The Internet of Audio Things: State of the Art, Vision, and Challenges

Luca Turchet, György Fazekas, *Member, IEEE*, Mathieu Lagrange, *Member, IEEE*, Hossein S. Ghadikolaei, *Member, IEEE*, and Carlo Fischione, *Senior Member, IEEE*

***Abstract*—The Internet of Audio Things (IoAuT) is an emerging research field positioned at the intersection of the Internet of Things, sound and music computing, artificial intelligence, and human–computer interaction. The IoAuT refers to the networks of computing devices embedded in physical objects (Audio Things) dedicated to the production, reception, analysis, and understanding of audio in distributed environments. Audio Things, such as nodes of wireless acoustic sensor networks, are connected by an infrastructure that enables multidirectional communication, both locally and remotely. In this article, we first review the state of the art of this field, then we present a vision for the IoAuT and its motivations. In the proposed vision, the IoAuT enables the connection of digital and physical domains by means of appropriate information and communication technologies, fostering novel applications and services based on auditory information. The ecosystems associated with the IoAuT include interoperable devices and services that connect humans and machines to support human–human and human–machines interactions. We discuss the challenges and implications of this field, which lead to future research directions on the topics of privacy, security, design of Audio Things, and methods for the analysis and representation of audio-related information.**

***Index Terms*—Auditory scene analysis, ecoacoustics, Internet of Audio Things (IoAuT), Internet of Sounds, smart city.**

I. Introduction

THE PARADIGM of the Internet of Things (IoT) refers to the augmentation and interconnection of everyday physical objects using information and communication technologies [1]–[3]. Recent years have witnessed an upsurge in IoT applications intersecting the areas of sound and music computing and semantic audio (see [4]–[6]). However, to date, the application of IoT technologies in audio contexts has received remarkably little attention compared to other domains, such as consumer electronics, healthcare, and geospatial analysis.

This article aims at creating a homogeneous and unified vision of the various efforts conducted in this domain, that we coin as the Internet of Audio Things (IoAuT). On the one hand, the creation of this vision strongly parallels similar efforts in the emerging field of the Internet of Musical Things (IoMusT) [7], where a number of devices for music production and consumption are connected within ecosystems that multiply possibilities for interactions between different stakeholders (including performers, audience members, and studio producers). On the other hand, this vision complements and extends IoMusT outlining requirements, applications, challenges, and opportunities that go well beyond the domain of music. In the specific context of this article, we highlight the difference between the terms "music," "audio," and "sound." With music, we exclusively refer to the musical stimuli, with audio we refer solely to the domain of the nonmusical auditory stimuli, whereas with sounds we intend the union of both music and audio. Consequently, we envision different IoT technologies and methods that address each of them.

First, we survey the existing technologies developed by practitioners across fields related to the IoAuT as proposed in this article. Second, we present a vision for the IoAuT and its motivations. We introduce the IoAuT as a novel paradigm in which smart heterogeneous objects (so-called *Audio Things*) can interact and cooperate between each other and with other smart objects connected to the Internet. The aim is to foster and facilitate audio-based services and applications that are globally available to users. Then, we reflect on the peculiarities of the IoAuT field, highlighting its unique characteristics in contrast to the IoT and IoMusT. Finally, we discuss the implications and challenges posed by the vision as well as we consider future directions.

Our focus is on technologies enabling the IoAuT as well as on current IoAuT research activities, drawing attention to the most significant challenges, contributions, and solutions proposed over the recent years. The result of our survey of the field reveals that at present, active research on IoAuT-related themes is rather fragmented, typically focusing on individual technologies or single application domains in isolation. *Ad hoc* solutions exist that are well developed and substantial, but their adoption remains low due to the issues of fragmentation and weak interoperability between existing systems. Such a fragmentation is potentially detrimental for the development and

Manuscript received April 1, 2020; revised May 1, 2020; accepted May 20, 2020. Date of publication May 25, 2020; date of current version October 9, 2020. The work of Mathieu Lagrange was supported by the Agence Nationale de la Recherche under Project ANR-16-CE22-0012. The work of György Fazekas was supported in part by the European Union H2020 Project under Grant 688382, in part by the Engineering and Physical Sciences Research Council under Grant EP/L019981/1, and in part by U.K. Research and Innovation under Grant EP/S022694/1. *(Corresponding author: Luca Turchet.)*

Luca Turchet is with the Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy (e-mail: luca.turchet@unitn.it).

