Hybrid deep learning for audio

Gaël RICHARD*

Professor, Telecom Paris, Institut polytechnique de Paris

*work with collaborators and in particular **K. Schulze-Forster**, C. Doire, R. Badeau

G. Richard

Content

- Context and motivation
- Towards hybrid deep learning
  - Some examples in other domains
  - Hybrid deep learning in audio
  - A specific example in unsupervised source separation
- Discussion and conclusion

Context and motivation

G. Richard

- 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

G. Richard

- 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 Hi-Audio :**

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

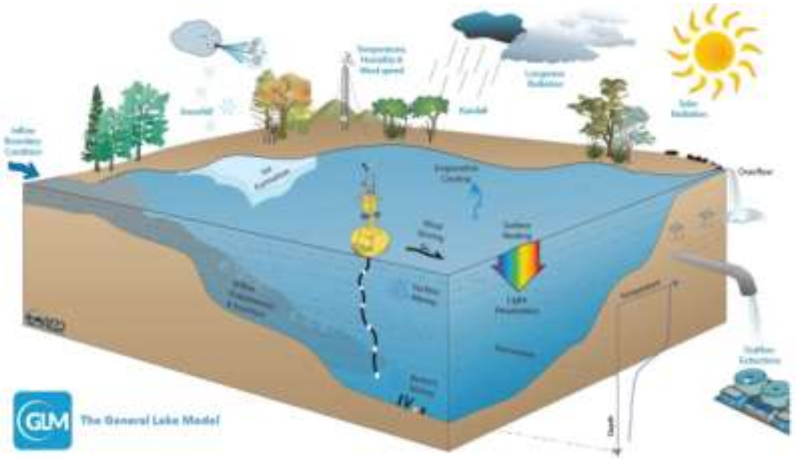
My approach: to build *Hybrid deep learning models*, by **integrating our prior knowledge** about the nature of the processed data.

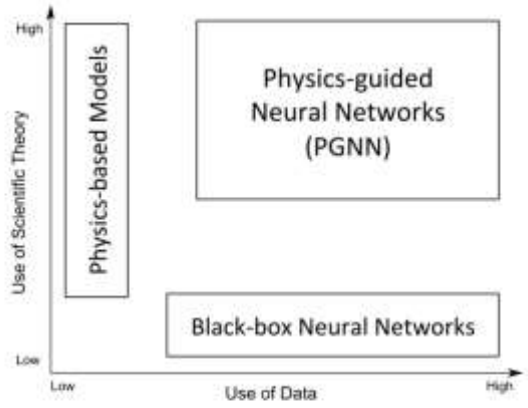
G. Richard

Towards Hybrid deep learning

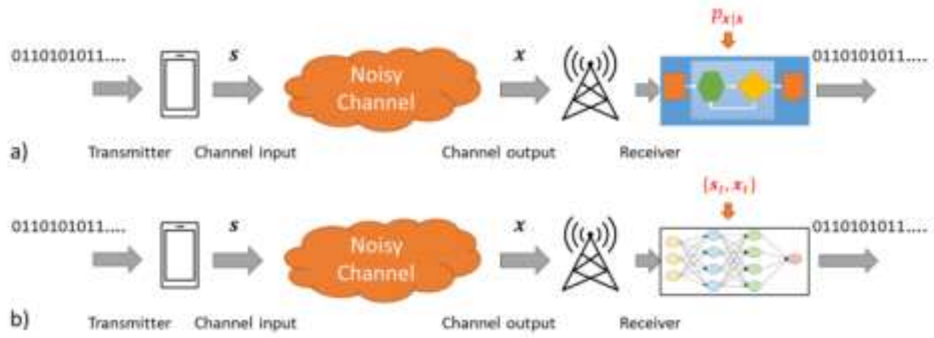
… some prior works.

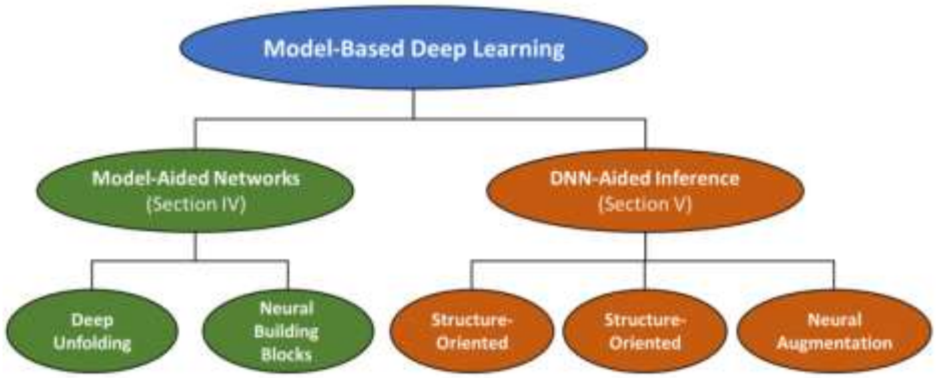
- Physics-guided neural networks in remote sensing [1],





- 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," arXiv, 2012.08405, 2020.

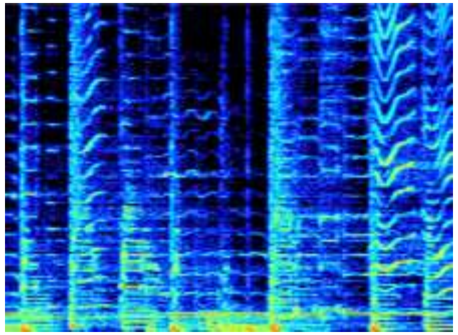
G. Richard

Why Hybrid deep learning is interesting for audio ?

- We often have some good apriori knowledge of the sources

- We have a long history of audio signal models:
  - **Audio perception:** hearing model, psychoacoustics,…
  - **Audio sound production:** source-filter, periodic/aperiodic, physical model,…
  - **Audio propagation:** room acoustics, reverberation, ...
  - **Audio signal models:** sparsity, factorisation, time-frequency representations, decomposition models …

- Audio requires specific deep neural networks architectures (compared to image processing)



G. Peeters, G. Richard, « Deep learning for audio» , Multi-faceted Deep Learning: Models and Data, Edited by Jenny Benois-Pineau, Akka Zemmari, Springer-Verlag, 2021 (to appear)

Towards Hybrid deep learning

… some prior works in Audio.

