Audio Quality Simulation and Inversion

Oct. 2023 - Sep. 2026

keywords: Audio Signal Processing, Audio Quality, Data Augmentation, Audio Effect Inversion

Thesis director(s)	:	Hichem Maaref (PR HDR, IBISC, 40%)
Co-advisor	:	Dominique Fourer (MCF, IBISC, 60%)
Laboratory	:	IBISC - Université d'Evry / Paris-Saclay
School	:	Collège doctoral 580 - STIC - Pôle A
Funding	:	ANR AAPG22 - Project AQUA-RIUS
Contact	:	dominique.fourer@univ-evry.fr

The ANR AQUA-RIUS Project:

As instrumental sound "timbre" is defined as all sound characteristics which are not related to pitch, loudness and duration, we consider here "audio quality" as everything related to the sound characteristics which is not related to the content sources. This therefore includes the choice of microphone, recording media (tapes, vinyls, digital and related potential artifacts), audio production chain (equalization, compression, reverberation) and diffusion (such as mp3 data-reduction). The project AQUARIUS proposes an exhaustive investigation of "audio quality" to provide analysis and synthesis tools leading to a more robust and unified data representation for audio signals. As Music information Retrieval (MIR) techniques aim at developing robust prediction methods related to the audio content and, independent from audio quality, now the present project aims to provide original contributions leading to a better understanding of the invariant properties of the learned audio features [1, 2]. Thus, we aim at improving the practices of researchers involved in real-world multimedia applications using machine learning algorithms. Indeed, this research project includes fundamental research related to signal processing and efficient data representation for machine learning with a consideration to real-world application scenarios for dataset audio tagging with a possible industrial valorization. The industrial collaborators of IRCAM and the first-rate expertise of the project contributors in audio signal processing and in machine learning (including deep learning) are a definite asset to tackle this project.

The project AQUA-RIUS will address the following scientific questions

- The analysis and modeling of audio quality with a focus on the capability to predict the effects applied during the audio signal production and diffusion chain.
- The simulation and the synthesis of audio quality effects with a consideration for making more robust machine learning algorithms through data augmentation and domain adaptation techniques to deal with several training datasets.
- The full control of the audio quality in order to cancel or to reverse production and diffusion effects.

Related work:

The litterature already reported a large number of audio effects which can be formally described as proposed in [34]. Gorlow and Reiss [12] show that complicated non-linear effects such as dynamic range compression can efficiently be reverted using a suitable mathematical model when its parameters (only 7 scalars) are known. Such audio reverse engineering techniques pave the way for a large number of applications for audio signal restoration, music remixing [8, 11] and the study of studio and DJ mixing practices [26], which we recently opened up as a new field of research [24, 31, 29]. More recent work now proposes to remove complicated non-linear effects such as distortion using deep learning methods [14]. This shows the relevance of deep neural networks in such tasks regarding the objectives of this project.

There, we proposed a new set of methods and a dedicated research dataset [26] leading to promising results for recovering the cue points and the effects parameters from an artistic DJ mix when the isolated music tracks are known. Our approach also enables to reverse the mixing process and to accurately estimate the signal transformations which were applied during the mixing process (which affect the resulting audio quality) since they are correctly identified.

Thesis goals:

Our objective is to revisit traditional approaches [3, 18] to restore audio signals and to enhance audio quality using recent deep learning methods such as conditioned-U-net [19]. Moreover, we will propose new model-based effect inversion or cancellation methods [12]. To this end, we aim at reversing the mixing process and/or targeted effects in real-world diffusion contexts. Hence, the project AQUA-RIUS will extend the promising work based on reverse engineering of studio and DJ mixes [26] by considering more complicated configurations such as those involving non-linear effects, unknown or partially-known tracks used during the mix session, and non-constant time-scaling effects.

I) Data Augmentation

In machine learning applications, the number and the diversity of the training examples is essential for the generalization and the robustness of the processing. Unfortunately, the constitution of a training dataset is a difficult and tedious task. The initial role of data augmentation [30] is to apply transformations on the training examples to artificially augment the size of the dataset and to improve the performances of processing, see [5, 32]. This approach has been successfully used in a large number of audio applications such as [23, 6]. It is clearly shown that applying the suitable chain of audio transformations and degradations on the training dataset can significantly improve the results when the tested examples are themselves transformed, as shown in [20]. The aim of this task is to extend this work based on data augmentation when used for audio applications. For this research, we will make use of the direct simulation toolbox for audio transformations, degradations and mixing.

Especially, the study of the invariance properties can be made in this context since most of the MIR applications aim at being robust to a large number of audio effects related to audio quality (e.g. artist or album effect [15]). The invariance is related to the ability of the features and/or of the estimated values to be weakly affected by a given transformation. For audio processing, it is interesting to see how the audio quality can be changed to achieve better and more robust results, and so, which transformations can be used or not. For example, as noted in [20] and contrarily to what we could expect, applying transformations for which the predicted label seems to be invariant, can improve the robustness.

