electronics

MDPI

Editorial

Machine Learning Applied to Music/Audio Signal Processing

Alexander Lerch [1,*] and Peter Knees [2]

[1] Center for Music Technology, Georgia Institute of Technology, Atlanta, GA 30332, USA
[2] Faculty of Informatics, Institute of Information Systems Engineering, TU Wien Informatics, Favoritenstraße 9-11, 1040 Vienna, Austria; peter.knees@tuwien.ac.at
* Correspondence: alexander.lerch@gatech.edu

Over the past two decades, the utilization of machine learning in audio and music signal processing has dramatically increased. Nowadays, novel approaches in the fields of Music Information Retrieval and Audio Signal Processing are largely dominated by machine learning solutions. Due to the unique challenges posed by (musical) audio signals, including the superposition of many sources overlapping in both time and frequency and the inherent semantic and hierarchical structure, the application of generic, domain-agnostic machine learning models is often not successful without modifications.

There is a wide range of tasks to be solved in audio signal analysis and processing, the majority of which require specifically adapted machine learning approaches. In this Special Issue, we have a fair subset of such tasks represented. Two papers in this collection address detecting the presence of the singing voice in musical audio. Zhang et al. [1] propose the use of a long-term recurrent convolutional network for this task to combine powerful feature extraction with contextual time sequence modeling. The authors show that the system is able to outperform existing state-of-the-art singing voice detection systems. Krause et al. [2] also present results on this task. They point out that nearly all state-of-the-art systems for singing voice detection are trained on popular music and investigate the performance of singing voice detection systems on opera recordings. In this scenario, they can show that a modern CNN-based approach does not significantly outperform a traditional, feature-based classification approach.

A prominent task in Music Information Retrieval is the transcription of pitches from the musical audio; while largely solved for monophonic audio signals, many challenges in polyphonic pitch transcription have yet to be solved. Gao et al. [3] focus on transcribing the singing voice and propose a two-stage system that first separates the vocals from the mix with a so-called high-resolution network and then transcribes the fundamental frequencies with an encoder/decoder network. A system for the transcription of bass in Jazz recordings is presented by Abeßer and Müller [4]. They investigate several architectural choices of a U-net deep neural network architecture and outperform existing methods for bass transcription with careful adapting parameters and training methodology. For the task of multi-pitch estimation, the transcription of all fundamental frequencies in a music recording, Taenzer et al. [5] show that an estimate of polyphony, i.e., the number of concurrently playing voices, can refine the predictions of a multi-pitch estimator. Hernandez-Olivan et al. [6] also work on music transcription and not only quantify the impact of timbre and onset envelope on the transcription accuracy but also present a model with improved performance by taking into account this extra information. Drum transcription focuses, as opposed to the transcription systems presented above, on the detection of percussive events in music. Vande Veire et al. [7] show that the results of a drum transcription system based on non-negative matrix factor deconvolution can be improved by forcing the estimated activation function to be "more binary" by introducing an extra regularization term and applying a sigmoid function to the activations. Tempo and beat detection has been an active topic of research in the Music Information Retrieval community for a long time. Pinto et al. [8] argue that current systems tend to be trained

Citation: Lerch, A.; Knees, P. Machine Learning Applied to Music/Audio Signal Processing. *Electronics* **2021**, *10*, 3077. https://doi.org/10.3390/electronics10243077

Received: 25 November 2021
Accepted: 26 November 2021
Published: 10 December 2021

on specific genres or training data that does not always reflect the user needs; they propose a system that enables a user to perform targeted fine-tuning of a state-of-the-art deep neural network based on a very limited temporal region of annotated beat locations. Carsault et al. [9] explore the estimation, prediction, and evaluation of chords in musical audio focused on the special use case of a real-time system for musical co-creativity.

There exist a wide variety of audio processing tasks beyond transcription. One area of interest is enhancing or restoring the quality of a low-quality audio signal. Lattner and Nistal [10] point out that the majority of current approaches for this task focus on non-music signals and address this white space with a Generative Adversarial Network (GAN) architecture for the task of enhancing music encoded at low bitrates. They show that stochastic generators seem to perform better than deterministic generators in increasing the audio quality.

Data-driven approaches, and deep learning approaches, in particular, are known for the need for a considerable amount of data to allow proper training. These training data, however, are not easily available for many audio and music tasks, as the manual annotation is often a time-consuming and tedious endeavor. Two papers in this Special Issue approach this problem from different angles. Venkatesh et al. [11] show how synthetically generated training data can be useful for the training of systems for the classification of music vs. speech in broadcast streams. Grollmisch and Cano [12] explore how semi-supervised learning, in particular the FixMatch algorithm, can improve the performance on three audio classification tasks, namely music, industrial sounds, and acoustic scenes.