G. Richard

- **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 specrograms

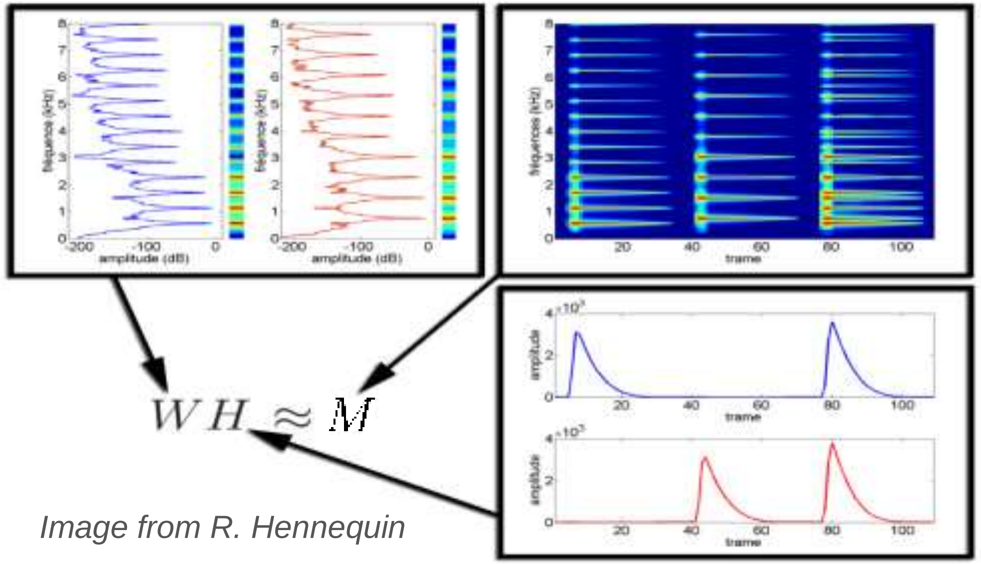


Image from R. Hennequin

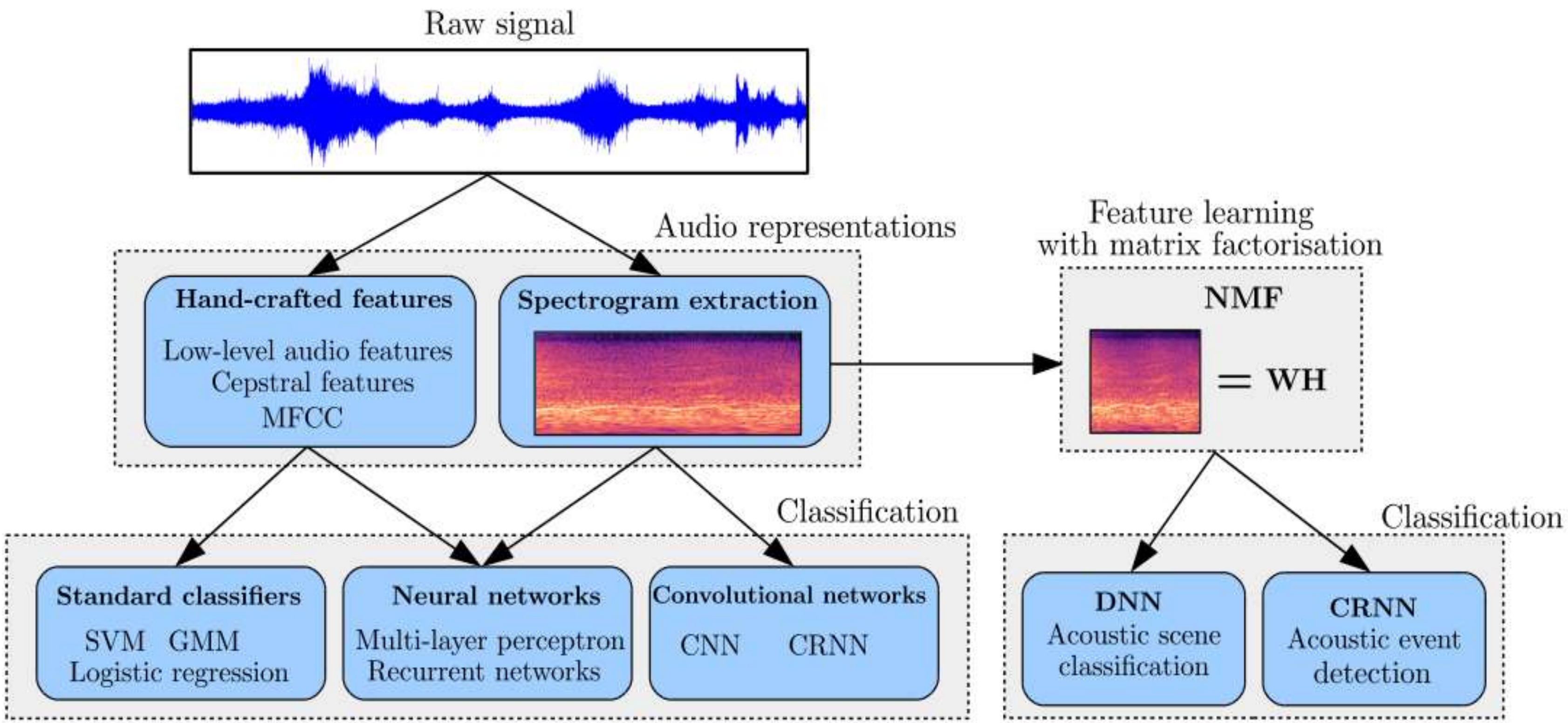
Lee, D.D. and Seung, H.S. (2001) Algorithms for Non-Negative Matrix Factorization. Advances in Neural Information Processing Systems, 13, 556-562.

Towards Hybrid deep learning

… some prior works in Audio.

G. Richard

- 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,

TELECOM Paris

IP PARIS

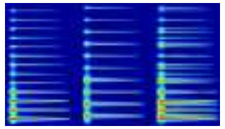
G. Richard

Towards Hybrid deep learning

… some prior works <u>in Audio</u>.

- **Deep NMF : the concept of deep unrolling**

  - Classic NMF



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

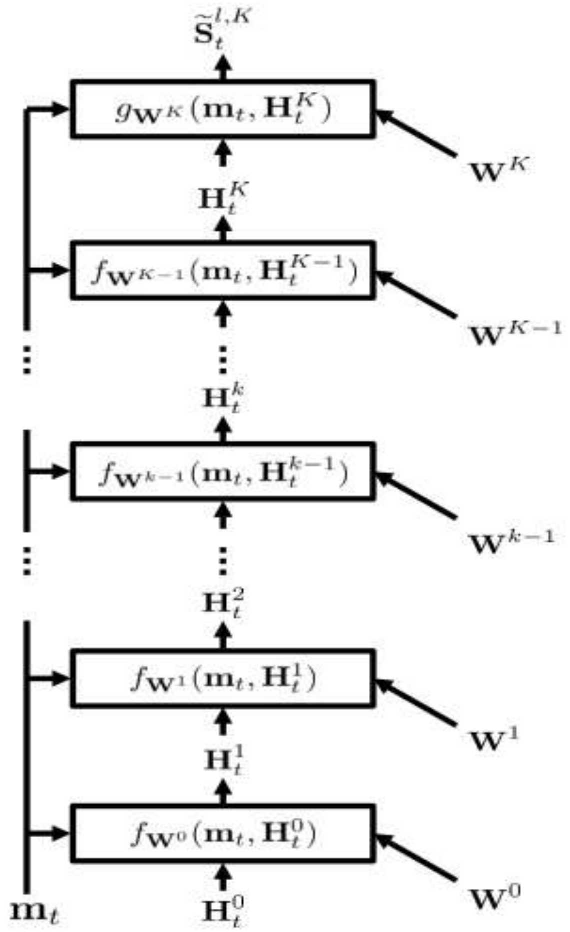
  - Minimizing a distance

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

  - ..towards iterative update rules

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

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



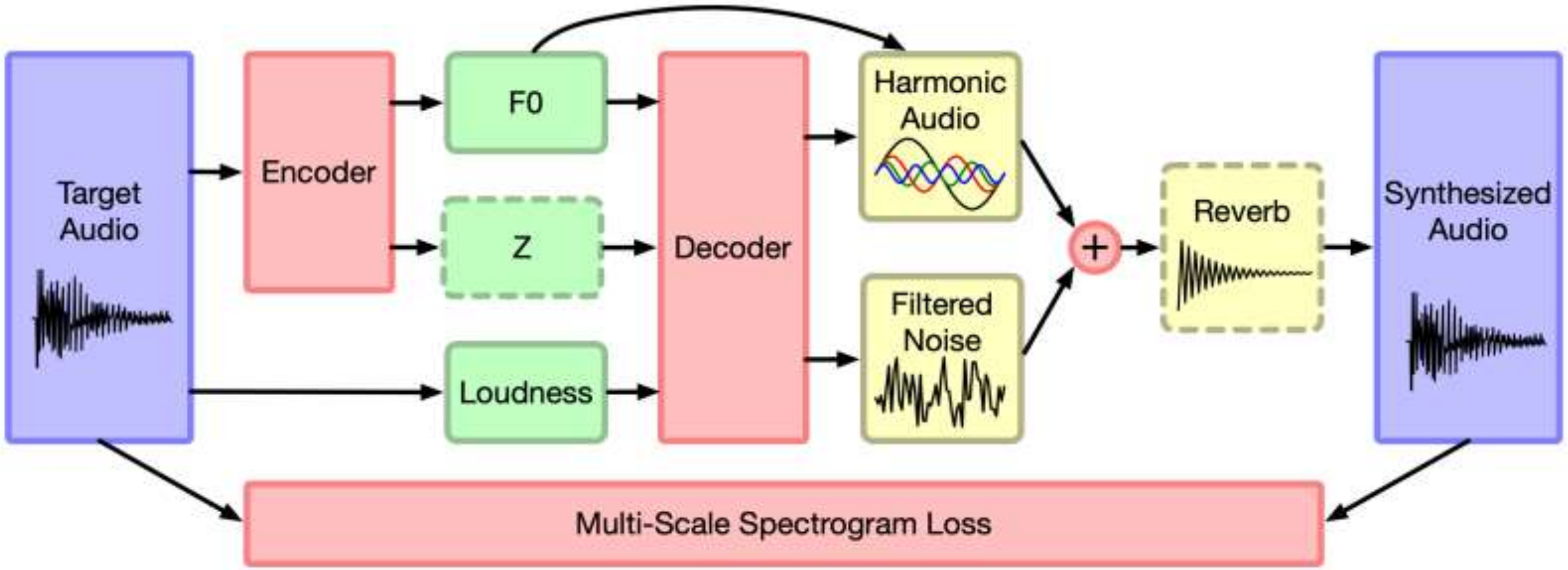
J. L. Roux & al., "Deep NMF for speech separation,," in IEEE Int. Conf. on Acous., Speech and Signal Proc. (ICASSP), 2015.

