Master 2 research internship - 2025
Audio Effect Chain Parameters Estimation using Differentiable Digital Signal Processing

Feb. 2025- Aug. 2025 (6 months)

Supervisors	:	Dominique Fourer, Geoffroy Peeters and Côme Peladeau
Team / Laboratory	:	SIAM / IBISC (EA 4526) - Univ. Évry/Paris-Saclay
Collaborators	:	LTCI TelecomParis
Fundings	:	ANR AQUA-RIUS Project (https://fourer.fr/aquarius/)
Contact	:	dominique.fourer@univ-evry.fr

1 Context

Since the rise of recording during the twentieth century, audio processors have been widely used in music production [1]. Processors usually fall into two categories: synthesizers which generate a new signal, and audio effects which transform an already existing signal. The term "audio effects" coins a variety of processes: spectral (equalization), temporal (delay, reverberation), time-variant filtering (chorus, flanger, phaser), dynamic processing (compression, limiter), non-linear processing (distortion) [2], etc.

2 Goal

The work revolves around the task of blind estimation of audio effects. Audio effects take as input an audio signal $\mathbf{x}$ and a set of parameters $\mathbf{v}$ and output an audio signal $\mathbf{y}$, as shown on figure 1.

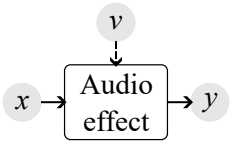

Figure 1: An audio effect transforms the input signal $\mathbf{x}$ into $\mathbf{y}$ according to some parameters $\mathbf{v}$.

Now, we aim to create a system which takes as input only the transformed signal $\mathbf{y}$ and computes an estimation of the parameters $\mathbf{v}$. Traditional methods based on deep learning rely on supervised learning with data pairs $\{\mathbf{v}, \mathbf{y}\}$. The considered neural network f_θ is trained to minimize a parameter-based metric between the estimated parameters $\hat{\mathbf{v}} = f_\theta(\mathbf{y})$ and the ground truth parameters $\mathbf{v}$. This metric can be for instance the mean-squared error $\|\hat{\mathbf{v}} - \mathbf{v}\|_2^2$. This approach is illustrated in figure 2.

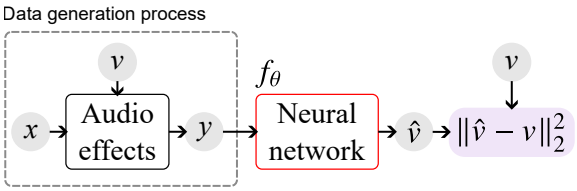

Figure 2: Traditional approach to train a neural network to perform a blind estimation of audio effects.

Recent approaches propose to implement audio processors as differentiable units g. This is called *differentiable digital signal processing*, or DDSP as coined by Engel et al. [3]. These differentiable processors allow the to compute the gradients of outputs w.r.t. the inputs signals and parameters, thus they can be inserted in deep learning. For example, in [4], we proposed to use *differentiable audio effects* (DDAFx) in our framework. We

trained our neural network so that the output of DDAFx matches the target sound by minimizing an audio metric m between the output of those DDAFx $\hat{\mathbf{y}} = g(\mathbf{x}, \hat{\mathbf{v}})$ and the ground truth signal $\mathbf{y}$:

$$\mathcal{L}(\mathbf{x}, \mathbf{y}, \theta) = m(g(\mathbf{x}, f_\theta(\mathbf{y})), \mathbf{x}) \tag{1}$$
$$= m(g(\mathbf{x}, \hat{\mathbf{v}}), \mathbf{x}) \tag{2}$$

These approaches illustrated in figure 3 have two advantages:

1. The ground truth parameters are no longer needed to train the network [5]
2. This allows the neural network to have better performance in terms of audio matching

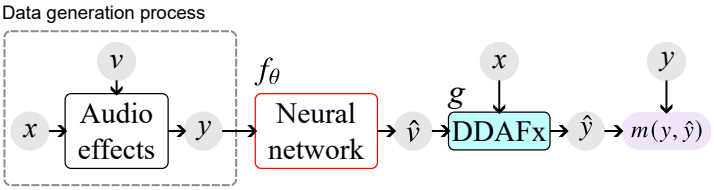

Figure 3: Approach proposed in [4] to train the neural network. The block DDAFx designates *differentiable audio effects*. DDAFx are not trainable, but they allow the gradients to "pass through" so that the neural network can be trained. See that now, only $\{\mathbf{x}, \mathbf{y}\}$ pairs are needed to train the network. The data generation process can now be hidden from us.

In this approach, the audio effects chain g has to be fixed. However, it could be beneficial to let the network predict an appropriate effects chain before estimating its parameters.
Some works try to tackle this issue but rely on labeled data: we want our approach to rely on differentiable audio effects and to use only $\mathbf{x}$ and $\mathbf{y}$. This could be done using a winner-takes-all training scheme [6], and would require to design and train an appropriate classifier. DDAFx typically suffer from high training costs [7], so attention could be brought on methods to keep the computation times low.

3 Required profile

- Strong knowledge in machine learning (deep learning) and signal processing (filtering, time-frequency analysis).
- Programming skills, particularly in Python and the *pytorch* library.
- High motivation, strong productivity, and a methodical approach to work.
- An interest in audio and music processing is a plus.

References

[1] Thomas Wilmering et al. "A History of Audio Effects". en. In: *Applied Sciences* 10.3 (Jan. 2020), p. 791. ISSN: 2076-3417. DOI: 10.3390/app10030791.

[2] Udo Zölzer, ed. *DAFX: Digital Audio Effects*. en. 2nd ed. Wiley, Mar. 2011. ISBN: 978-0-470-66599-2. DOI: 10.1002/9781119991298.

[3] Jesse Engel et al. "DDSP: Differentiable Digital Signal Processing". en. In: *Proc of ICLR*. Addis Ababa, Ethiopia: ICLR, Apr. 2020.

[4] Côme Peladeau and Geoffroy Peeters. "Blind Estimation of Audio Effects Using an Auto-Encoder Approach and Differentiable Digital Signal Processing". In: *Proc. of IEEE ICASSP*. Seoul, South Korea: IEEE, Apr. 2024, pp. 856–860. DOI: 10.1109/ICASSP48485.2024.10448301.

[5] Christian J. Steinmetz et al. "Automatic Multitrack Mixing with a Differentiable Mixing Console of Neural Audio Effects". In: *Proc. of IEEE ICASSP*. Toronto, Ont., Canada: IEEE, June 2021, pp. 71–75. DOI: 10.1109/ICASSP39728.2021.9414364. (Visited on 03/18/2024).

[6] Stefan Lee et al. “Stochastic Multiple Choice Learning for Training Diverse Deep Ensembles”. In: *Proc. of NeurIPS*. Vol. 29. Curran Associates, Inc., 2016.

[7] Chin-Yun Yu et al. “Differentiable All-Pole Filters for Time-Varying Audio Systems”. In: *Proc. of DAFx*. Surrey, United Kingdom: DAFx, 2024, pp. 345–352.