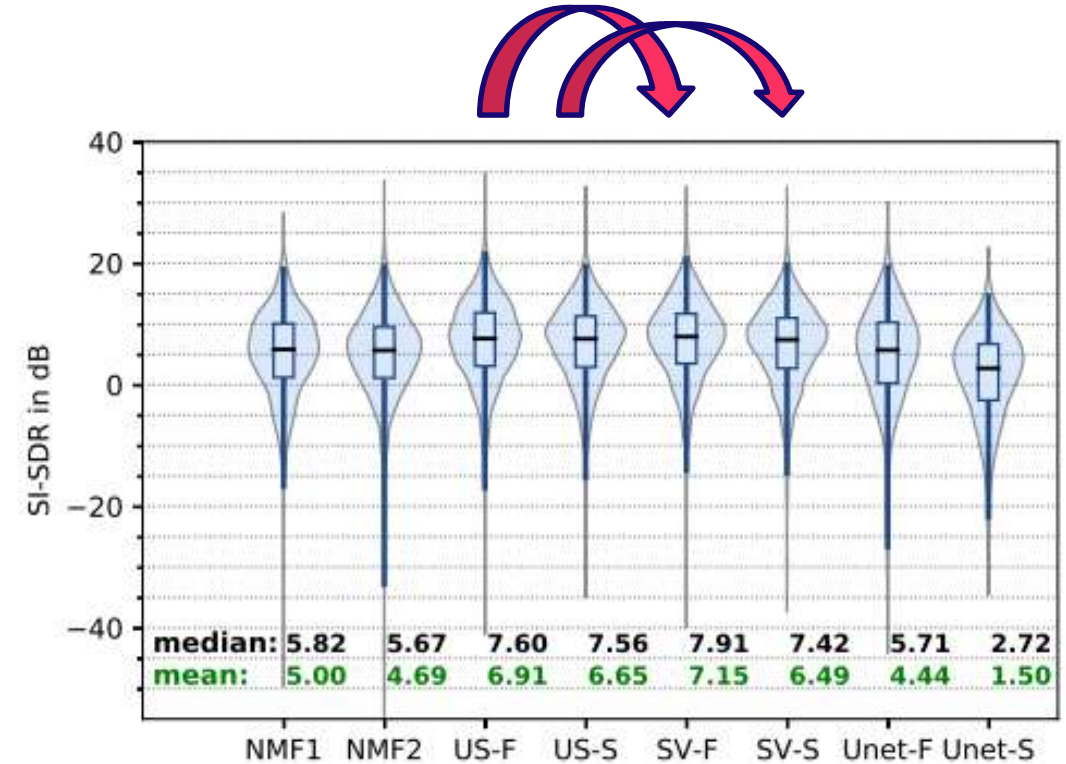


Global architecture overview



Some results

- Unsupervised (US) \approx supervised (SU)