Towards Hybrid deep learning

… some prior works in Audio

G. Richard

- 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 Hybrid deep learning

… some prior works <u>in Audio</u>

- 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)

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

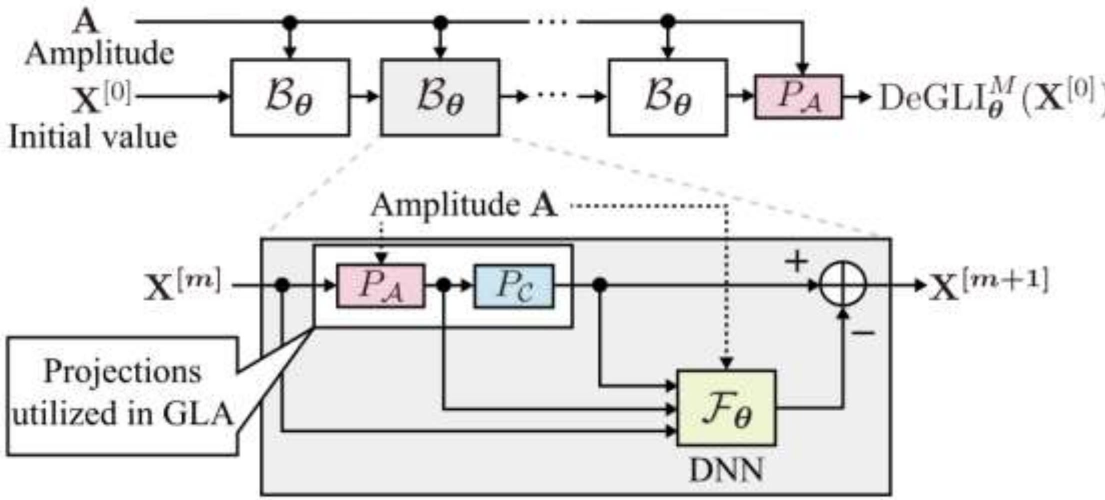
  - *Implemented as an iterative algorithm*

$$\mathbf{X}^{[m+1]} = P_{\mathcal{C}}(P_{\mathcal{A}}(\mathbf{X}^{[m]}))$$

$$P_{\mathcal{A}}(\mathbf{X})[\omega,\tau] = A[\omega,\tau])\frac{X[\omega,\tau]]}{|X[\omega,\tau]]|} \qquad P_{\mathcal{C}}(\mathbf{X}) = \mathcal{G}(\mathcal{G}^{\dagger}(\mathbf{X}))$$

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

Deep griffin-Lim



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,

Towards Hybrid deep learning

… some prior works in Audio

- ….And other very recent examples for virtual analog modelling combining Ordinary Differential Equations (ODEs) and neural network to learn the derivative function….
- … at DAFx 2022 !!

Proceedings of the 25th International Conference on Digital Audio Effects (DAFx20in22), Vienna, Austria, September 6-10, 2022

VIRTUAL ANALOG MODELING OF DISTORTION CIRCUITS USING NEURAL ORDINARY DIFFERENTIAL EQUATIONS

Jan Wilczek *
WolfSound
Katowice, Poland
jan.wilczek@thewolfsound.com

Alec Wright and Vesa Välimäki †
Acoustics Lab
Dept. Signal Processing and Acoustics
Aalto University, Espoo, Finland
alec.wright@aalto.fi

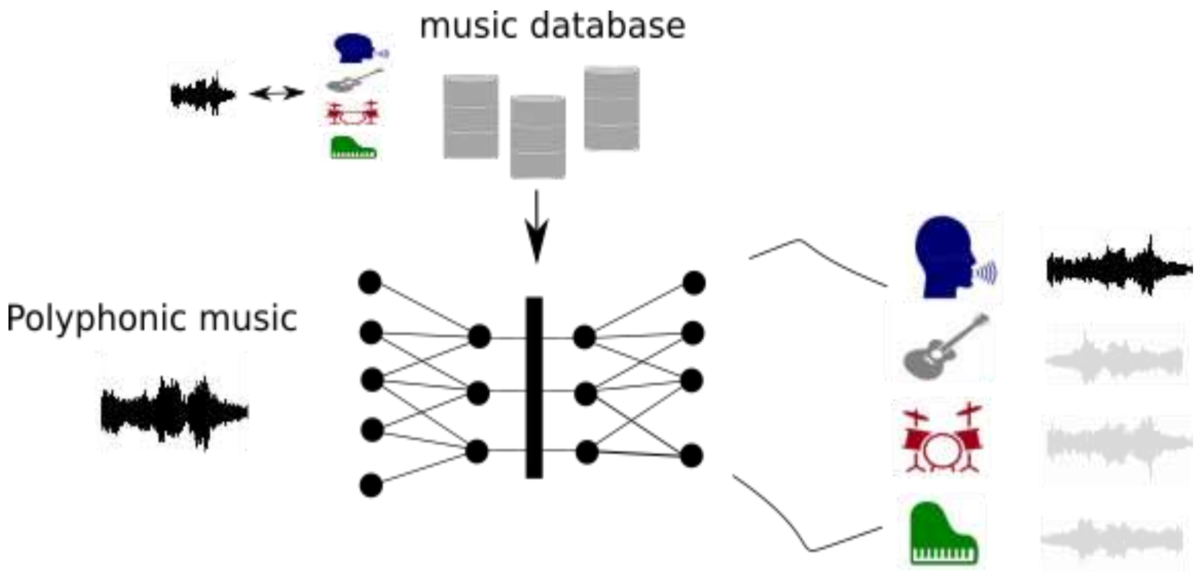
Emanuël A. P. Habets
International Audio Laboratories Erlangen ‡
Erlangen, Germany
emanuel.habets
@audiolabs-erlangen.de

G. Richard

Towards Hybrid deep learning

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

- For example in music source separation



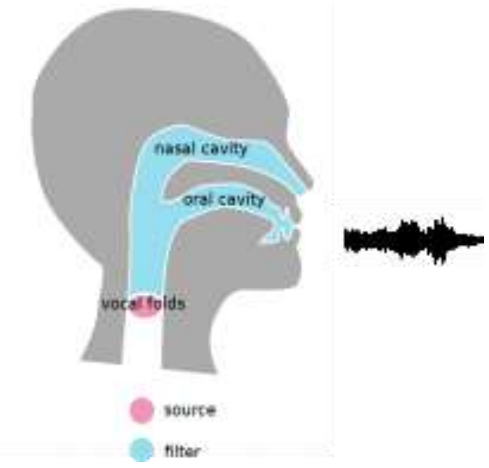
Main limitations:

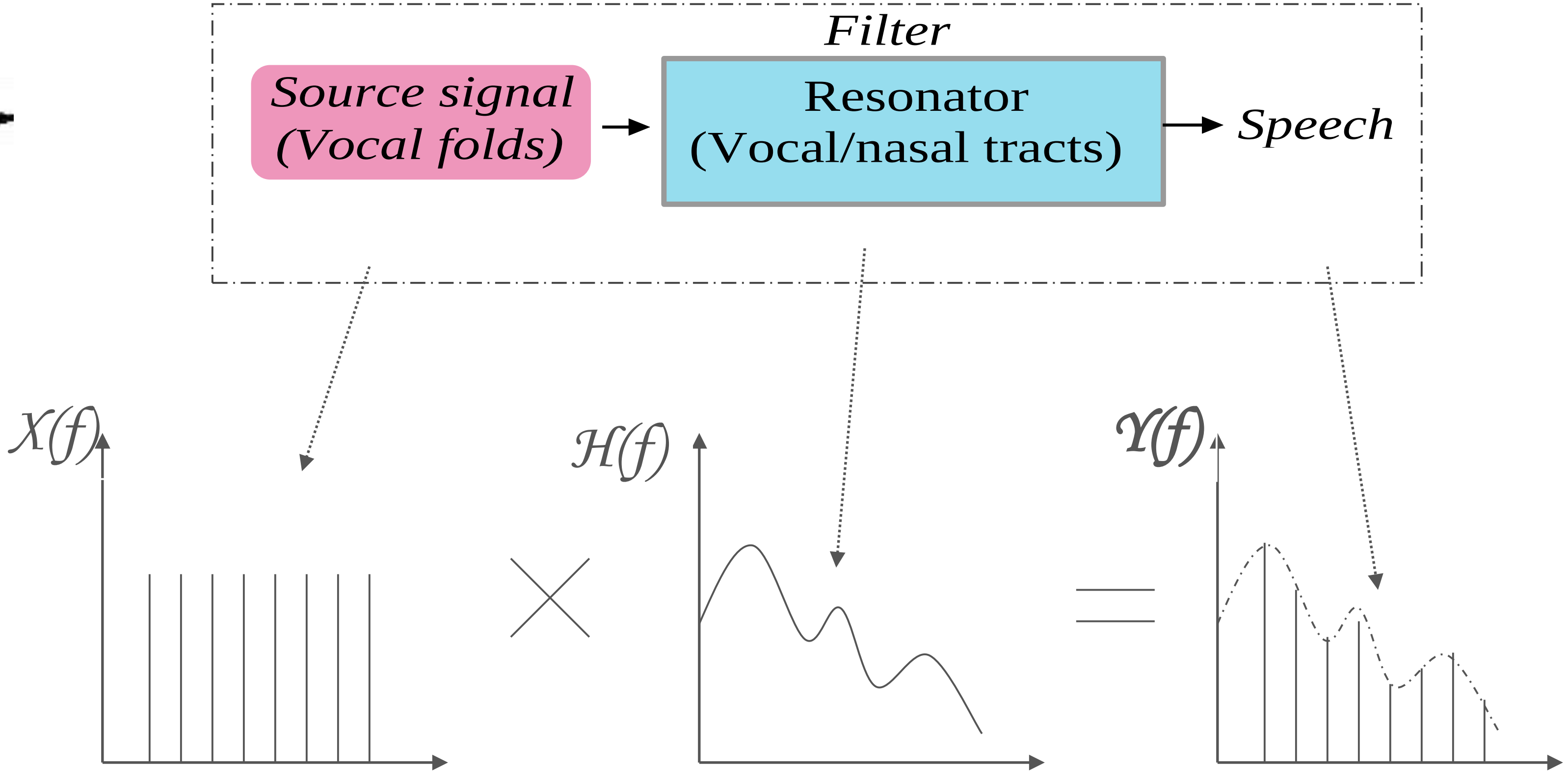
- *Difficulty to obtain « aligned » data*
- *Knowledge learned (only) from data*
- *Complexity: overparametrized models*
- Overconsumption regime
- **Non-interpretable/non-controllable**

