Guest Editorial: Special Issue on Computational Intelligence for End-to-End Audio Processing

COMPUTATIONAL Audio Processing techniques have been largely applied in diverse areas of strategic interest, like entertainment, art, human-machine interfaces, security, forensics, and healthcare. Developed services in these fields are characterized by a progressive increase of complexity, interactivity and intelligence, and the employment of Computational Intelligence techniques have enabled the achievement of a remarkable degree of automation with excellent performance.

The typical methodology adopted in these tasks consists in extracting and manipulating useful information from the audio stream to pilot the execution of target services. Such an approach is applied to different kinds of audio signals, from music to speech, and from sound to acoustic data. In the last few years, a new emerging computational intelligence paradigm has become popular among scientists working in the field. Broadly named as *end-to-end learning*, it consists in omitting any handcrafted intermediary algorithms in the solution of a given problem and directly learning all needed information from the sampled dataset. Due to its flexibility and versatility, this approach has attained a certain success in the Computational Audio Processing field.

Deep neural architectures are often used in these contexts and fed with raw audio data, whereas supervised, weakly-supervised or unsupervised training algorithms are employed to find a suitable data representation across the different abstraction layers to solve the task under study, i.e., classification, regression, prediction and detection. At the same time, an increasing attention has been given by the scientific community to the development of end-to-end solutions to synthesize raw audio streams, like speech or music.

The goal of this Special Issue is to understand how and to what extent novel Computational Intelligence techniques based on the emerging end-to-end learning paradigm can be efficiently employed in Digital Audio, in the light of all aforementioned aspects. In line with the mission of the IEEE Computational Intelligence Society Task Force in Computational Audio Processing (http://ieeeciscap.dii.univpm.it/), the organizers of this Special Issue have strived to bring the focus on the most recent advancements in the Computational Intelligence field, and on their applicability to Digital Audio problems from the end-to-end learning perspective.

The Issue collects seven original contributions, which cover some of the aforementioned topics providing to the reader an insightful panoramic view of the most recent research achievements. The selection of the present papers is the result of a rigorous review procedure, where at least three independent reviewers were involved with each paper, and up to three review rounds were performed before final acceptance for publication.

The first contribution of the Issue is by Ravanelli *et al.*, and deals with one of the most popular topics in Computational Audio Processing, i.e., automatic speech recognition (ASR), which has largely benefited from the recent advances in Deep Learning. Modern speech recognizers often employ acoustic models based on recurrent neural networks (RNNs), which are able to deal with long-term dependencies. In this contribution, the authors revise one of the most popular RNN models, namely gated recurrent units (GRUs), and propose a simplified architecture that turned out to be very effective for ASR in challenging environments characterized by significant noise and reverberation. The contribution is two-fold. First, the role played by the reset gate has been analyzed, showing that a significant redundancy with the update gate occurs, thus shoving the authors to remove it from the GRU design, leading to a more efficient neural architecture. Second, thehyperbolic tangent activation has been replaced with the Rectified Linear Unit (ReLU) one. This variation couples well with batch normalization, thus resulting in a systematic improvement of performance. Results show that the new architecture, called Light GRU not only reduces the training time by more than 30% over a standard GRU, but also consistently improves the recognition accuracy across different tasks, input features, noisy conditions, as well as across different ASR paradigms, ranging from standard Deep Neural Network (DNN) Hidden Markov Model (HMM), i.e., DNN-HMM speech recognizers to end-to-end Connectionist Temporal Classification (CTC) models.

Then follows the contribution of Salvati *et al.*, which again deals with noisy and reverberant acoustic environments and focuses on the speaker localization task. In particular, they apply convolutional neural networks (CNNs) to minimum variance distortion-less response (MVDR) localization schemes, investigating the direction of arrival (DOA) estimation problem applying a uniform linear array (ULA). CNNs are used to process the multichannel data from the ULA and to improve the data fusion scheme, which is performed in the steered response power (SRP) computation. CNNs improve theincoherent frequency fusion of the narrowband response power by weighting the components, reducing the negative impact of noise and reverberation.

Digital Object Identifier 10.1109/TETCI.2018.2809178

The use of CNNs avoids the handcrafted selection of acoustic cues by exploiting the computation performed in their convolutional layers. An extensive experimental campaign allowed demonstrating the superior localization performance of the proposed SRP beamformer with respect to other state-of-the-art techniques.

