applied sciences

MDPI

Editorial

Special Issue on Deep Learning for Applications in Acoustics: Modeling, Synthesis, and Listening

Leonardo Gabrielli [1,*], György Fazekas [2] and Juhan Nam [3,*]

[1] Department of Information Engineering, Università Politecnica delle Marche, 60131 Ancona, Italy
[2] Center for Digital Music, Queen Mary University, London E1 4FZ, UK; g.fazekas@qmul.ac.uk
[3] Korea Advanced Institute of Science and Technology, Daejeon 34141, Korea
* Correspondence: l.gabrielli@staff.univpm.it (L.G.); juhan.nam@kaist.ac.kr (J.N.)

Keywords: deep learning; machine listening; audio processing

1. Introduction

The recent introduction of Deep Learning has led to a vast array of breakthroughs in many fields of science and engineering. The data-driven approach has gathered the attention of research communities and has often been successful in yielding solutions to very complex classification and regression problems.

In the fields of audio analysis, processing and acoustic modelling, Deep Learning has been swiftly adopted, initially borrowing their methods from the image processing and computer vision field, and then finding creative and innovative solutions to suit domain-specific needs of acoustic research. 2D convolutional operators and adversarial approaches are two well-known developments of deep image processing that have been borrowed and adapted for audio processing. However, the nature of acoustic signals poses challenges in terms of representation that require novel solutions. The anisotropic nature of the time-frequency representation of audio signals with the Short-Time Fourier Transform, the multi-scaled nature of musical events and the effect of psychoacoustics prompt researchers to adopt new approaches and architectures.

Another issue introduced by the statistical nature of Deep Learning and the large datasets involved, with respect to traditional Digital Signal Processing, is the reproducibility of results and the understanding of the learned models. These highlight the importance of maintaining rigorous research practices and put more emphasis on the advancement of knowledge rather than the empirical problem-solving that Deep Learning research has initially focused on and succeeded at.

These issues informed the process of writing our Call for Papers and we are glad to introduce the papers accepted for publication in this Special Issue. In the editing process we have received many contributions of high value. Some of these have not been deemed compatible with the topics of our call for papers and have been published elsewhere, while only a few of the received papers haven't met the quality criteria.

2. Deep Learning for Applications in Acoustics

The topics covered by the published papers cover sound synthesis, generative music, spatial audio, bioacoustics, audio scene classification and more. We would like to start our survery by highlighting those topics that are rarely addressed by the audio and acoustics communities, but pose new interesting challenges.

The paper from Narváez and Percybrooks [1] addresses the need for generating heart sound datasets for training neural networks. The goal is to expand the currently limited datasets with synthesized sounds that can be classified by neural networks as normal, as opposed to abnormal. In this work, Generative Adversarial Networks (GANs) are employed to generate heart sounds and the Empirical Wavelet Transform (EWT) is

Citation: Gabrielli, L.; Fazekas, G.; Nam, J. Special Issue on Deep Learning for Applications in Acoustics: Modeling, Synthesis, and Listening. *Appl. Sci.* **2021**, *11*, 473. https://doi.org/10.3390/app11020473

Received: 2 December 2020
Accepted: 16 December 2020
Published: 6 January 2021

employed to clean the generated waveforms from noise, thus reducing the training time. A quantitative assessment is provided using a mel-cepstral distortion metric, and the results are compared to another method present in the literature with a positive outcome. Additionally, several heart sound classification methods have been compared with the generated dataset.

The paper from Merchan et al. [2] deals with the classification of vocalization spectrograms of Antillean Manatee, a large endangered sea mammal. In order to estimate its population for scientific purposes, the authors turn to audio classification. Their previous studies successfully managed to identify and count these mammals in the area of Panama by analyzing audio recordings with traditional DSP techniques. In this work, they aim to improve the robustness of the classification by using Deep Learning methods. They explore the use of different types of Convolutional Network architectures and different input features. The authors also report a set of experiments aimed at understanding the role of the dataset preparation, the importance of cross-validation, discussing the trained features of the network and comparing to a method based on traditional FFT processing.

We turn now our attention to more traditional audio applications, such as Acoustic Scene Classification and filter design. A review of Deep Learning techniques for Acoustic Scene Classification is provided by Abeßer [3]. This field has gathered attention in the last years, and of particular importance is the annual Detection and Classification of Acoustic Scenese and Events (DCASE) competition. This timely review by Abeßer outlines about a hundred papers in the field, providing a perspective of current research trends.

In Pepe et al. [4], multichannel equalization is investigated with a novel approach based on nonlinear optimization, were Neural Networks are employed to generate the FIR coefficients that minimize a target function based on spectral flatness. The results exceed traditional methods including Frequency Deconvolution, obtaining extremely flat frequency responses at multiple positions in a car environment with several loudspeakers.

Spatial audio is covered by two papers with interesting results. In [5], Zieliński and colleagues compare the performance of humans and machines in detecting a spatial audio scene. Specifically, the task is to assess the position of a music ensemble that is located in front of the listener, to the back or around her/him. Human subjects undertook the test remotely over the Internet, therefore with unknown listening devices. Their accuracy is relatively low. Several Deep Learning algorithms are tested, showing very high classification results under known conditions, however scoring similarly to humans when the electro-acoustic conditions are unknown to the machine. This opens new interesting challenges for Machine Listening models.

In [6], Tsui et al. investigate the restoration of high-frequency distortions in spherical harmonic interpolation using neural network models. Spherical harmonic interpolation is an effective method to upsample a sparse head-related transfer function (HRTF) measurement to a dense set of HRTFs. Depending on the number of sparse measurement points and spherical harmonic order, however, it can lose high-frequency information, which may cause timbre difference and weaken localization. They attempt to remedy the issue using a combination of convolutional auto-encoder (CAE) and denoising auto-encoder (DAE). The evaluation based on perceptual difference and localization models shows the approach to be promising.