G. Richard

The source filter model

an efficient speech production model





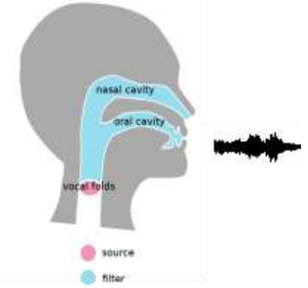
Fant, G. Acoustic theory of speech production, 1960, The Hague, The Netherlands, Mouton.

G. Richard

Towards Hybrid deep learning

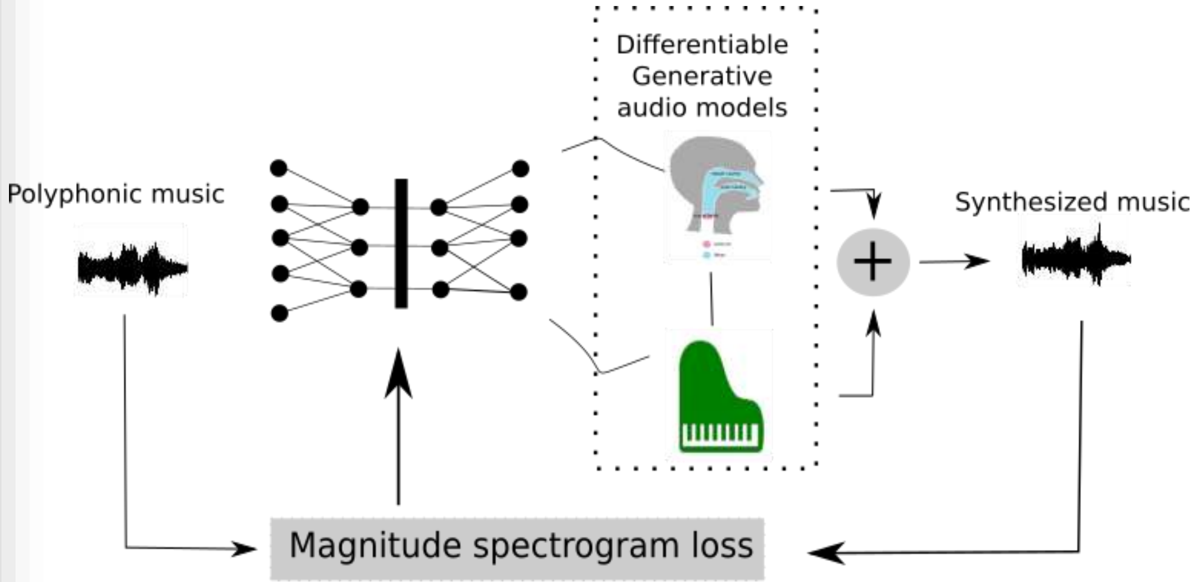
… by **integrating our prior knowledge** about the nature of the processed data.