György Fazekas is with the Center for Digital Music, Queen Mary University of London, London E1 4NS, U.K.

Mathieu Lagrange is with the French National Center for Scientific Research, University of Nantes, 44035 Nantes, France.

Hossein S. Ghadikolaei is with the Machine Learning and Optimization Laboratory, EPFL, 1015 Lausanne, Switzerland.

Carlo Fischione is with the Department of Network and Systems Engineering, KTH Royal Institute of Technology, 10044 Stockholm, Sweden.

Digital Object Identifier 10.1109/JIOT.2020.2997047

successful adoption of IoAuT technologies, a recurring issue within the more general IoT field [1]–[3]. As a consequence, this article not only seeks to bridge existing research areas and communities and foster cross-collaborations but also aims to ensure that IoAuT-related challenges are tackled within a shared, pluralist, and system-level perspective.

We believe that the IoAuT has the potential to foster new opportunities for the IoT industry, paving the way to new services and applications that are able to exploit the interconnection of the digital and physical realms, especially in the smart home [8] and smart city context [9]. Nevertheless, for IoAuT technologies to emerge and be adopted by end users, a number of technical and human interaction-related challenges need to be addressed. These include low-latency communication infrastructures and protocols, embedded IoT hardware specialized for audio, dedicated application programming interfaces (APIs), and software relying on specific ontological principles and semantic audio processes [10], [11], as well as the design of novel devices dedicated to audio analysis, production or consumption, employing appropriate signal processing, machine learning, deep learning, and artificial intelligence technologies. This article aims to identify and discuss the challenges arising in this novel vision of the IoAuT.

II. Internet of Audio Things: Concept and Vision

The IoAuT is an emerging field positioned at the intersection of IoT [1]–[3], human–computer interaction [12], [13], and artificial intelligence applied to audio contexts [14]. The IoAuT can be seen as a specialization of the IoT, where one of the prime objectives is to enable processing and transmission of audio data and information. The IoAuT enables the integration and cooperation among heterogeneous devices with different sensing, computational, and communication capabilities and resources. We clarify that in the context of the IoAuT, sensing is not only referred to audio signals via microphones but also to other sources providing quantities tracked by sensors, for instance, measuring vibrations or pressure variations.

We define an Audio Thing as "*a computing device capable of sensing, acquiring, processing, actuating, and exchanging data serving the purpose of communicating audio-related information.*" With "audio-related information" we refer to "*data sensed and processed by an Audio Thing, and/or exchanged with a human or with another Audio Thing.*" We define the IoAuT as "*the ensemble of interfaces, protocols, and representations of audio-related information that enable services and applications for the communication of audio-related information in physical and/or digital realms.*"

The IoAuT may be structured into ecosystems, just like the general IoT domain [15], [16]. An *IoAuT ecosystem* forms around commonly used IoAuT hardware and software platforms as well as standards. From the technological perspective, the core components of an IoAuT ecosystem are of three types.

1) *Audio Things:* Audio Things are entities that can be used to produce audio content or to analyze phenomena associated with auditory events, and can be connected to a local and/or remote network and act as a sender and/or a receiver. An Audio Thing can be, for example, a node in a wireless acoustic sensor network (WASN), a device responding to a user's gesture with auditory feedback, or any other networked device utilized to control, generate, or track responses to auditory content (see the examples of Audio Things used in the systems described Section III). We position Audio Things as a subclass of things, therefore they inherit characteristics of things in the IoT context, such as sensors, actuators, connectivity options, and software to collect, analyze, receive, and transmit data.
2) *Connectivity:* The IoAuT connectivity infrastructure supports multidirectional wired and wireless communication between Audio Things, both locally and remotely. The interconnection of Audio Things over local networks and/or Internet is achieved by the means of hardware and software technologies, as well as standards and protocols governing the communication.
3) *Applications and Services:* Various types of applications and services can be built on top of the connectivity, targeting different users according to the purpose of the Audio Things (e.g., human agents monitoring events, patients, and doctors). Such applications and services may have an interactive or a noninteractive nature. To establish interactive audio applications, real-time computations have particular importance. Analogously to the IoT field, the IoAuT can leverage Web APIs and Web-of-Things architectures [17]. Services can be exposed by Audio Things via Web APIs. Applications are part of a higher layer in the Web of Audio Things architecture letting users interact with content or Audio Things directly.