Several papers in this special issue deal with music-related tasks such as the generation of bass scores [7], chord progressions [8], and the analysis of rhythmic patterns [9].

Grachten et al. [7] present a variational gated auto-encoder model for conditional generation of bass notes. They focus on temporally stable and controllable sequence generation, which is critical in human-collaborative music creation. The temporal stability is secured by the novel use of cross-reconstruction which makes cross predictions between two outputs and two inputs in the nearby position. The controllability is provided by allowing users to take the audio track as input and explore variations from the reference on two dimensional conditional latent space. They also show that the latent space is

learned to generate semantically disentangled bass patterns that require a relative pitch representation, and sensitivity to harmony, instrument timbre, and rhythm.

In Navarro-Caceres et al. [8] the authors propose an assistive music composition system that is able to satisfy multiple criteria and find multiple candidates in chord sequence generation using an Artificial Immune System, a biologically inspired AI technique based on Genetic Programming for multi-objective optimisation. The generated chord progressions fit a desired tonal tension profile, taking for instance consonance or the melodic attraction of successive chords into account, thus aiding the composition process. The usability of the system as an assistive tool was demonstrated in a small subjective user study.

Similarly to harmony in music, rhythmic patters can also be considered in a hierarchy. Pesek et al. in [9] puts forward a layered *compositional* hierarchical model or rhythm, where each successive layer is a composition of one of more parts of the previous layer. The relative spacing and scaling of part relations are modelled by Gaussians, leading to a tempo-invariant rhythmic pattern representation. An unsupervised learning algorithm is proposed for parameter estimation and constructing the model using a greedy optimisation approach. The utilities of the model are demonstrated in rhythm pattern extraction using the Ballroom dataset commonly used in music information retrieval evaluations such as MIREX, and in pieces with varying tempo and time signature. Its robustness to changes make the model applicable to the analysis of a broad range of music genres.

Finally, one paper deals with speech processing, more specifically, with the conversion of the timbre of a source speech into that of a target speaker [10]. Xiaokong et al. propose a method based on bidirectional long short-term memory (BLSTM) neural networks for conversion of vocal tracks in presence of noise. A novel architecture with statistical filtering and sub-band cepstrum conversion and fusion is introduced reducing the impact of noise on the input speech. Additionally, a cepstrum filter is proposed to further improve the quality of the converted voice.

3. Summary

The topics targeted by this Special Issue are wide-ranging but one trend that is evident, is the need for advancements in the neural network architectures and the handling of training data, to put research goals forward. It is quite evident that the adoption of available techniques, borrowed from the image processing field, is not sufficient for the audio community and a renovated attention must be put in the analysis of the networks with ablation studies and in the conceiving of new type of layers and mathematical operators. To conclude, we would like to acknowledge the efforts of the authors, and express our appreciation to those who publicly shared samples and data from their experiments.

Funding: This research received no external funding.

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

References

1. Narváez, P.; Percybrooks, W.S. Synthesis of Normal Heart Sounds Using Generative Adversarial Networks and Empirical Wavelet Transform. *Appl. Sci.* **2020**, *10*, 7003. [CrossRef]
2. Merchan, F.; Guerra, A.; Poveda, H.; Guzmán, H.M.; Sanchez-Galan, J.E. Bioacoustic Classification of Antillean Manatee Vocalization Spectrograms Using Deep Convolutional Neural Networks. *Appl. Sci.* **2020**, *10*, 3286. [CrossRef]
3. Abeßer, J. A Review of Deep Learning Based Methods for Acoustic Scene Classification. *Appl. Sci.* **2020**, *10*, 2020, doi:10.3390/app10062020. [CrossRef]
4. Pepe, G.; Gabrielli, L.; Squartini, S.; Cattani, L. Designing Audio Equalization Filters by Deep Neural Networks. *Appl. Sci.* **2020**, *10*, 2483. [CrossRef]
5. Zieliński, S.K.; Lee, H.; Antoniuk, P.; Dadan, O. A Comparison of Human against Machine-Classification of Spatial Audio Scenes in Binaural Recordings of Music. *Appl. Sci.* **2020**, *10*, 5956. [CrossRef]
6. Tsui, B.; Smith, W.A.P.; Kearney, G. Low-Order Spherical Harmonic HRTF Restoration Using a Neural Network Approach. *Appl. Sci.* **2020**, *10*, 5764. [CrossRef]

7. Grachten, M.; Lattner, S.; Deruty, E. BassNet: A Variational Gated Autoencoder for Conditional Generation of Bass Guitar Tracks with Learned Interactive Control. *Appl. Sci.* **2020**, *10*, 6627. [CrossRef]
8. Navarro-Cáceres, M.; Merchán Sánchez-Jara, J.F.; Reis Quietinho Leithardt, V.; García-Ovejero, R. Assistive Model to Generate Chord Progressions Using Genetic Programming with Artificial Immune Properties. *Appl. Sci.* **2020**, *10*, 6039, doi:10.3390/app10176039. [CrossRef]
9. Pesek, M.; Leonardis, A.; Marolt, M. An Analysis of Rhythmic Patterns with Unsupervised Learning. *Appl. Sci.* **2020**, *10*, 178. [CrossRef]
10. Miao, X.; Sun, M.; Zhang, X.; Wang, Y. Noise-Robust Voice Conversion Using High-Quefrency Boosting via Sub-Band Cepstrum Conversion and Fusion. *Appl. Sci.* **2020**, *10*, 151. [CrossRef]