Moreover, data augmentation is an interesting track to reduce the risk of overfitting or Horse Effect, see e.g. [27, 28, 17]. For some training datasets, of classification e.g., the predicted labels can be correlated to an unwanted property of the annotated examples, which may provide illusory good results when the test examples have the same correlations. This problem is often met in a cross dataset scenario, or with low quality signals. In this case, the use of data augmentation diversifies the sound properties of the training dataset and makes possible its decorrelation with the labels to improve the robustness and the performance in a real-world case. But, the transformations used must be carefully selected in a way that the sound properties which truly characterize the labels remain unchanged.

II) Generative Models

Our objective is to explore the simulation and the inversion of audio quality using generative models based on deep neural networks.

Among the proposed approaches, we propose to investigate variational autoencoders (VAE) [22], generative adversarial networks (GAN) [10] and its adaptation CycleGAN [33] which are intensively used in a large variety of computer vision applications. Here, we believe that they can also be used to simulate realistic signals with a targeted audio quality while obtaining a latent representation allowing a possible control of the audio quality and an inversion.

The literature proposes several studies based on generative deep learning architecture for generating music from a constrained latent representation [13]. Such approaches could be adapted to take in consideration audio quality to simulate for example studio or artistic effects which are of interest. Disentangling the

salient information conveyed by VAEs can then be obtained using for example an adversarial loss function as previously proposed by Fader networks applied to computer vision [16].

Another idea is the use of generative model as proposed in [4] for inverse poisoning which has shown its efficiency for improving the robustness and the efficiency of trained classical deep neural network architectures using the same evaluation dataset. Thus, this task will propose to adapt such approaches to existing audio classification methods. Moreover, simulating audio quality is of interest for music production as a creative effect to convey a desired atmosphere (e.g. *vintage, warm, hollow, round*, or *distant*).

III) Signal Restoration, Music Unmixing and Re-mixing

In this task, we propose to study in detail the inversion of audio effects and sound degradations. The aim is to restore a signal with its original audio quality before transformation. This topic has many interests such as the removal of defects and noise [3] of altered or damaged audio files. This is also of interest for canceling of a specific audio effect to go back to the original and raw version [8, 11].

First, we propose a classical approach based the inversion or the cancellation of an effect through mathematical modeling. Second, we will address this problem through the Deep Learning approaches which can be used to approximate [7] and/or reverse an arbitrary effect operator using for example a specific neural architecture [9].

We also propose to extend our previous research on reverse engineering of studio and DJ mixes [26], by considering more complicated mix configurations such as:

(1) involving non-linear effects,
(2) unknown or partially-known tracks used during the mix session,
(3) non-constant time-scaling: here, an iterative analysis-by-synthesis approach can allow to approach the applied time-varying speed changes very precisely, leading to better performance of the methods for reverse DJ mixing.

This is also useful for the dataset generation task of the project, since the unmixed tracks isolated from a DJ mix can include diffusion transformations (EQ, delay, distortion effects) which can be useful for a comparison with the original tracks [25].

As a result of the previous simulation and the inversion tasks, one interesting application is the re-mixing of a recording. This task is dedicated to the sound transformation of a musical piece to change its mixing style. For example, given a recent music release mixed in a modern style, applying an inversion of the effects we will be able to obtain a neutral version which can be then re-mixed to a different style, such as in the 60's fashion.

The automatic creation of a re-mixing application has a great benefit for data augmentation, for many recognition tasks which are insensitive to the audio quality (e.g. automatic transcription, instrument recognition, rhythm analyses) [21].

Required profile

- good machine learning and signal processing knowledge
- mathematical understanding of the formal background
- excellent programming skills (Python, Matlab, C++)
- good motivation, high productivity and methodical works
- an interest for audio processing, AI and new technologies

References

[1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[2] Alberto Bietti and Julien Mairal. Group invariance, stability to deformations, and complexity of deep convolutional representations. *arXiv preprint arXiv:1706.03078*, 2017.

[3] Olivier Cappé. *Techniques de réduction de bruit pour la restauration d'enregistrements musicaux*. PhD thesis, Paris, ENST, 1993.

[4] Adrien Chan-Hon-Tong. Symmetric adversarial poisoning against deep learning. In *IPTA 2020*, Paris, France, November 2020.

[5] E. I. Chang and R. P. Lippmann. Using voice transformations to create additional training talkers for word spotting. In *Advances in Neural Information Processing Systems (NIPS)*, pages 875–882, 1995.

[6] Alice Cohen-Hadria, Axel Roebel, and Geoffroy Peeters. Improving singing voice separation using Deep U-Net and Wave-U-Net with data augmentation. In *Proc. EUSIPCO 2019*, September 2019.

[7] Balázs Csanád Csáji et al. Approximation with artificial neural networks. *Faculty of Sciences, Etvs Lornd University, Hungary*, 24(48):7, 2001.