Fig. 1 depicts the main components of an architecture supporting IoAuT ecosystems. The data flow can be grouped into: 1) streams from the Audio Things, which includes audio streams and messages consisting of features extracted from the audio signals captured by the Audio Thing's microphones or other sensors producing audio signals like measurement streams and 2) audio streams arriving to the Audio Thing that are rendered as sounds by means of loudspeakers, as well as control messages governing the behavior of the Audio Thing. An example of the first type of data flow is represented by the data produced by nodes of WASNs (which typically have limited or no capability of receiving feedback messages). An example of the second type of data flow is the messages sent by a remote doctor to the smart sonic shoes described in [5].

A. Relation to Other Fields

The IoAuT has strong connections with and could be seen as a subfield of the Internet of Media Things (IoMT), which is defined as a network of things capable of sensing, acquiring, actuating, or processing media or metadata [18]. This is currently under exploration by MPEG.[1] We consider the IoAuT as a subfield of the IoMT (which in turn is a subfield of the IoT) and we position it at the intersection with the IoMusT (see

[1]ISO/IEC 23093 (IoMT): https://mpeg.chiariglione.org/standards/mpeg-iomt

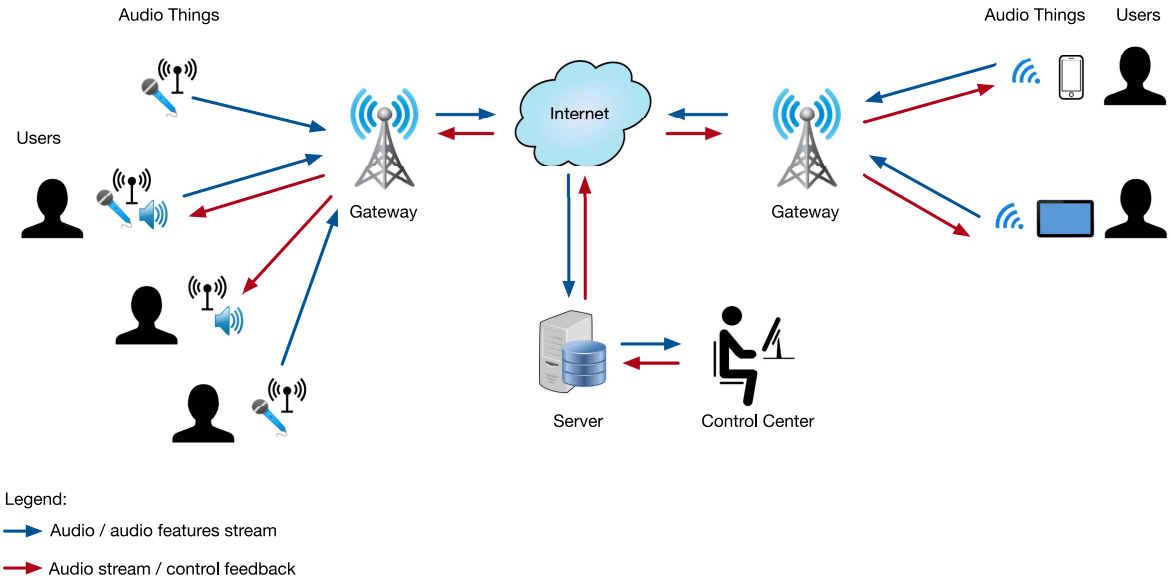

Fig. 1. Schematic representation of an architecture supporting IoAuT ecosystems.

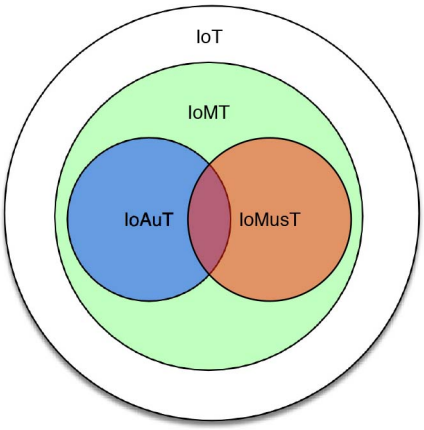

Fig. 2. Schematic representation of the relation between the IoAuT and the fields of IoT, Internet of Multimedia Things (IoMT), and IoMusT.

