Introduction to the Issue on Data Science: Machine Learning for Audio Signal Processing

I. Introduction

AUDIO signal processing is currently undergoing a paradigm change, where data-driven machine learning is replacing hand-crafted feature design. This has led some to ask whether audio signal processing is still useful in the "era of machine learning." This special issue aims to promote progress, systematization, understanding, and convergence of applying machine learning in the area of audio signal processing. Specifically, this issue is focused on work that demonstrates novel applications of machine learning techniques in the area of sound, speech and music signal processing, as well as methodological considerations of merging machine learning with audio signal processing.

This special issue covers various topics, from a review of deep learning in the audio domain (Purwins *et al.*) to speech recognition (Bavu *et al.*), from voice activity detection (Ariav and Cohen) to musical brain state decoding (Ntalampiras and Potamitis), from music information retrieval (Kim *et al.*) to bioacoustic classification for species monitoring (Thakur and Rajan), from polyphonic acoustic event detection (Vesperini *et al.*) to heart sound segmentation (Oliveira *et al.*) and from speech enhancement (Wood and Rouat) to source separation (Le Roux *et al.*).

II. Overview of Methods

The articles in this issue employ a variety of machine learning methods – conventional and new. These methods occur in novel combinations, are applied to new domains or are extended by audio-specific modifications.

Poorjam *et al.* suggest an outlier detection method for robust estimation of parameters (mean and covariance). Hidden Markov models are used in brain state decoding for music information retrieval (Ntalampiras and Potamitis), in conjunction with a linear time invariant model. Hidden semi-Markov models are used for heart sound segmentation (Oliveira *et al.*). Wood and Rouat use non-negative matrix factorization (NMF) for speech enhancement. Ivry *et al.* combine a deep encoder-decoder network to map the input signal to a lower-dimensional embedding and then train a support vector machine. Thakur and Rajan combine a dynamic kernel method based on Gaussian mixtures with NMF and an iterative ("deep") version thereof.

Various prominent deep learning methods are applied to or adapted to the audio domain. Long short term memory (LSTM) is used for source separation (Sun *et al.*) or for voice activity detection (Ariav and Cohen) in conjunction with a deep residual net and a WaveNet encoder incorporating both the auditory as well as the visual modality. Sample CNN is employed for music classification (Kim *et al.*), capsule networks for sound event detection (Vesperini *et al.*). Encoder-decoder models, such as autoencoders, are used for speech activity detection (Ariav and Cohen), for learning modulation filters to perform speech recognition in noisy conditions (Agrawal and Ganapathy) and for speech upsampling (Eskimez *et al.*) in conjunction with generative adversarial networks (GANs).

Deep learning methods are also extended by introducing new audio-specific layers, e.g. the learnable passband biquadratic filterbank (Bavu *et al.*) or layers for estimating complex time-frequency masks for source separation based on discrete representations (Le Roux *et al.*).

III. Overview of the Articles

The article contributed by Purwins *et al.* presents a broad overview of how deep learning has and continues to make impacts in applications of audio signal processing, specifically speech, music and acoustic environments. They review several applications centered around analysis, synthesis and transformation. The article also provides a nice introduction to major themes in audio signal processing with machine learning, such as task definitions, feature extraction, deep models, the use of data for training, and the evaluation of models. The article ends with a survey of several timely questions that remain open.

A. Analysis and Classification

The contribution of Bavu *et al.* nicely illustrates how well-studied concepts of digital signal processing can inform the specification and training of machine learning models. They propose an end-to-end model for audio classification where the first layer of the model is essentially a parameterized IIR filter bank expressed as a recurrent neural network. This is followed by convolutional layers operating on frame-wise magnitudes, and then fully connected layers to a decision function. With proper handling of stability issues on the filter bank parameters during training, the entire network can be optimized toward a specific classification task, e.g., speech command recognition, or environmental sound classification. The resulting models are competitive with the state of the art, and benefit from having fewer parameters.

Robustness against noise has always been a critical aspect of various speech or signal processing systems. Poorjam *et al.* propose a novel outlier identification approach. The approach

Digital Object Identifier 10.1109/JSTSP.2019.2914321

Read the Math

To read:
→ HMM
→ Encoder-decoder for speech activity recognition
→ linear time invariant model
→ parameterized IIR filter bank

uses the deterministic minimum covariance determinant to identify outliers in speech databases. In an unsupervised manner, Agrawal and Ganapathy use a deep variational autoencoder to learn modulation filters in the time-frequency domain. They show how this improves the robustness of supervised and semi-supervised speech recognition under the presence of additive and reverberation noise. In Ivry *et al.* the authors develop a deep neural network based encoder-decoder architecture to address the transient and stationary noise for voice activity detection. More specifically, diffusion maps are used in the encoder model to compress the high-dimensional acoustic signal information into a low-dimensional representation. The paper presents both a real-time and an off-line version of the proposed architecture for different application scenarios and experimentally shows their generalization capability. Ariav and Cohen extend voice activity detection to a multimodal setting which includes not only speech signals but also corresponding video that captures the speakers mouth region. They incorporate the dual input into an end-to-end deep neural network where a variant of WaveNet encoder extracts audio features from the raw waveform and a deep residual network extracts visual features from the image sequences. Both features are fused to form a joint representation and it is further encoded via LSTMs to make temporal predictions of voice activity. They show that the multimodal approach achieves better performance compared to unimodal models in challenging acoustic environments including high levels of noise and transients.