[8] J. Reiss D. Barchiesi. Reverse engineering of a mix. *Journal of the Audio Engineering Society*, 58:563–576, 2010.

[9] Aidan N Gomez, Mengye Ren, Raquel Urtasun, and Roger B Grosse. The reversible residual network: Backpropagation without storing activations. *arXiv preprint arXiv:1707.04585*, 2017.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[11] S. Gorlow and S. Marchand. Reverse engineering stereo music recordings pursuing an informed two-stage approach. In *Proc. Digital Audio Effects Conf. (DAFx'13)*, 2013.

[12] S. Gorlow and J. D. Reiss. Model-based inversion of dynamic range compression. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(7):1434–1444, July 2013.

[13] Jay A Hennig, Akash Umakantha, and Ryan C Williamson. A classifying variational autoencoder with application to polyphonic music generation. *arXiv preprint arXiv:1711.07050*, 2017.

[14] Johannes Imort, Giorgio Fabbro, Marco A Martínez Ramírez, Stefan Uhlich, Yuichiro Koyama, and Yuki Mitsufuji. Removing distortion effects in music using deep neural networks. *arXiv preprint arXiv:2202.01664*, 2022.

[15] Y. E Kim, D. S Williamson, and S. Pilli. Towards quantifying the album effect in artist identification. In *International Society on Music Information Retrieval Conference (ISMIR)*, 2006.

[16] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. Fader networks: Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems*, pages 5967–5976, 2017.

[17] S. Lapuschkin, S. Waldchen, A. Binder, G. Montavon, W. Sämek, and K.R. Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1–8, 2019.

[18] N. López, Y. Grenier, G. Richard, and I. Bourmeyster. Single channel reverberation suppression based on sparse linear prediction. In *Proc. IEEE ICASSP 2014*, pages 5182–5186, 2014.

[19] G Meseguer-Brocal and G Peeters. Conditioned-U-Net: Introducing a control mechanism in the U-Net for multiple source separations. In *International Society on Music Information Retrieval Conference (ISMIR)*, 2019.

[20] Rémi Mignot and Geoffroy Peeters. An analysis of the effect of data augmentation methods: Experiments for a musical genre classification task. *Transactions of the International Society for Music Information Retrieval*, 2(1):97–110, 2019.

[21] Enrique Perez-Gonzalez and Joshua D Reiss. Automatic mixing. *DAFX: Digital Audio Effects,*, pages 523–549, 2011.

[22] Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. Variational autoencoder for deep learning of images, labels and captions. In *Advances in neural information processing systems*, pages 2352–2360, 2016.

[23] J. Schlüter and T. Grill. Exploring data augmentation for improved singing voice detection with neural networks. In *International Society on Music Information Retrieval Conference (ISMIR)*, pages 121–126, 2015.

[24] Diemo Schwarz and Dominique Fourer. Towards extraction of ground truth data from DJ mixes. In *International Society on Music Information Retrieval Conference (ISMIR)*, Suzhou, China, oct 2017.

[25] Diemo Schwarz and Dominique Fourer. A dataset for DJ-mix reverse engineering. In *International Symposium on Music Information Retrieval (ISMIR), late breaking demos*, Paris, France, September 2018.

[26] Diemo Schwarz and Dominique Fourer. Methods and Datasets for DJ-Mix Reverse Engineering. In Kronland-Martinet R., Ystad S., and Aramaki M., editors, *Perception, Representations, Image, Sound, Music*, volume 12631 of *Lecture Notes in Computer Science (LNCS)*, pages 31–47. Springer, Cham, March 2021. extended version of CMMR 2019 conference article hal-02172427.

[27] B. L. Sturm. A simple method to determine if a music information retrieval system is a horse. *IEEE Transactions on Multimedia*, 16(6):1636–1644, 2014.

[28] Bob Sturm. The horse inside: Seeking causes behind the behaviours of music content analysis systems. *CoRR*, abs/1606.03044, 2016.

[29] K. Taejun, C. Minsuk, S. Evan, Y. Yi-Hsuan, and N. Juhan. A computational analysis of real-world DJ mixes using mix-to-track subsequence alignment. In *International Society on Music Information Retrieval Conference (ISMIR)*, 2020.

[30] D. A Van Dyk and X.-L Meng. The art of data augmentation. *Journal of Comp. and Graph. Statistics*, 10(1), 2001.

[31] Lorin Werthen-Brabants. Ground truth extraction & transition analysis of DJ mixes. Master's thesis, Ghent University, Belgium, 2018.

[32] S. C Wong, A. Gatt, V. Stamatescu, and M. D McDonnell. Understanding data augmentation for classification: when to warp? In *Proc. IEEE int. conf. on digital image computing: techniques and applications (DICTA)*, pages 1–6, 2016.

[33] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[34] Udo Zölzer. *Digital Audio Signal Processing*. John Wiley & Sons, 2005.