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

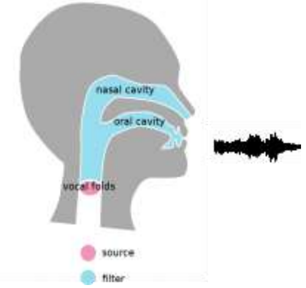


Towards Hybrid deep learning

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

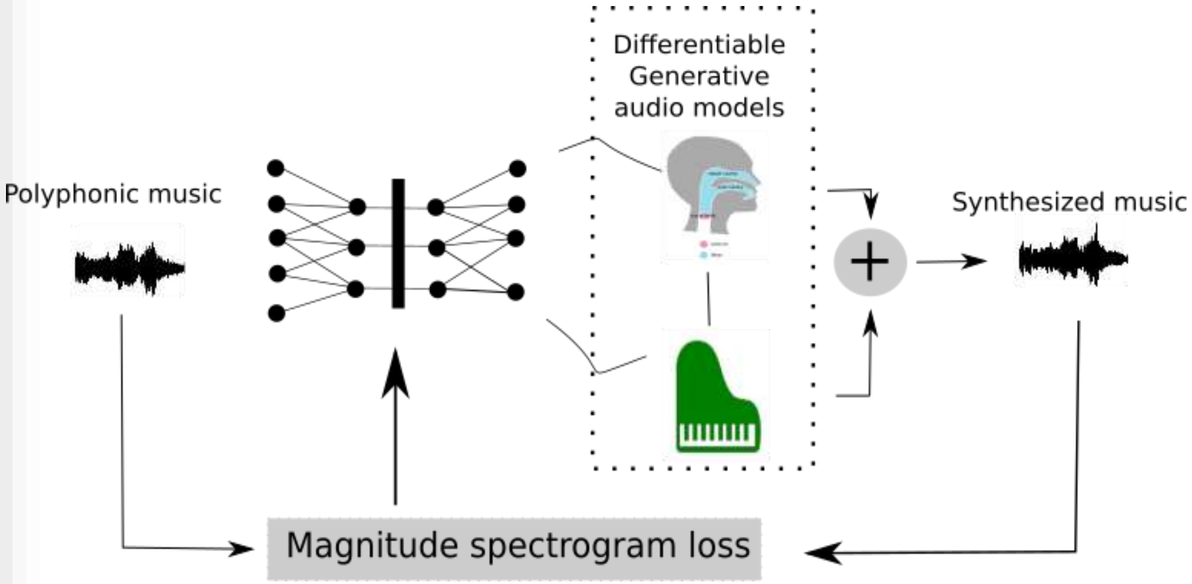
G. Richard

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



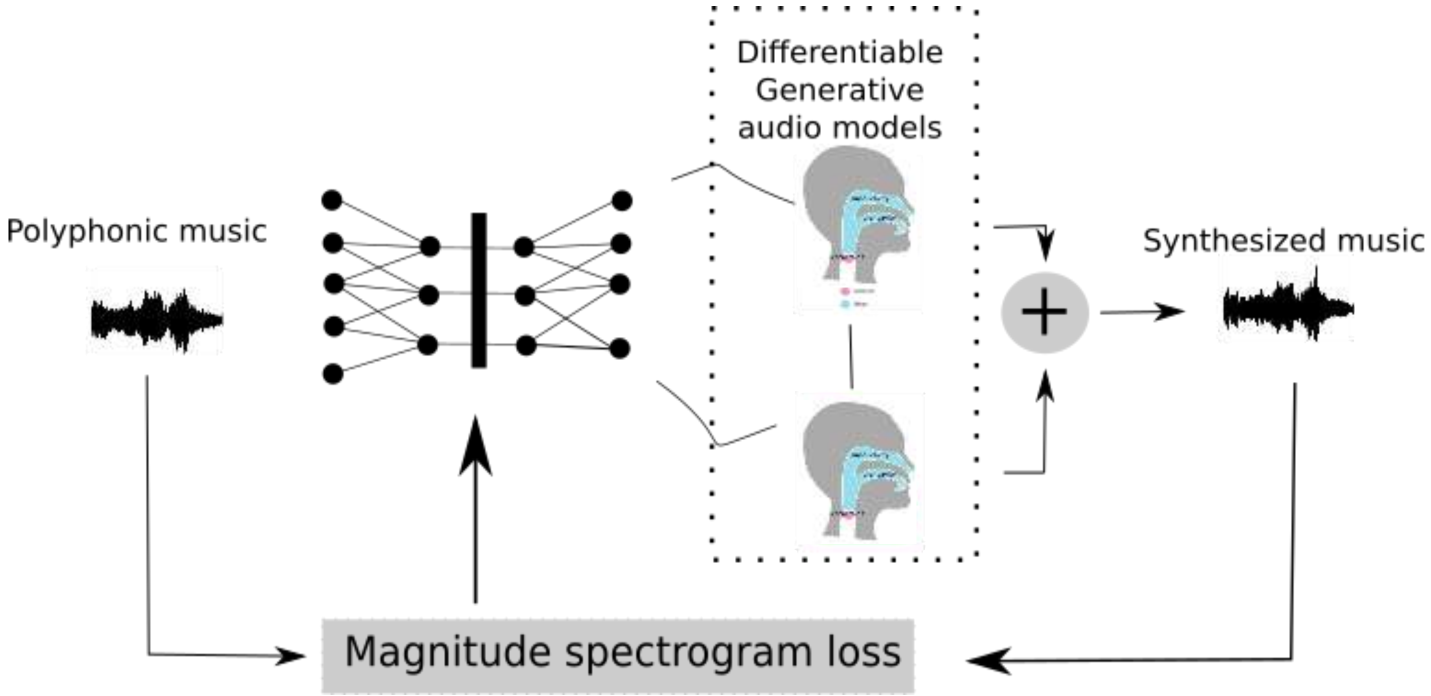
A new paradigm

- 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*)

Towards Hybrid deep learning

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

- Preliminary work on source separation (choir singing)



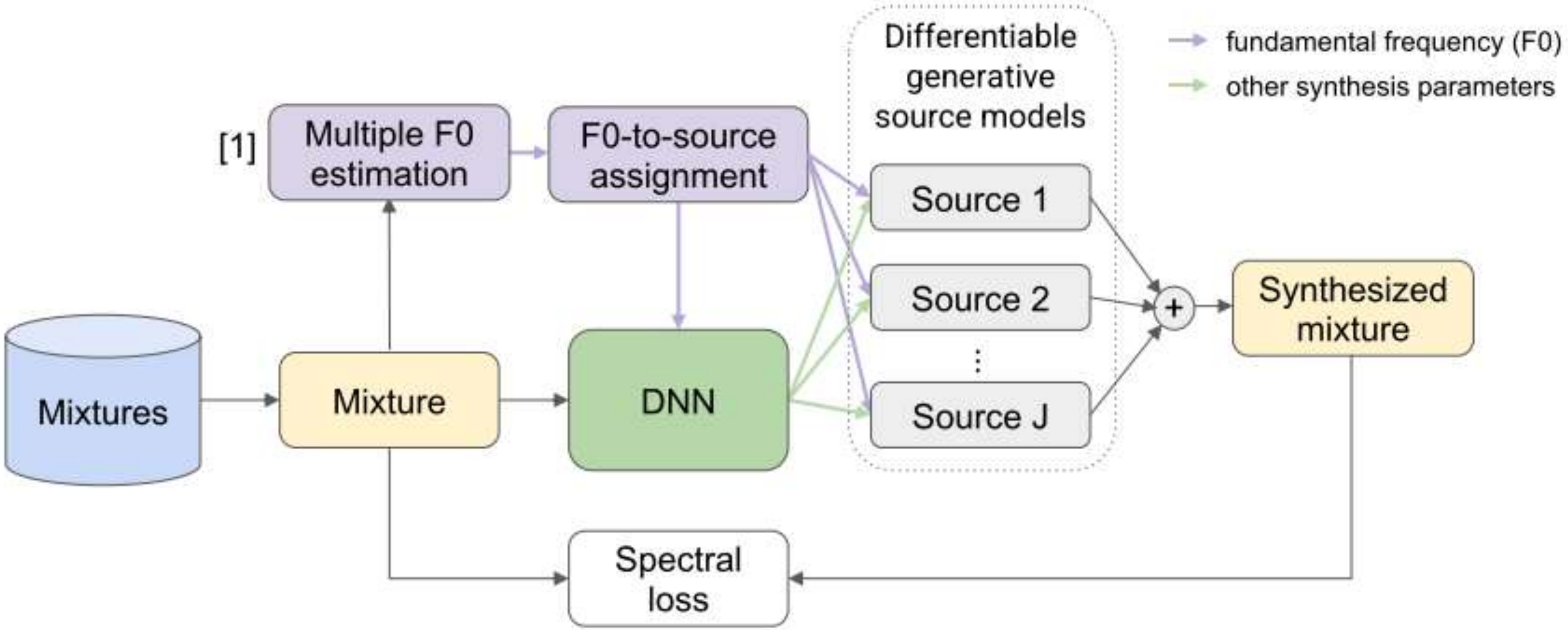
K Schulze-Forster, CSJ Doire, G Richard, R Badeau [Unsupervised Audio Source Separation Using Differentiable Parametric Source Models](), arXiv preprint arXiv:2201.09592

G. Richard

G. Richard

Unsupervised learning strategy

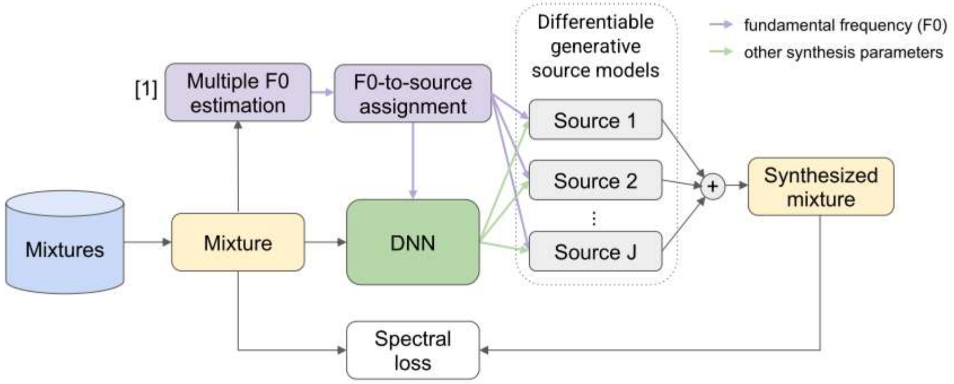
(e.g. no need of the individual source signals)