Fig. 2). The IoAuT differentiates from the IoMT for its focus on audio applications, whereas the IoMT also deals with other multimedia aspects, such as video. Similar to what the Web of Things[2] represents for the IoT, we use the term *Web of Audio Things* to refer to approaches taken to provide an application layer that supports the creation of IoAuT applications.

In contrast to the IoT, the IoAuT may pose stringent requirements and challenges related to the collection, analysis, and communication of audio-related information. For instance, a distributed array of microphones in a WASN might need to be synchronized tightly with low-latency communications to detect audio events in real time. Current IoT protocols and systems are insufficient to tackle this challenge. Along the same lines, the IoAuT demands novel analytic tools specific to the audio domain, which should be able to process large amounts of audio-related data and extract meaningful information given tight temporal constraints (e.g., for monitoring or surveillance purposes) and pose specific challenges in the areas of real-time signal processing and machine learning (see Sections IV-C and IV-F). In the same vein, current data models devised for the representation of the IoT domain are not adequate to describe the knowledge related to IoAuT ecosystems, which has the potential to foster interoperability across heterogeneous Audio Things.

It is important to highlight the distinctive features of the IoAuT with respect to the IoMusT.

1) The IoAuT does not have musical purposes, whereas the focal points of the IoMusT are live music performance, music pedagogy, studio productions and, in general, interactions between specific stakeholders, such as performers, composers, audience members, and studio producers. The purposes of stakeholders in the IoMusT are radically different from those of the stakeholders of the IoAuT. Music is a creative activity, and creativity is an aspect that is scarcely addressed in the IoAuT. As a consequence, most of the implications and challenges of the two fields are different (e.g., requirements of ultralow-latency transmission of musical content to guarantee credible interactions between performers). Nevertheless, some applications lie at the intersection of the two fields (see [19] where a wearer of a sensor-equipped garment could interact with an online repository of audio content, in a musical performance context).
2) The IoMusT is not a subfield of the IoAuT because, according to the vision reported in [7], the IoMusT is inherently multisensory, encompassing haptic feedback and virtual reality as communication media that extend the musical layer. Conversely, the IoAuT deals exclusively with the audio signal.
3) The level of human involvement is generally different in the two fields. First, whereas almost all audio signals within the IoMusT are generated or ultimately used by humans, IoAuT applications can make use of audio signals not related to human activities (e.g., monitoring environmental sounds such as birds). Second,

[2]https://www.w3.org/WoT/

in the IoMusT, most of the times, a human listener is involved in the interactions of the technology with the sonic content (e.g., the audience member enjoys the music of performers remotely connected; the music student listens to classes technologically mediated by smart instruments; the studio producer listens to the content retrieved from cloud-based repositories). Conversely, in several IoAuT applications (e.g., traffic monitoring and surveillance), the listening aspect performed by humans can be absent for the technology to work, and a system may completely rely on automatic processes.

4) The IoAuT may encompass activities, processes, applications, and services that are not present or are radically different in the IoMusT. For instance, sonification processes are normally absent in the IoMusT (e.g., the sonification [20] of human movements for rehabilitation purposes). Conversely, creative aspects typical of the IoMusT contrast with objective measurements that characterize most of IoAuT systems and applications. In addition, the context around the IoMusT stakeholders is different from the one that is around IoAuT stakeholders or the one used by them (e.g., environmental sounds of a city), and context-aware systems [21] may be radically diverse in the two fields. This necessarily involves different ontologies to represent the underlying knowledge as well as algorithms for context reasoning. Along the same lines, proactive services based on such context-aware systems are also diverse.
5) The quality of service for IoMusT applications may radically differ from those in the IoAuT. In the IoAuT, some nodes and/or sensors may be inactive for long periods of time, yet a system remainds operational, whereas, in the IoMusT, it is essential that each node, sensor, or actuator is running perfectly during user interaction. Also, in the IoAuT, the network may be utilized for very long periods of time (e.g., a WASN deployed in a smart city may run uninterruptedly for several months or years), whereas in the IoMusT, it is typically utilized to ensure the stakeholder interactions with the desired musical content (e.g., remote performances may last a few hours).
6) In the IoMusT, the audio signals need to be captured and reproduced in high quality to ensure credible musical interactions between stakeholders. In the IoAuT, this stringent constraint may not hold true for some systems and applications. For instance, some nodes in WASNs involved in surveillance applications embed low-cost microphones and analog-to-digital converters, which may have much lower sampling rates and resolutions.
7) The typical application of artificial intelligence also differs between the two fields. In the context of IoMusT, it is more common for AI technologies to be directly embedded in a single musical thing or a relatively restricted number of musical things, which have to extract, process, or transmit semantic metadata related to a musical audio signal. In the envisioned IoAuT context, it is typically expected that AI has to extract and process information obtained several from spatially distributed low-cost sensors, although single or multisensor embedded applications are also possible.