(b) $J = 4$ sources



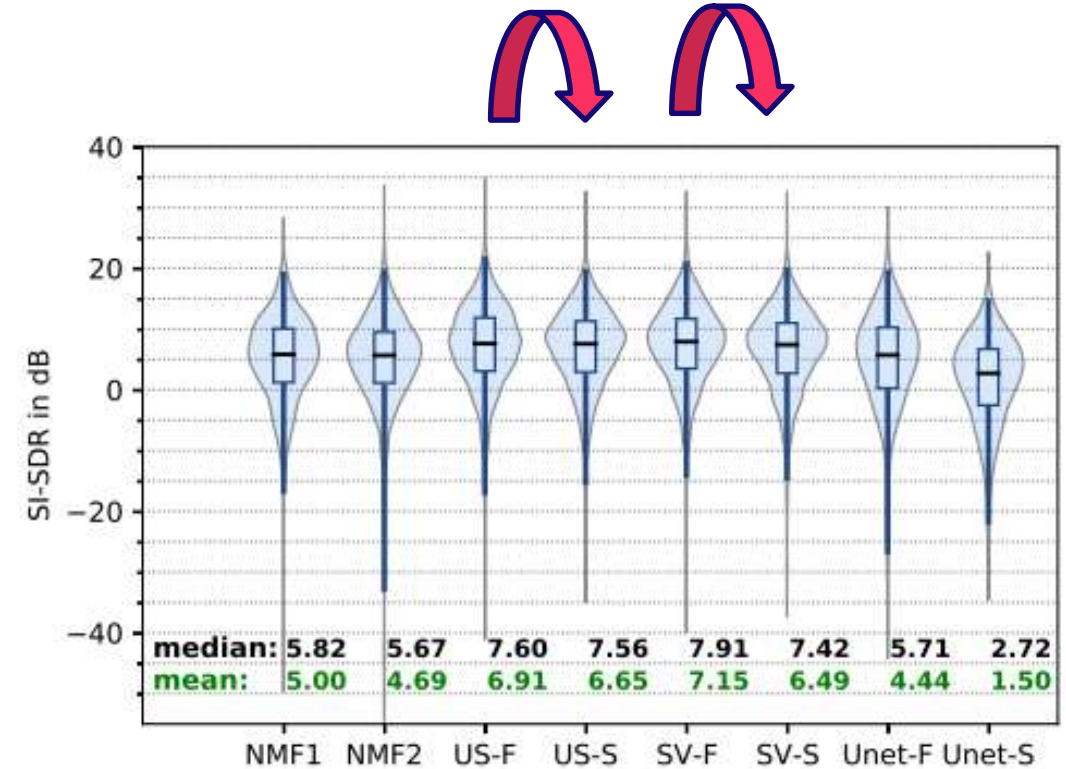
NMF1: S. Ewert and M. Müller, "Using score-informed constraints for NMF-based source separation," in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing. IEEE, 2012, pp. 129–132.

NMF2: J.-L. Durrieu, B. David, and G. Richard, "A musically motivated mid-level 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) \approx 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



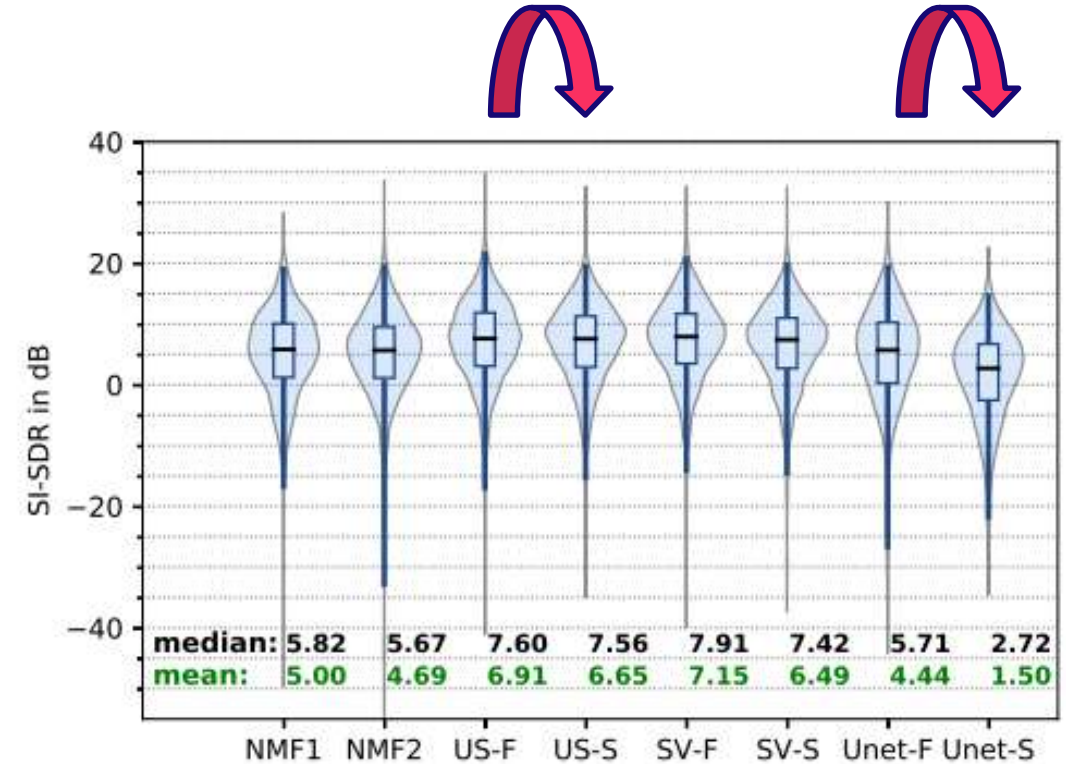
NMF1: S. Ewert and M. Müller, "Using score-informed constraints for NMF-based source separation," in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing. IEEE, 2012, pp. 129–132.