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

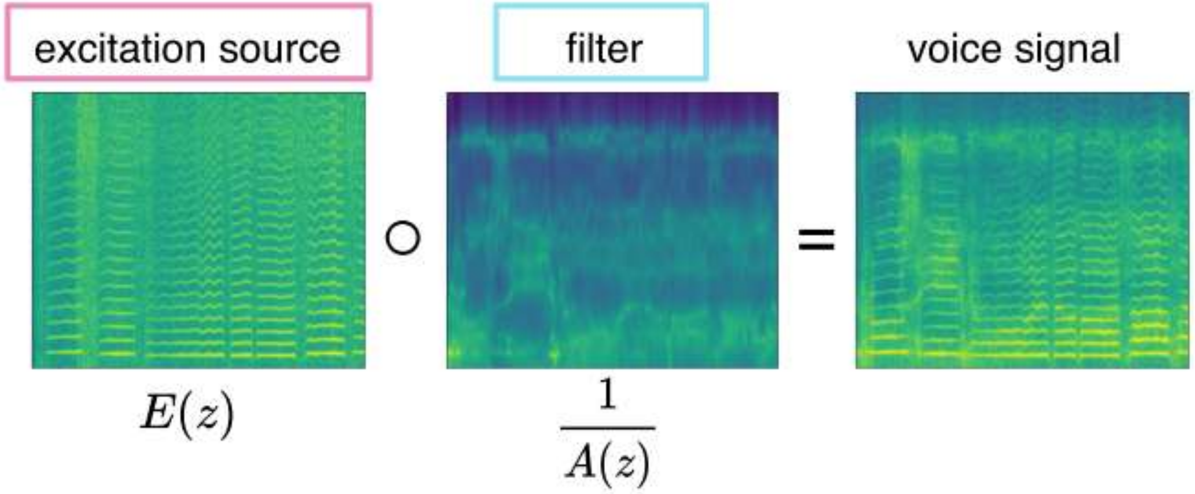
Parametric source models

G. Richard



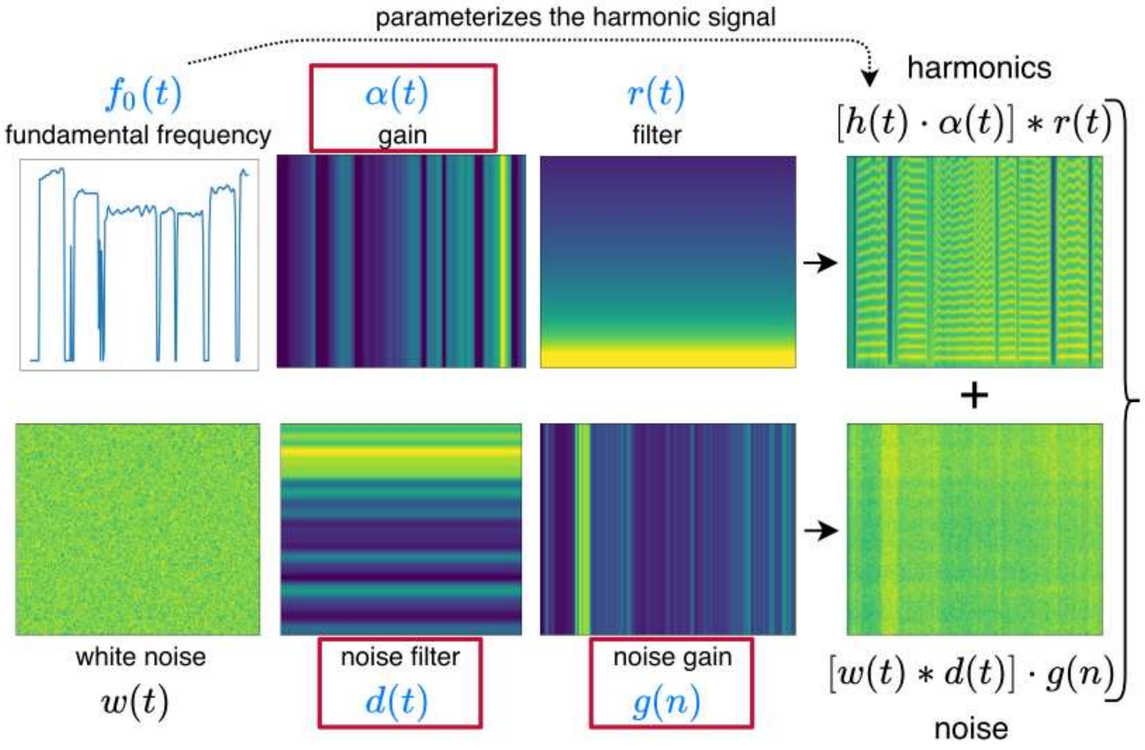
Singing voice as a source / filter model :

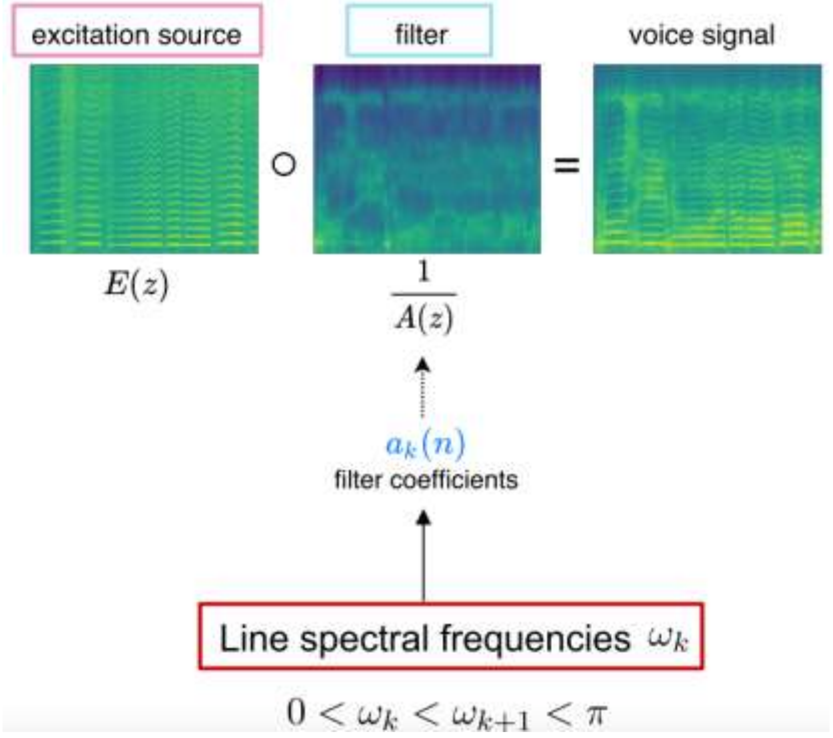
- source = vibration of vocal folds
- Filter = resonances of vocal/nasal cavities



J. Engel, C. Gu, A. Roberts et al., “DDSP: Differentiable digital signal processing,” in Proc. Int. Conf. Learning Representations, 2019

Parametric source models





The Line Spectral frequencies (LSF)

G. Richard

- An alternative to the linear prediction coefficients

$$y(n) = x(n) + e(n) = x(n) - \sum_{i=1}^{P} a_i \; x(n-i)$$

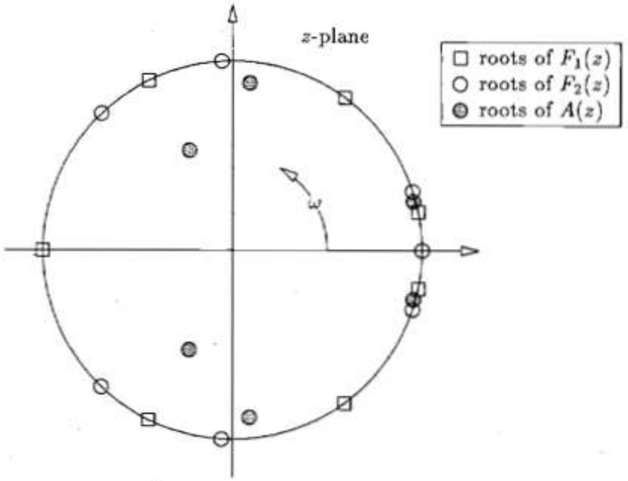
$$y(n) = x(n).(1 - \sum_{i=1}^{P} a_i z^{-1})$$

- Use of two auxilliary polynomials

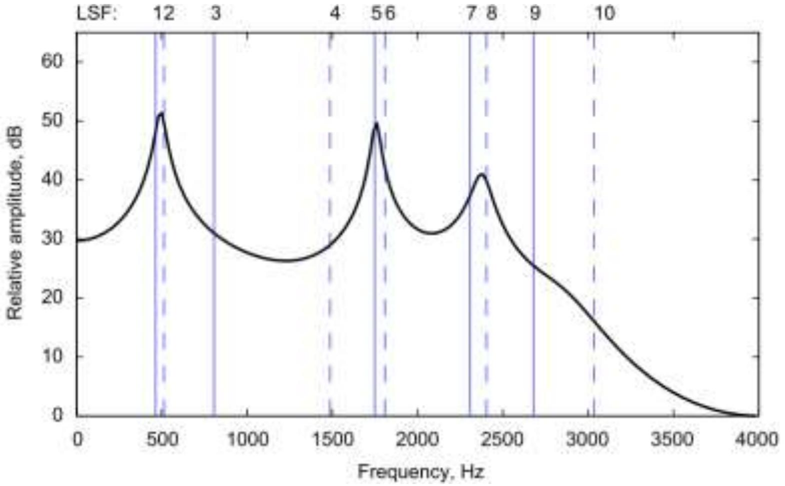
$$Y(z) = A(Z)E(z) = (1 - a_1 z^{-1} - \cdots - a_P z^{-P})E(z)$$

$$F_1(z) = A(Z) + z^{-(P+1)} A(z^{-1})$$

$$F_2(z) = A(Z) - z^{-(P+1)} A(z^{-1})$$



An example of LPC spectrum (from [2])



[1] P. Kabal and R. P. Ramachandran, "The computation of line spectral frequencies using Chebyshev polynomials," IEEE/ACM Trans. on Audio, Speech, and Language Processing, vol. 34, no. 6, pp. 1419–1426, 1986.
[2] I. McLoughlin, Line spectral pairs, Signal Processing 88 (2008) 448–467