Besides the IoMusT, the IoAuT differentiates from other related technological areas present in the audio domain.

1) *WASNs:* Current WASNs typically employ embedded systems and network communication protocols not specifically conceived for audio processing tasks [22], which are instead key in the IoAuT. In addition, the IoAuT differentiates from today WASNs paradigms for the extensive use of semantic audio methods [11] able to extract structured meaningful information from the captured audio signal.
2) *Sonification:* The field of sonification [20] typically does not focus on networked scenarios involving embedded systems, where information to be sonified or resulting from the sonification activity is communicated across devices. In the IoAuT, applications may comprise the extension of traditional sonification methods toward networked scenarios, especially involving embedded systems.
3) *Semantic Audio:* The field of semantic audio [11] has rarely found application in IoT contexts dealing with the audio signal, and this is particularly true for the nonmusical domain. Typically, it does not focus on embedded systems, which are at the heart of the IoAuT. In the IoAuT, semantic audio methods are useful for advanced interoperability purposes across heterogeneous Audio Things.
4) *Embedded Audio:* Current embedded systems specific to audio processing offer a little range of connectivity options and scarce hardware–software methods supporting advanced machine learning algorithms. In the IoAuT vision, the connectivity component of embedded systems is crucial to devise advanced applications leveraging edge computing techniques while seamless accounting for privacy and security aspects.

Whereas the IoAuT stems from the technologies and paradigms listed above, it differentiates from them for a broader and holistic vision able not only to encompass all of them in a unified domain but also to extend them toward novel avenues. In the next section, these aspects are discussed in relation to the state of the art.

III. State of the Art

This section reviews key studies on which our IoAuT vision is based.

A. Wireless Acoustic Sensors Networks

One of the most compelling and important extensions of the IoT to the audio domain is represented by wireless acoustic sensors networks (WASNs) [22], [23]. These are networks of tiny and low-power autonomous nodes that are equipped with microphone-based sensing, processing, and communicating facilities. Such nodes are based on *embedded audio* platforms, i.e., embedded systems dedicated to digital audio processing (see the Bela board [24]), where a variety of audio software

runs on single-board computers, such as the Raspberry Pi or the Beaglebone [25].

Network architectures can be considered depending on the task at hand and the technical and ethical constraints that may be encountered (see [26] for a thorough discussion). One of the most typical application domain of WASNs is that of *acoustic monitoring* or *acoustic scene analysis* [14], [27]–[30], including urban noise pollution monitoring [31], environment surveillance (see [32]), anomalies detection [33], and wildlife monitoring [34]. For the last case, the WASN paradigm leads to the emergence of a new disciple called *ecoacoustics* [35] where scientists go beyond the single animal call analysis to gather statistics computed over large scale both in time and space [36], particularly relevant for ecosystems health monitoring.

