Editorial: Reconstruction of Audio From Incomplete or Highly Degraded Observations

THE RESTORATION of audio content, in particular speech and music from degraded observations, is a challenging and long-standing problem in audio processing. In particular this holds for severe degradations and incomplete observations, which are regularly encountered in practice. Traditional restoration techniques, such as interpolation or filtering, are often not applicable or perform poorly in this case. The advent of sparse signal processing in the very beginning of this century and, even more recently, of (deep) machine learning has opened wide new research and design opportunities for audio restoration, among many other signal processing problems. With the aid of such contemporary tools, researchers have recently been able to achieve unprecedented success in recovering or significantly improving the quality of severely degraded audio. As the field advances very quickly, the potential for improvement, as well as exploration, is hardly exhausted.

This special issue has been organized to gather contributions that would serve both as a comprehensive primer on the state-of-the-art, and a showcase of current developments within the field. Finally, this special issue features 11 original articles out of the total of 27 submissions received worldwide. The articles cover a number of important scenarios including degradation of signals in the time or the time-frequency domain, elimination of pulse-like disturbances, speech enhancement or phase recovery. Different methodological approaches are used, from time-frequency representations, through operator learning, to deep neural networks. In the call for papers, authors were invited to provide their code and data with their manuscript; thanks to the commitment of the authors, most articles are indeed enhanced by providing these materials, thus contributing to the sustainability, the reproducibility and the reliability of the research in the proposed topics.

The issue opens with the survey article *A Survey and an Extensive Evaluation of Popular Audio Declipping Methods* by Záviška, Rajmic, Ozerov and Rencker. It is devoted to the problem of clipping, or saturation, in audio signals, which has attracted much interest in recent years. The article surveys important contributions in this research area, focusing in particular on methods based on sparse signal representations, which has become the dominant methodology for this problem. Audio declipping algorithms are presented in terms of assumptions that are made about the audio signal, the modeling domain, and the optimization algorithm. Furthermore, the survey provides an extensive numerical evaluation of selected, popular declipping algorithms, performed on real audio data. The algorithms are evaluated in terms of different objective criteria, including perceptual metrics of sound quality. The article is accompanied by a repository containing the evaluated methods and a web page with audio excerpts.

While most of today's methods for audio declipping are based on the so-called *synthesis model*, the article *Sparse Analysis Model Based Dictionary Learning for Signal Declipping* by Li, Rencker, Dong, Luo, Plumbley and Wang proposes a novel algorithm based on the alternative analysis model. This sparse analysis-model-based declipping algorithm (SAD) learns the analysis dictionary/operator from the training clipped data, and at the same time estimates the declipped signal. The algorithm utilizes the recently introduced Analysis Simultaneous codeword optimization (ASimCO) and adapts it to the declipping task. Despite increased computational burden, the numerical experiments on speech and music demonstrate improved performance in signal to distortion ratio, compared to state-of-the-art methods A-SPADE, which uses a fixed dictionary, and ConsDL, which is synthesis model-based and uses a learned dictionary.

Time-frequency representations based on windowed Fourier and cosine dictionaries, or more generally on filter banks, are ubiquitous in audio processing. This is true for addressing severe audio degradations as well. Degradation may occur as an artifact of time-frequency processing, demanding further post-processing steps such as phase retrieval. The latter attempts to reconstruct audio from the magnitude of a time-frequency representation when the measured phase is invalid, e.g., after processing, or it is even unavailable.

In *Deep Griffin–Lim Iteration: Trainable Iterative Phase Reconstruction Using Neural Network* by Masuyama, Yatabe, Koizumi, Oikawa and Harada, the reconstruction of missing phases from spectrograms is addressed by revisiting the well-known Griffin and Lim (GL) algorithm with neural networks. The proposed architecture relies on a *sub-block* entity composed of a double projection similar to one GL iteration, coupled to a neural network. By pre-training the neural network on a simple refinement task, the authors decouple the number of iterations during runtime from the training process. As in the classic GL algorithm, the number of iterations can thus be adjusted on the fly to balance computational complexity and reconstruction quality. This is in stark contrast to prior work, which requires *unrolling* a number of iterations fixed at training time. The proposed architecture also contributes to the important challenge

Digital Object Identifier 10.1109/JSTSP.2021.3052087

1932-4553 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

of modeling complex-valued time-frequency data with neural networks. The solution proposed by the authors consists of the use of convolutional layers with gated mechanisms to form the complex-valued neural network in each sub-block.

The task of time-frequency phase retrieval is usually framed as a minimization problem with respect to a quadratic loss function. The article *Phase Retrieval With Bregman Divergences and Application to Audio Signal Recovery* by Vial, Magron, Oberlin and Févotte is concerned with the generalization of existing solvers for the phaseless reconstruction problem to a larger class of loss functions, more specifically beta-divergences, which have been proven useful in various audio processing tasks. Algorithms based on gradient descent, alternating direction method of multipliers and alternating projections are presented and evaluated under clean and noisy conditions. In particular in the noisy setting, promising results are achieved.

Incomplete or highly degraded time-frequency observations may occur in settings different from phaseless measurements. In *Time-Frequency Fading Algorithms Based on Gabor Multipliers* by Krémé, Emiya, Chaux, and Torrésani, a non-linear time-frequency filtering framework is proposed to address problems where localized time-frequency areas must be faded out. It is based on the formulation of an optimization problem in the signal domain, in which the time-frequency coefficients of the solution should satisfy two properties: in the targeted time-frequency areas, their energy should be controlled finely while outside those areas, they should fit the time-frequency coefficients of the observation. The solutions involve the so-called Gabor multipliers and some properties of their eigendecomposition, from which algorithms are derived.