Two articles in this issue apply signal processing and machine learning to biological data: electroencephalograms (EEG, brain activity) and phonocardiograms (heart sounds). The article of Ntalampiras and Potamitis investigates the classification of musical audio signals using features computed from autoregressive modeling of mean EEG measurements of the human brain while listening to music. They train hidden Markov models for each class on these features and show that similarities exist between human subjects to a degree that meter and instrumentation can be inferred from averaged brain activity. The article of Oliveira *et al.* applies machine learning to the segmentation of heart sounds into different events and periods of activity (e.g., first beat and systolic period). They propose an unsupervised approach using hidden semi-Markov models and show evidence that its performance – when tuned to a specific subject – can equal that of a supervised approach.

Vesperini *et al.* employ capsule neural networks (CapsNets) that have been used in image processing and have been shown to perform better than convolutional neural networks (CNNs), especially regarding scarce robustness to affine transformations and detection of overlapping images. In this paper, CapsNets are used for polyphonic sound event detection and are shown to be effective and superior to CNNs for this task.

Two papers propose effective approaches to acoustic signal classification. The one by Thakur and Rajan uses a combination of the characteristics of matrix factorization with the discriminative abilities of kernel methods. The method was evaluated using four different bioacoustic databases. Results show the effectiveness of the proposed approach. The second paper by Kim *et al.* uses end-to-end learning with CNNs for audio classification. The algorithm operates on raw waveforms and has a small number of filters. In addition to evaluation on various tasks, the paper shows detailed analyses that provide a better understanding of the system's architecture.

B. Synthesis and Enhancement

One of the major problems for supervised machine learning methods is the requirement of high quality human-labelled data. Unfortunately, in practice, it is expensive to obtain such data. However, data without labels is usually accessible. Unsupervised learning methods using such data have always been an interesting research area. Wood and Rouat develop a real-time method for enhancing stereo mixtures of speech and real-world noise, combining NMF with generalized cross-correlation. The NMF dictionary is learned in a purely unsupervised way. Phase has always been an important information from signals that many systems tend to ignore. The authors use phase information to further improve the method they propose. Besides accuracy, latency is another crucial evaluation criterion for signal processing systems. The authors incorporate an asymmetric STFT windowing scheme to largely reduce the latency for applications in hearing aids. Additionally, the implementation of the algorithm has been open-sourced to help the research community to advance faster.

A recently introduced unsupervised learning algorithm, GAN, has opened a new paradigm of generating high-dimensional data such as image and audio. Eskimez *et al.* apply GANs to speech super-resolution, which is a task to expand the audio bandwidth by synthesizing the missing high-frequency content. They set up the generator network to have a convolutional autoencoder architecture with 1-D convolution kernels. It takes a narrow-band log power spectrogram of the speech input and generates high-frequency content of the log-power spectrogram. The output is concatenated with the narrow-band spectrogram to be compared with the full wide-band spectrogram in the discriminator for adversarial training.

Many audio processing systems use an audio representation based on the magnitude spectrogram of the audio signal and neglect the phase. However, in tasks like speech enhancement and source separation, the quality of the reconstructed signal in the time domain suffers if no suitable estimate of the phase is used. Sun *et al.* and Le Roux *et al.* estimate time frequency masks for source separation, including an explicit estimation of the phase. For this purpose, Sun *et al.* use complex signal approximation and a LSTM neural network to account for the long-term speech context. The model is evaluated with noise and speech interference. Le Roux *et al.* suggest the use of codebooks to estimate masks in the time-frequency representation, encoding phase and magnitude separately or jointly. To implement this, they propose layers that generalize sigmoid and convex softmax activation layers. They introduce methods to optimize those codebooks, train the network that implements them, and explain how to include those methods in an end-to-end learning framework. They also provide an in-depth parameter exploration of their model.

To explore:
→ Difference between multimodal & unimodal approach
→ GAN for speech recognition and speech super-resolution

IV. Outlook

There are many future challenges for machine learning in the audio domain, new and old, including the interpretation of learned models in high-dimensional spaces, problems associated with data-poor domains, adversarial examples, high computational requirements, and research driven by companies using large in-house datasets that is ultimately not reproducible. We hope that this special issue broadly shows the current state of the art in machine learning for audio signal processing, and identifies numerous directions calling for deeper investigation.

Hendrik Purwins, *Lead Guest Editor*
Department of Architecture, Design & Media Technology
Faculty of IT and Design
Aalborg University Copenhagen
Copenhagen 2450, Denmark

Bob Sturm, *Guest Editor*
Division of Speech, Music and Hearing
School of Electronic Engineering and Computer Science
Royal Institute of Technology (KTH)
Stockholm 114 28, Sweden

Bo Li, *Guest Editor*
Google Inc.
Mountain View, CA 94043 USA

Juhan Nam, *Guest Editor*
Graduate School of Culture Technology
Korea Advanced Institute of Science and Technology (KAIST)
Daejeon 34141, South Korea

Abeer Alwan, *Guest Editor*
Speech Processing and Auditory Perception Laboratory
Electrical and Computer Engineering, University of California, Los Angeles (UCLA)
Los Angeles, CA 90095 USA