A prominent example of WASNs for acoustic monitoring of urban areas is SONYC, a system that integrates sensors, machine listening, and data analytics to monitor and analyze urban noise pollution in New York [6], [30], [37]. Another kind of network implementations is also considered in various other places in the world. In Germany, the Stadtlaerm project [38], [39] aims at triggering events from a given taxonomy provided the input signal received by the sensors. Events can be "traffic," "shouting," etc. In that case, the complete processing chain is implemented on the sensor node, namely, recording, analysis, and classification. The main benefit of this type of architecture is that the data to be transmitted from the sensors to the servers has a very low bit rate and can be directly interpreted by humans. Some drawbacks are present. First, each processing step has to be energy efficient since it is embedded in the sensor. For the same reasons, modifying the processing chain, for example, updating the taxonomy of events, can be cumbersome as it requires a complete update of the embedded software. The DYNAMAP project [40] studies the development of such a network in two major cities in Italy. In France, the CENSE project focuses on the deployment of dense networks that transmit high-level spectral features [4], [41] which are designed to: 1) respect the privacy of the citizen [26] and 2) permit a quality of the description of the sound scene that goes well beyond the use of averaged acoustic pressure level that is commonly considered for those applications [42].

When considering WASN for urban areas monitoring, three main components are of importance to gain knowledge from the data gathered. First, the microphones shall be well calibrated and durable. Since most WASN are based on the "many but low cost" paradigm, the microphones must be relatively cheap. MEMS capsules, such as the ones used in smartphones are a relevant choice, although their durability for long time periods remains unknown [43].

Second, the sensors shall be reliable enough in order to obtain regularly sampled data in time and space [31], [44], [45]. Designing the topology of the network is also of crucial importance and needs to balance many constraints that are enforced by urban regulations [46]. Most WASNs are static, meaning that the sensors are not moving but some alternatives are considered, for example, by taking into account buses [47] and more importantly considering smartphones [48]–[61]. The latter case is particularly tempting, as the sensors are densely present in urban areas. However, the quality of the data has to be questioned as in any crowdsourcing paradigm [62], for instance, because the calibration of the microphone is of great importance for noise mapping applications [63]–[66]. Along the same lines, unmanned aerial vehicles (such as drones) also represent an opportunity for moving acoustic sensing. Recent examples of the use of these technologies include applications for search and rescue scenarios [67] and for ecoacoustic monitoring [68]. It is plausible to hypothesize that in the future drone-based networks will emerge, which leverage the acoustic information for surveillance, environment monitoring, and related applications.

Third, the gathered data shall be filtered [69], mined, and displayed [70]. For this purpose, data management systems need to be deployed [71], [72] and skillfully used. This final step is nontrivial and is currently researched extensively. It is mandatory to select relevant data to motivate given actuation. The challenge here is that the data analyst shall be able to mine a large amount of data that is diverse in terms of content and structure both in space and time [73].

Most large-scale WASNs do not consider the acoustic relations between the audio content captured at different nodes, which, for instance, can be exploited for source localization. Nevertheless, for smaller scale WASNs, or for more advanced nodes, source localization (single or multiple) can be performed using various techniques and algorithms (see [74]–[80]).

B. Sonification and the Internet of Things

A handful of works have explored the use of *sonification* techniques in conjunction with the IoT. Sonification is essentially a technique that consists of the transformation of data into sounds [20]. Sonification is referred to as the use of nonspeech audio to convey information. More specifically, sonification is the transformation of data relations into perceived relations in an acoustic signal for the purposes of facilitating communication or interpretation [81].

The first example of this category of works is the one reported in [82]. The authors sonified the electricity consumption of various appliances in the home, which were enhanced with a device able to monitor the amount of electricity used and were equipped with wireless connectivity to a base unit. This system aimed at enhancing users' awareness of electricity consumption for sustainability purposes.

A second example is represented by the work reported in [83] within the context of the so-called "Industry 4.0." Bederson [84] developed a preliminary prototype of a sonification-based system for acoustic monitoring of manufacturing processes and production machines, using the approach of "auditory augmented reality." The system uses an array of microphones placed onto a production machine (such as a 3-D printer) and is able to detect normal states or anomalies of the manufacturing process from the sound of the monitored machine. The classification of these states is based on machine learning algorithms running on a remote cloud, the result of which is communicated as continuous auditory stimuli to a