In *Parameter Tuning-Free Missing-Feature Reconstruction for Robust Sound Recognition*, Liu and Qu propose a matrix completion scheme for recovering unreliable time-frequency measurements. The method is applied to improve DNN-based recognition and classification of speech and other audio data in the presence of different types of distortion, more specifically Gaussian and environmental noises. In the presented experiments, employing the proposed method as a preprocessing step in DNN-based feature recognition shows promising results. Notably, in contrast to previous methods for reconstructing missing time-frequency features, the proposed method does not require (hyper-) parameters to be tuned for optimal performance, and is therefore straightforward to apply.

The reduction of a quantity of hyper-parameters difficult to interpret is addressed also in the article *Bayesian Restoration of Audio Degraded by Low-Frequency Pulses Modeled via Gaussian Process* by Carvalho, Ávila and Biscainho. They consider the problem of restoration of audio signals degraded by long pulses with significant low-frequency content, a common defect found when reproducing old vinyl and gramophone recordings. The authors introduce a novel Bayesian approach with model parameters estimated via Markov-Chain Monte Carlo (MCMC) algorithm. This approach performs on par with the state-of-the-art in terms of perceptual quality of restored signals, while it circumvents also the need to know the location of the pulse beforehand, which is a drawback of previous works.

The article *Dictionary Learning for Sparse Audio Inpainting* by Tauböck, Rajbamshi, and Balazs develops a new sparsity-based algorithm for filling medium-sized gaps in audio signals. The recent Sparse Audio Inpainter (SPAIN), originally suggested to be applied segment-wise on the signal, is adapted into a global algorithm here, improving on the state-of-the-art. More importantly, the authors propose a dictionary learning procedure based on pre-defined deformations of a fixed Gabor dictionary. The learning step is designed such that it increases the sparsity of the time-frequency representation of the audio signal in question, resulting in boosting the performance of the modified Sparse Audio Inpainter in terms of objective metrics.

In *GACELA -: A Generative Adversarial Context Encoder for Long Audio Inpainting of Music,* Marafioti, Majdak, Holighaus and Perraudin developed a framework for performing audio inpainting of long segments of lost data from its time-frequency context. Based on a generative adversarial network conditioned of mel-scale time-frequency coefficients, the method generates STFT coefficients, which are converted to an audio signal using phaseless reconstruction with phase-gradient heap integration. The individual stages of the method are evaluated numerically, and the overall generation quality is studied based on extensive perceptual tests.

The work *On Filter Generalization for Music Bandwidth Extension Using Deep Neural Networks* by Sulun and Davies is concerned with a novel data augmentation method for DNN-based bandwidth extension for musical signals. Since various lowpass filters are used in practice to restrict the bandwidth of musical audio, bandwidth extension schemes considering only a specific type of lowpass filter suffer from filter mismatch when applied to unmatched data. To overcome the detrimental effect on quality caused by filter mismatch, this study proposes to augment the training data of DNN-based bandwidth extension by using various lowpass filters. It is demonstrated that this augmentation leads to better generalization to unseen filters.

The article *A Multi-Scale Feature Recalibration Network for End-to-End Single Channel Speech Enhancement* by Xian, Sun, Wang and Naqvi present a deep-neural-network-based method for end-to-end single channel speech enhancement. More specifically, they propose a multi-scale feature recalibration convolutional encoder-decoder with the bidirectional gated recurrent unit (BGRU) architecture. The multi-scale recalibration 2-D convolutional layers are used to extract both local and contextual signal features. In addition, a gating mechanism is used in the recalibration network to control the information flow among the layers, enabling them to retain speech and suppress noise. The fully connected layer is employed to compress the output of the multiscale 2-D convolutional layer with a small number of neurons, capturing the global information and improving parameter efficiency. The BGRU layers employ forward and backward GRUs, which help to exploit the interdependency among the past, current and future frames to improve predictions. The experiments confirm that the proposed method outperforms several state-of-the-art methods.

Finally, we would like to mention the article *Sparsity-based audio declipping methods: selected overview, new algorithms, and large-scale evaluation* by Gaultier, Kitić, Gribonval and Bertin. This article, unfortunately, could not be included in this issue despite an excellent evaluation. The manuscript presents a general algorithmic framework for declipping that covers

several existing as well as newly proposed approaches. Using a large-scale benchmark and a smaller scale formal listening test, guidelines are provided for various clipping levels, both for speech and music. The article will appear in one of the forthcoming regular issues of IEEE/ACM Transactions on Audio Speech and Language Processing.

The guest editor team thanks all the authors and reviewers whose contributions and effort have made this special issue possible. We would also like to thank Prof. Lina Karam, the Editor-in-Chief, for her kind support and suggestions, and to Mikaela Langdon from the IEEE publication office for keeping our special issue on track at different stages of the process.

PAVEL RAJMIC, *Lead Guest Editor*
Department of Telecommunications,
Faculty of Electrical Engineering and Communication,
Technická 12, 616 00 Brno,
Czech Republic

NANCY BERTIN, *Guest Editor*
Univ Rennes, CNRS, IRISA
F-35000 Rennes,
France

VALENTIN EMIYA, *Guest Editor*
Aix Marseille Université, Université de Toulon, CNRS, LIS,
Marseille,
France

NICKI HOLIGHAUS, *Guest Editor*
Acoustics Research Institute
Austrian Academy of Sciences, 1040 Wien
Austria

ALEXEY OZEROV, *Guest Editor*
InterDigital
35576 Cesson-Sévigné
France