Although neural networks have shown superior performance over traditional approaches in a wide variety of tasks, there is an inherent problem with the interpretability of such systems; it is hard to explain reasons for individual decisions the systems makes and the impact of individual underlying factors on the result can often not be determined. Zinemanas et al. [13] address this issue by presenting a system for audio classification that explains its predictions based on the similarity of the input to a set of learned prototypes in a latent space. Krug et al. [14] introduce a new technique for visualizing and understanding neural networks for audio speech recognition, in which neuron activation profiles can visualize acoustic concepts learned by the model.

Last but not least, Zeng and Lau [15] address the harmonization of melodies in the symbolic score domain (as opposed to audio). They apply a reinforcement learning approach to the problem that learns a structured representation of the input melody to leverage phrase information for the harmonization task.

To summarize, the contributions in this Special Issue on machine learning for audio and music cover a wide range of topics that mirror the variety of tasks and approaches in the larger field. As such, it presents an excellent snapshot of current tasks and challenges and introduces novel solutions to some of the key problems in the field. We, the editors, thank all the authors for their valued contributions that make this a truly "special" issue of Electronics.

Author Contributions: Both authors contributed equally. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, X.; Yu, Y.; Gao, Y.; Chen, X.; Li, W. Research on Singing Voice Detection Based on a Long-Term Recurrent Convolutional Network with Vocal Separation and Temporal Smoothing. *Electronics* **2020**, *9*, 1458. [CrossRef]
2. Krause, M.; Müller, M.; Weiß, C. Singing Voice Detection in Opera Recordings: A Case Study on Robustness and Generalization. *Electronics* **2021**, *10*, 1214. [CrossRef]
3. Gao, Y.; Zhang, X.; Li, W. Vocal Melody Extraction via HRNet-Based Singing Voice Separation and Encoder-Decoder-Based F0 Estimation. *Electronics* **2021**, *10*, 298. [CrossRef]
4. Abeßer, J.; Müller, M. Jazz Bass Transcription Using a U-Net Architecture. *Electronics* **2021**, *10*, 670. [CrossRef]

5. Taenzer, M.; Mimilakis, S.I.; Abeßer, J. Informing Piano Multi-Pitch Estimation with Inferred Local Polyphony Based on Convolutional Neural Networks. *Electronics* **2021**, *10*, 851. [CrossRef]
6. Hernandez-Olivan, C.; Zay Pinilla, I.; Hernandez-Lopez, C.; Beltran, J.R. A Comparison of Deep Learning Methods for Timbre Analysis in Polyphonic Automatic Music Transcription. *Electronics* **2021**, *10*, 810. [CrossRef]
7. Vande Veire, L.; De Boom, C.; De Bie, T. Sigmoidal NMFD: Convolutional NMF with Saturating Activations for Drum Mixture Decomposition. *Electronics* **2021**, *10*, 284. [CrossRef]
8. Pinto, A.S.; Böck, S.; Cardoso, J.S.; Davies, M.E.P. User-Driven Fine-Tuning for Beat Tracking. *Electronics* **2021**, *10*, 1518. [CrossRef]
9. Carsault, T.; Nika, J.; Esling, P.; Assayag, G. Combining Real-Time Extraction and Prediction of Musical Chord Progressions for Creative Applications. *Electronics* **2021**, *10*, 2634. [CrossRef]
10. Lattner, S.; Nistal, J. Stochastic Restoration of Heavily Compressed Musical Audio Using Generative Adversarial Networks. *Electronics* **2021**, *10*, 1349. [CrossRef]
11. Venkatesh, S.; Moffat, D.; Miranda, E.R. Investigating the Effects of Training Set Synthesis for Audio Segmentation of Radio Broadcast. *Electronics* **2021**, *10*, 827. [CrossRef]
12. Grollmisch, S.; Cano, E. Improving Semi-Supervised Learning for Audio Classification with FixMatch. *Electronics* **2021**, *10*, 1807. [CrossRef]
13. Zinemanas, P.; Rocamora, M.; Miron, M.; Font, F.; Serra, X. An Interpretable Deep Learning Model for Automatic Sound Classification. *Electronics* **2021**, *10*, 850. [CrossRef]
14. Krug, A.; Ebrahimzadeh, M.; Alemann, J.; Johannsmeier, J.; Stober, S. Analyzing and Visualizing Deep Neural Networks for Speech Recognition with Saliency-Adjusted Neuron Activation Profiles. *Electronics* **2021**, *10*, 1350. [CrossRef]
15. Zeng, T.; Lau, F.C.M. Automatic Melody Harmonization via Reinforcement Learning by Exploring Structured Representations for Melody Sequences. *Electronics* **2021**, *10*, 2469. [CrossRef]