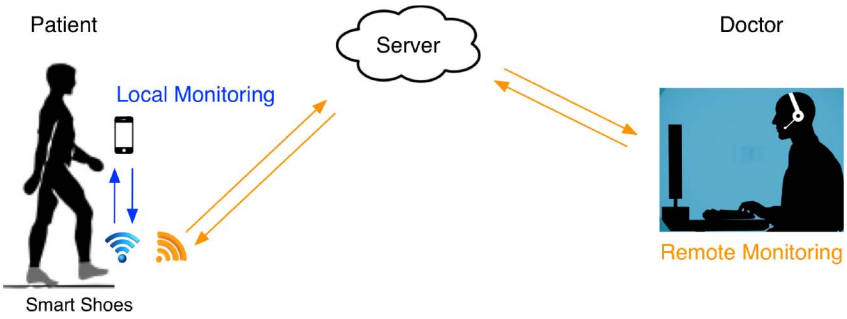

Fig. 3. Schematic representation of the local and remote interactions enabled by the system reported in [5].

worker operating near the machines, thanks to a wireless link to the connected headphones.

A third example is reported in [5], where a pair of smart sonic shoes is connected to the Internet to explore novel forms of sound-based motor therapies. This article is positioned in the context of remote patient monitoring [85], and more specifically, is conceived for telerehabilitation of motor disabilities [86]. As opposed to the previous two systems described in this section, such a work uses the approach of *interactive sonification* [87], which deals with the involvement of dynamic human interaction in the generation or exploration of information transformed into sound. The described prototype of smart sonic shoes is able to transform each footfall into a sound simulating walking on various surface materials [88], can collect data about the gait of the walker, as well as its sound production can be controlled via a remote doctor (see Fig. 3). The purpose of these shoes is to guide and improve walking actions in rehabilitation contexts due to the ability of sound to modulate the gait of a person (see [89]–[92]). The use of this portable device could enable patients to perform sound-based rehabilitation exercises while being comfortable in their homes. Patients and their families could be provided with cost-effective tools to autonomously monitor the progress of therapy. Doctors could be enabled to remotely monitor each patient and control the sonic feedback at each exercise. This has the potential to prevent patients to visit frequently the hospital by decreasing the cost for both patients and hospitals.

C. Auditory Augmentation of Connected Objects

Researchers have also focused on the sonic augmentation of everyday objects by means of tangible devices equipped with motion sensors, microphones, speakers, and wireless connectivity. A notable example in this category is StickEar [93], a small device attachable to an object, which encompasses the wireless sensor network technology and enables sound-based interaction. The device was conceived to empower people with the ability to deploy acoustic tags on any objects or space, and be informed of acoustic cues that may be produced by an object or a location. Applications of this device with sound-based input/output capabilities include remote sound monitoring, remote triggering of sound, autonomous response to sound events, and controlling of digital devices using sound.

D. Acoustic Data Transmission

Recent years have witnessed the emergence of the technology of device-to-device acoustic data transmission, which provides a means of proximity communication between co-located devices as an alternative to more widespread and common solutions such as electromagnetic communications. In more detail, information to be transmitted is encoded into inaudible ultrasonic sound waves that can be picked up by conventional microphones (which enables the adoption of this technology into portable solutions such as smartphones running dedicated apps). The nature of the information can range from text messages to images, and the technology could be used for payment transfers, user authentication, and smart city applications such as digital locks. At present, two main companies are leading such technological developments, Chirp[3] [94] and Trillbit.[4] Various online documents refer to this technology as an enabler for an *Internet of sounds*, envisioning it as a standard for IoT communications given the scalability of the solution.

E. Semantic Audio

Semantic audio is an interdisciplinary field providing techniques to extract structured meaningful information from audio [11]. It typically goes beyond simple case-specific audio analyses, for instance, the detection of a single type of event in an audio stream, as well as more complex audio feature extraction, classification, or regression problems. It does so by combining signal analysis to extract quantifiable acoustic features from audio, machine learning techniques to map acoustic features to perceptually, environmentally, or musically meaningful features, and structured representations that place these features into possibly multirelational or heterogeneous hierarchies [10], [95] using, for example, semantic Web ontologies [96].

Semantic audio is a core concept in the IoAuT because it provides the means for both analyzing and understanding the content of audio streams or recordings as well as communicating this information between Audio Things. These devices are typically situated in complex distributed environments, consisting, for instance, of networks of standalone sensors, embedded systems in mobile sensing and communication devices, as well as data and control centers. This creates

[3]https://chirp.io/
[4]https://www.trillbit.com/

the need for complex and versatile yet interoperable audio analysis and representation techniques which is at the heart of semantic audio.