CNNs are used also in the work contributed by Hou *et al.*, in which the speech audio-visual enhancement (SE) problem is addressed. The authors take inspiration from the multimodal learning paradigm and propose a CNN based algorithm to process the audio-visual streams in a single unified model to perform speech enhancement. The proposed network, namely an audio-visual deep CNN (AVDCNN), is structured as an audio-visual encoder-decoder network, in which audio and visual data are first processed using individual CNNs, and then fused into a joint network to generate enhanced speech (the primary task) and reconstructed images (the secondary task) at the output layer. The model is trained in an end-to-end manner, and parameters are jointly learned through back-propagation. Results show that the AVDCNN model yields a notably superior performance compared with an audio-only CNN-based SE model and two conventional SE approaches, confirming the effectiveness of integrating visual information for SE purposes. In addition, the AVDCNN model also outperforms an audio-visual SE model already proposed in the literature.

Next, Ntalampiras focuses on wireless acoustic sensor networks (WASN), which represents a hot topic among the scientific community. The author observes that in the related literature several non-stationary phenomena related to the WASN operating conditions are not adequately considered and modeled. This work providesa problem formulation systematizing such issues and on top of that builds a sound classification system able to consider the presence of multiple sensor faults and of environmental noise. The proposed classifier is based on an Echo State Network operating at the sensor data level, while the decisions are combined at a higher level via a correlation-based dependency graph. The author carried out a thorough experimental campaign utilizing data coming from a WASN composed of 23 sensors aiming at the acoustic classification of moving vehicles.

Then, one finds three contributions related to the field of digital music processing. In the first one, Choi *et al.* study the application of a DNN to a specific music classification problem, namely music tagging. In this article, the authors investigate specific aspects of neural networks – in particular the effects of noisy labels. A large music tagging dataset is analyzed to investigate the reliability of training and evaluation. Using a trained network, the authors compute label vector similarities, which are then compared to ground-truth similarity. The obtained results show that neural networks can be effective despite relatively large error rates in ground-truth datasets, while conjecturing that label noise can be the cause of varying tag-wise performance differences. Lastly, valuable insights into the relationships between music tags are also provided.

The further two works target sound synthesis. Matthew *et al.* focus on a specific task, namely the automatic programming of sound synthesizers, which involves finding parameters for sound synthesizers using algorithmic methods. Sound matching is one application of automatic programming, where the aim is to find the optimal parameters for a synthesizer in order to emit a sound as close as possible to a certain target. The authors describe and compare several sound matching techniques used to automatically program the Dexed synthesizer, which is a virtual model of a Yamaha DX7. The techniques are a hill climber, a genetic algorithm and three DNNs. A sound matching task based on six sets of sounds, derived from increasingly complex configurations of the Dexed synthesis algorithm, has been defined. A bidirectional long short-term memory (BLSTM) network resulted to have the best performance and was able to match sounds closely in 25% of the test cases. This network was also able to match sounds in near real time, which provides a significant speed advantage over recent competitive techniques, based on search heuristics. The authors also describe their own open source framework, which can be adapted to different synthesizers and algorithmic programming techniques.

Finally, in the work of Gabrielli *et al.* an interesting end-to-end learning based approach for physics-based acoustic modeling is discussed and experimentally validated. In the state of the art, many numerical algorithms have been studied and proposed for sound synthesis. They are able to simulate complex physical phenomena in real-time with an acceptable computational cost, indeed reaching the market with commercial products. The authors observe that sound synthesis based on physical models could benefit greatly from automated methods that require less specific know-how and save the sound-designer valuable time. In this work, a novel neural approach based on the end-to-end learning paradigm is proposed for parameter estimation in physics-based sound synthesis and it allows achieving good results. The approach is presented in a formal way and an application to a practical use case is reported. Methodological issues, such as dataset generation, are also investigated.

To conclude, as guest editors, we would like to thank all authors for their contributions to this special issue. We also would like to express our sincere appreciation to all reviewers for their time and efforts. Finally, our gratitude goes to the IEEE Transactions on Emerging Topics in Computational Intelligence Editor-in-Chief, Professor Yew-Soon Ong, and the members of the Editorial board for their substantial support in the whole organizing and reviewing procedure. We look forward to more exciting changes in this field, which faces major reshaping at the time by such approaches as the featured end-to-end learning, ultimately approaching human and super-human audio processing capabilities.

S. SQUARTINI, *Guest Editor*
Department of Information Engineering
Università Politecnica delle Marche,
Ancona 60131, Italy

A. UNCINI, *Guest Editor*
Department of Information Engineering,
Electronics and Telecommunications
Sapienza University of Rome
Rome 00185, Italy

B. SCHULLER, *Guest Editor*
GLAM – Group on Language Audio & Music
Imperial College London,
London SW7 2AZ, U.K.

C.-K. TING, *Guest Editor*
Department of Computer Science
and Information Engineering
National Chung Cheng University
Chiayi 621, Taiwan