NMF2: J.-L. Durrieu, B. David, and G. Richard, "A musically motivated mid-level 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) \approx 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



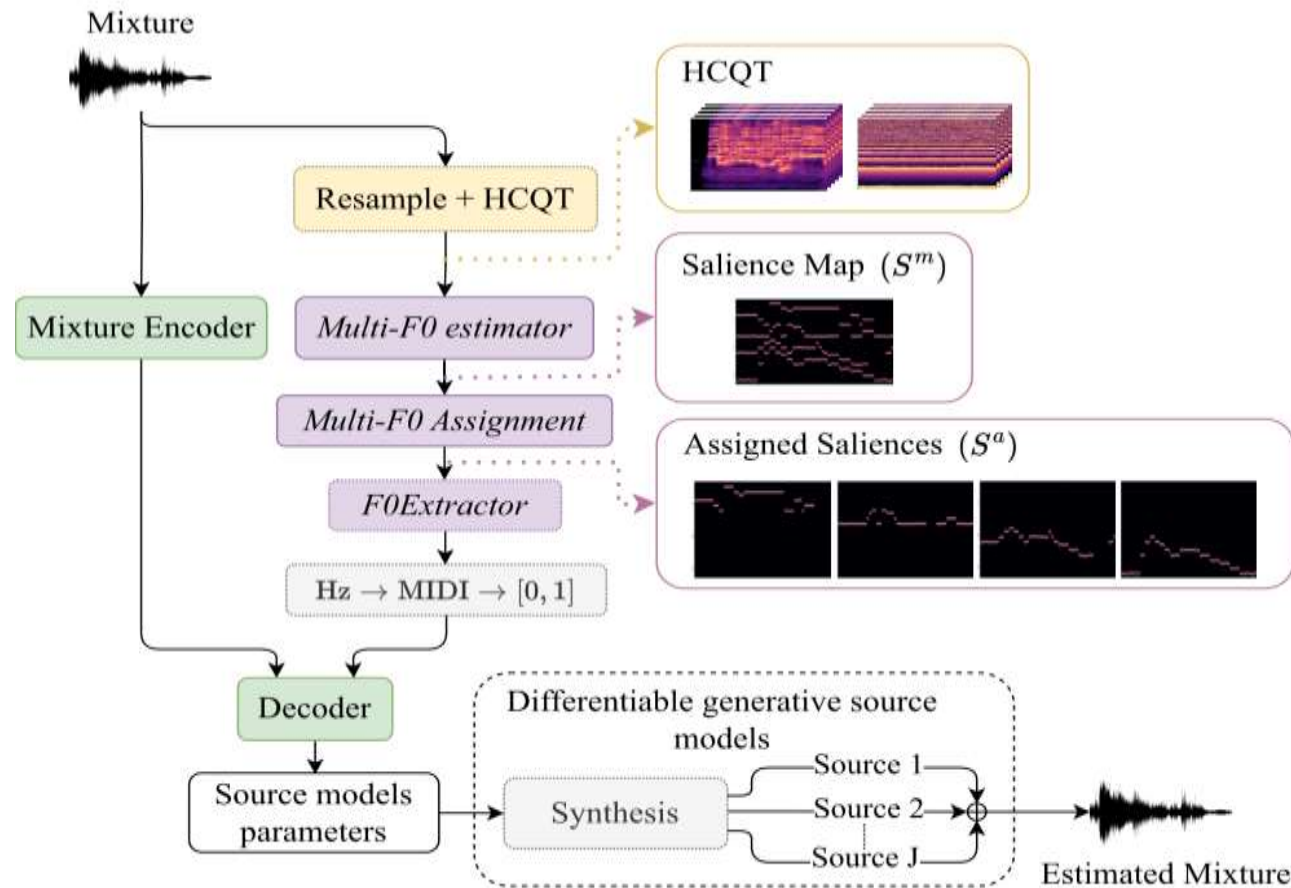
NMF1: S. Ewert and M. Mueller, "Using score-informed constraints for NMF-based source separation," in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing. IEEE, 2012, pp. 129–132.