Unsupervised learning strategy

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

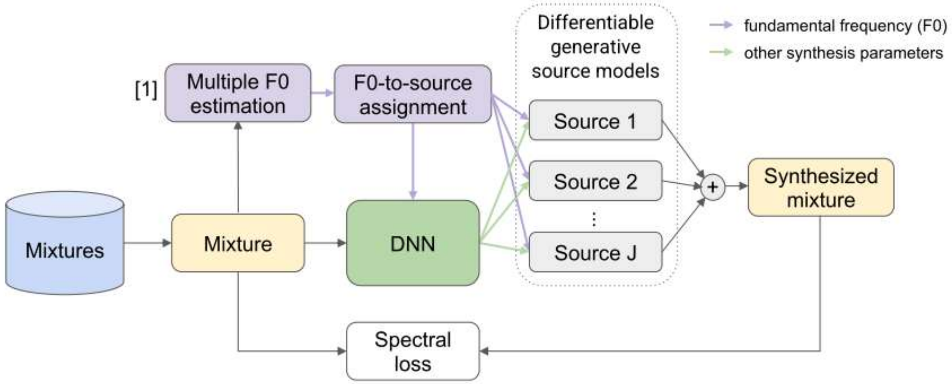
G. Richard

- A multi-scale spectral loss

$$\mathcal{L}_{rec} = \sum_c \mathcal{L}_c$$

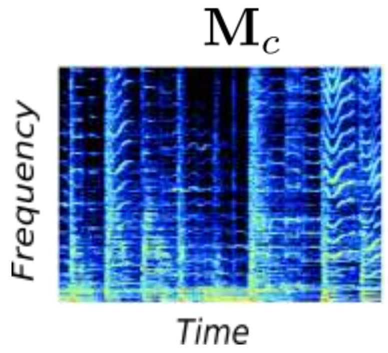
With

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



Where $\mathbf{M}_c$ and $\tilde{\mathbf{M}}_c$ denote the magnitude spectrograms of the input mixture and its estimate, respectively,

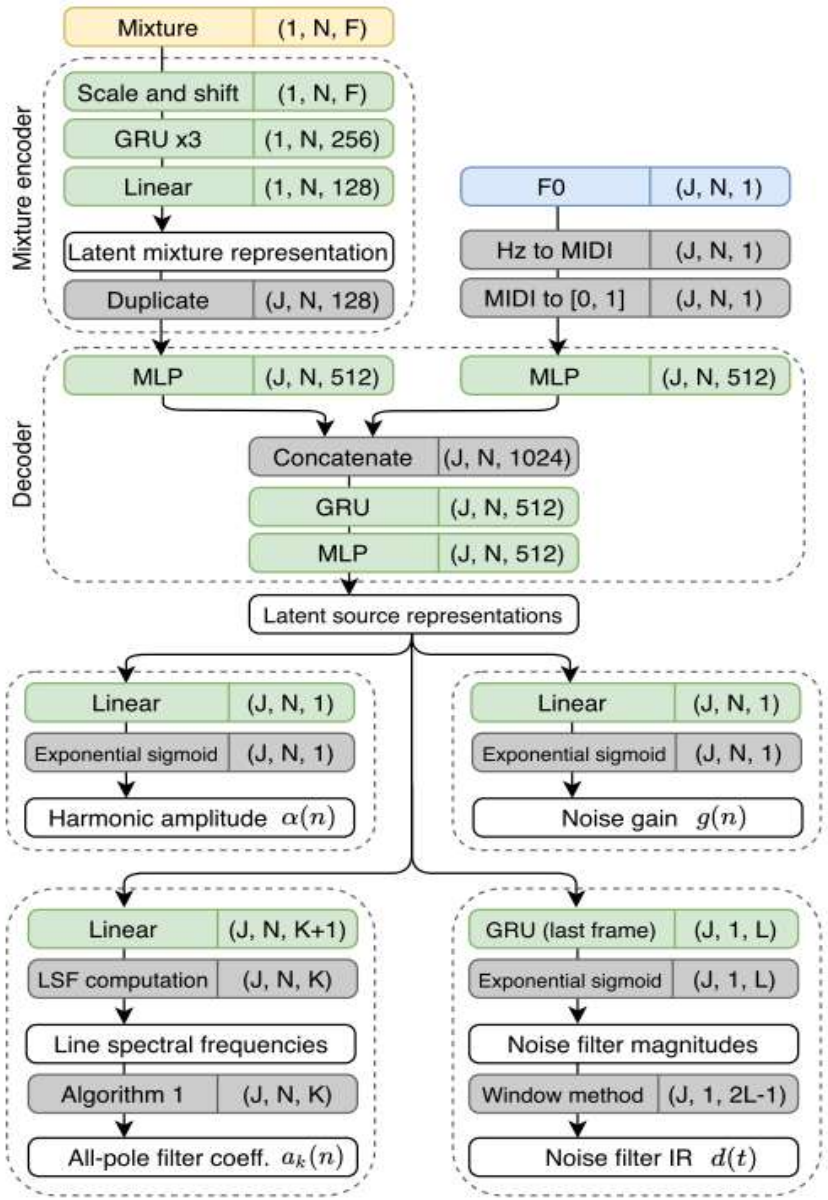
and c = [2048, 1024, 512, 256, 128, 64] indicates the FFT size used to compute the STFT. The frames overlap by 75%.

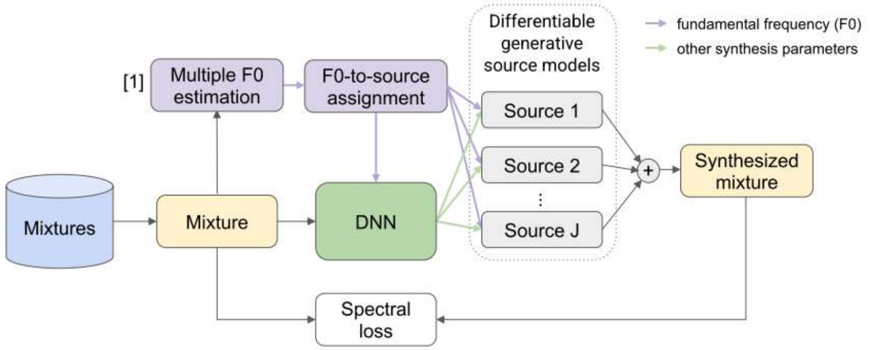


[1] H. Cuesta, B. McFee, and E. Gómez. Multiple f0 estimation in vocal ensembles using convolutional neural networks. ISMIR, 2020.

Global architecture overview

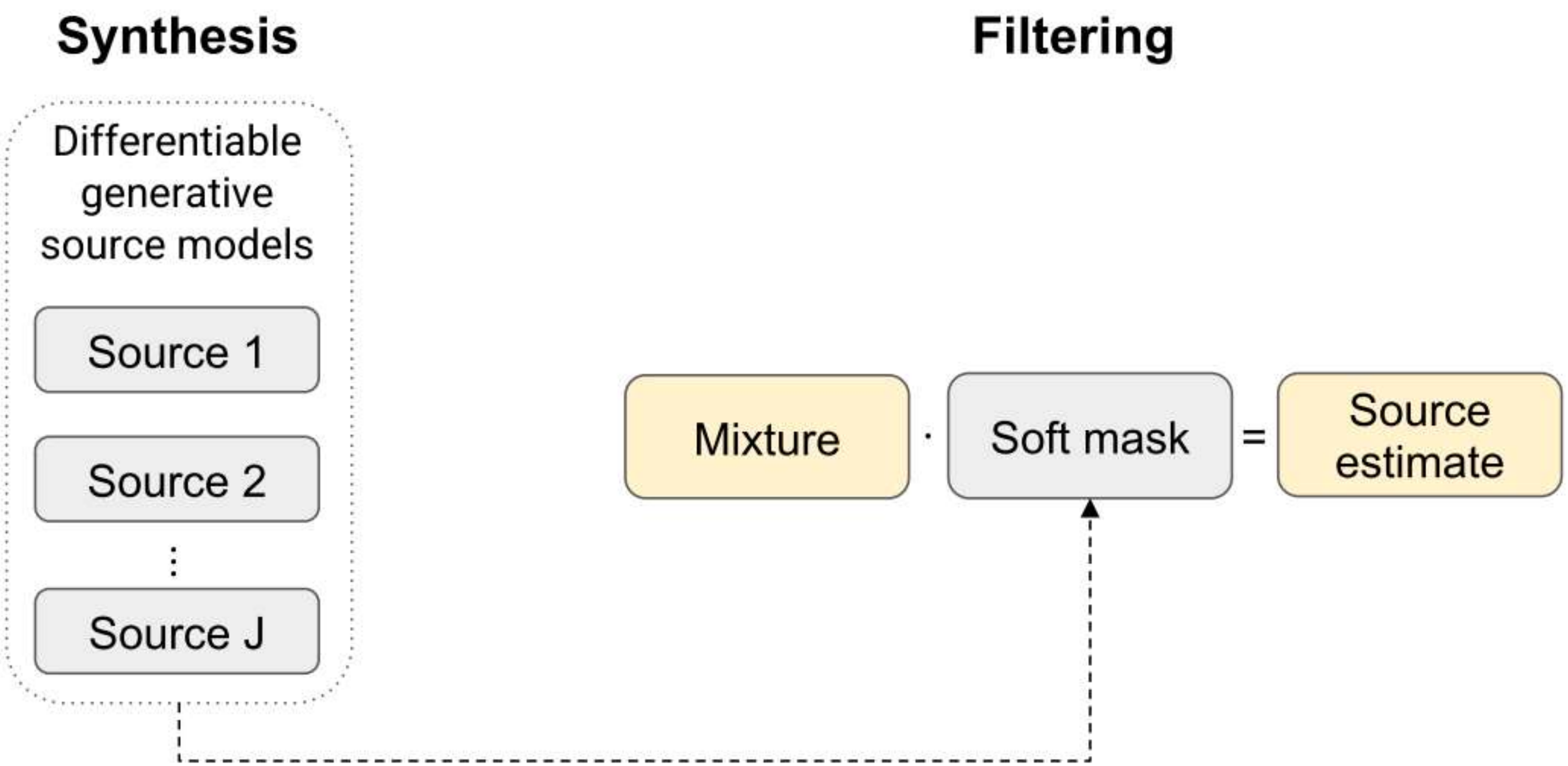
G. Richard





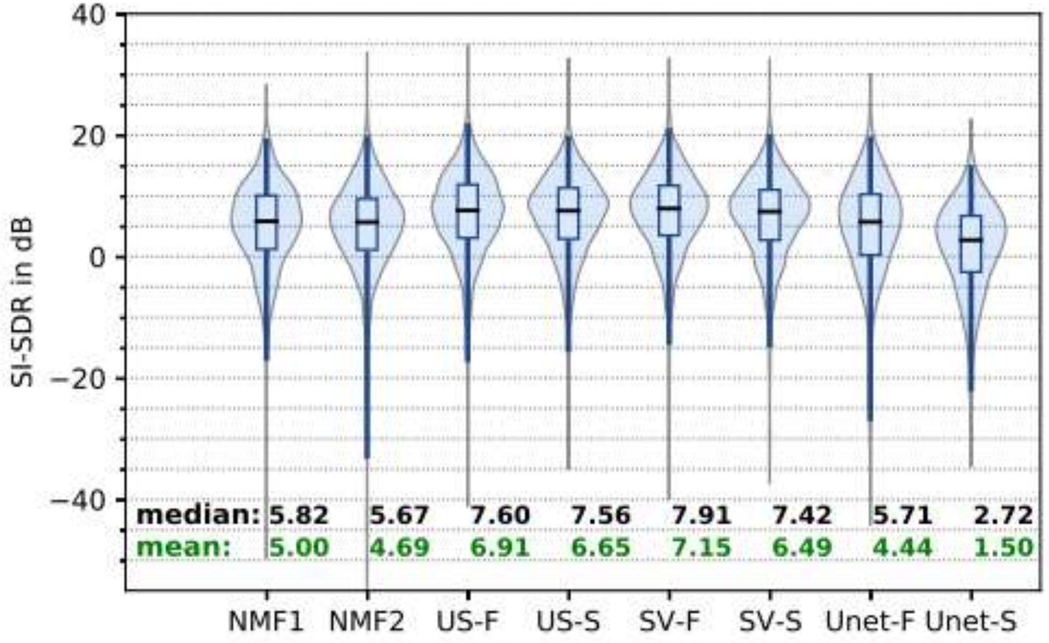
Synthesis or filtering

G. Richard



Some results

- Unsupervised (US) ≈ supervised (SU)

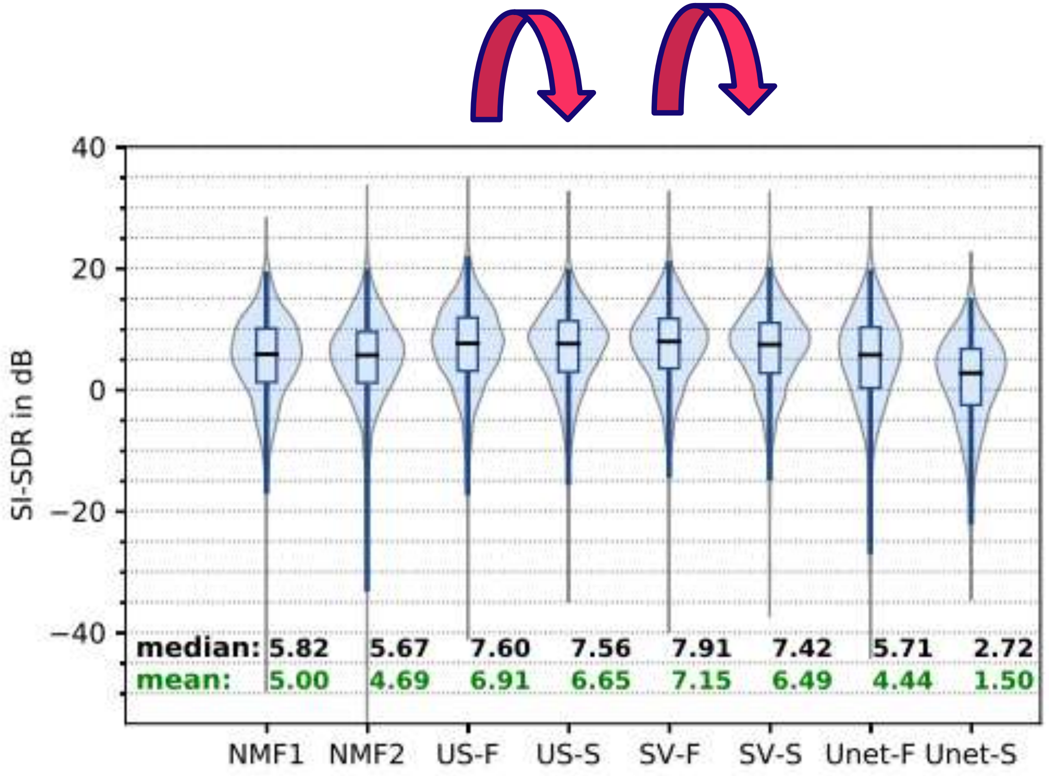


(b) $J = 4$ sources

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)

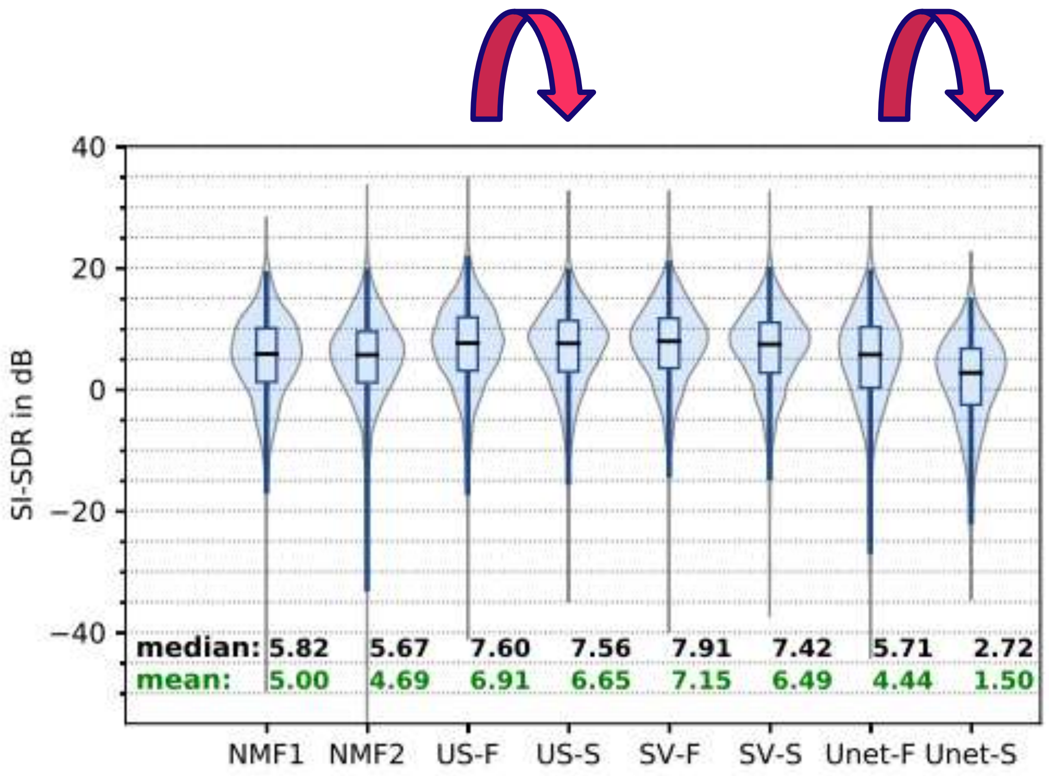


(b) $J = 4$ sources

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.

A short audio demo and some take aways

G. Richard

- **A short demo at**
  - https://schufo.github.io/umss/
  - Ou lien local
- **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

Kilian's Shulze forster web site: https://schufo.github.io/

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**