NMF2: J.-L. Durrieu, B. David, and G. Richard, "A musically motivated mid-level 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.

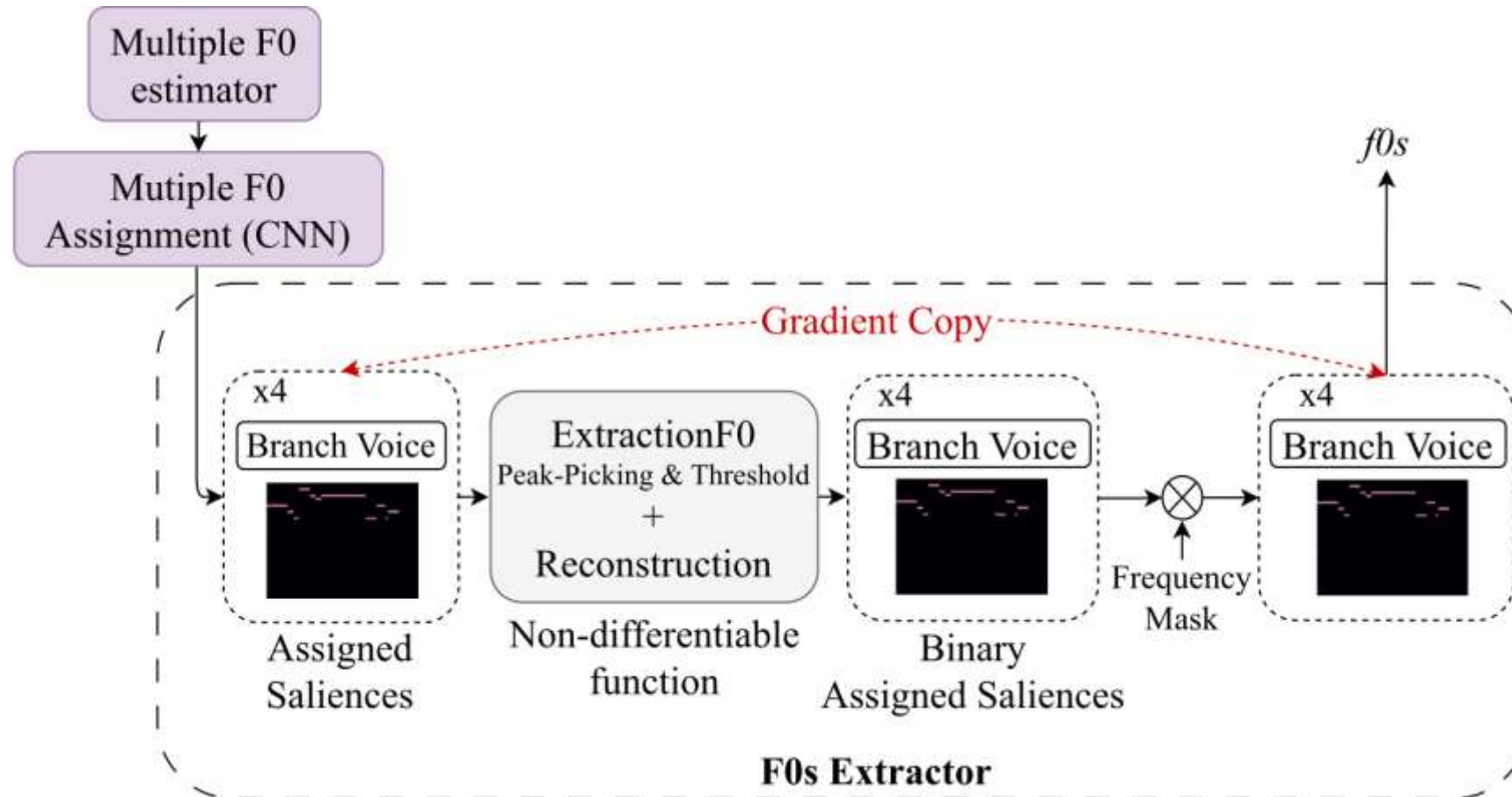
Towards a fully differentiable model for unsupervised singing voice separation

- Integration of multi-F0 extractor and automatic voice assignment



Towards a fully differentiable model for unsupervised singing voice separation

- Extraction of F0 sequences from assigned salience maps.



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)
- ... 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	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

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



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

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**