There are relevant examples of systems relying on semantic audio. For instance, the Audio Commons ecosystem [97] provides a mechanism to combine generic audio and content repositories within creative application scenarios [19], [98]–[102] that include sounds collected from the broader environment. A key concept in these systems is the combination of the two primary aspects of semantic audio: 1) machine analysis and automatic tagging of content and 2) its representation in an appropriate semantic hierarchy for interoperability [10], [103]. Tagging comes with its own challenges owing to noisy annotations in relevant labeled data sets, lack of temporal accuracy in the annotations, i.e., often only weakly labeled data are available, as well as the presence of multiple sound sources in an audio stream or recording [104]–[106].

Detection is followed by annotation within a semantic hierarchy that supports efficient communication and interoperability. This requires a shared conceptualization of low- to high-level acoustic features, as well as meaningful labels across different audio-related domains. Several ontologies have been proposed for these purposes, including those for audio features [107], effects and transformations [108], mobile sensing in the audio context [109], as well as ontologies that bind complex workflows and signal routing in audio processing environments [110] and ontologies that bind distributed content repositories together [103].

F. Web-Based Digital Audio Applications

The Web Audio API is one of the most recent among the technologies for audio applications on the Web and its use is becoming increasingly widespread [111]. It enables real-time sound synthesis and processing on Web browsers simply by writing the JavaScript code. It represents a promising basis for the creation of distributed audio applications such as those envisioned in the IoAuT. Different from Java or Flash, which are implemented in the form of browser plugins, the Web Audio API is implemented by the browser itself. Moreover, the Web Audio API is a World Wide Web Consortium (W3C) proposed standard.[5]

Recently, Web Audio technologies have been employed in embedded systems, thus bridging the realm of smart objects with that of audio applications leveraging the Web. An example in this category is reported in [112]. The authors proposed a preliminary system consisting of a network of nodes based on the Raspberry Pi platform. Each node runs a Web Audio application that could exploit a number of libraries previously built for mobile-based applications (e.g., for synchronization purposes [113]), with the purpose of implementing a distributed architecture for musical performances.

Along the same lines, Skach *et al.* [19] proposed a system that links Web-based digital audio technologies and embedded audio. Their system consists of a sensor- and actuator-equipped garment allowing for the interactive manipulation of musical and nonmusical sounds retrieved from online sound repositories. Specifically, the authors developed a jacket-based and trousers-based prototype for body-centric sonic performance, which allows the wearer to manipulate sounds through gestural interactions captured by textile wearable sensors. The data tracked by such sensors control, in real time, audio synthesis algorithms working with content downloaded from Audio Commons,[6] a Web-based ecosystem for repurposing crowd-sourced audio such as the Freesound.org[7] repository (see Fig. 4). The prototype enables creative embodied interactions by combining e-textiles with Web-based digital audio technologies.

Fig. 4. Schematic representation of the sensor- and actuator-equipped garment presented in [19], which interacts with the audio content repository Freesound.org.

To date, a number of promising projects have demonstrated how audio-based applications can be bridged into the Web browser via the Web Audio API. A large amount of these projects have focused on the musical domain (see [114]–[116]). A noticeable exception is represented by the FXive project [117], an online real-time sound effect synthesis platform. Various algorithms are used to synthesize everyday sounds, ranging from models for the contact between objects [118] to models for footstep sounds [88]. FXive represents a service targeting designer of sound effects, with the aim of replacing the need for reliance on sound effect sample libraries in the sound design. Designers of sound effects rather than searching for sound libraries and attempting to modify the retrieved sound samples to fit a desired goal, can directly shape their sounds by using the online service.

IV. CHALLENGES

The IoAuT inherits many challenges of the general field of IoT (see [119]). In addition to these, the practical realization of the envisioned IoAuT poses specific technological and personal data-related challenges. The realization of the

[5]https://www.w3.org/TR/webaudio/

[6]http://audiocommons.org

[7]http://freesound.org

IoAuT vision described in Section II occurs through the evolution of the network and services' infrastructure as well as of the capabilities of Audio Things connecting to them. We identify eight areas that currently hinder many interesting IoAuT application scenarios: 1) connectivity; 2) interoperability and standardization; 3) machine analysis of audio content; 4) data collection and representation of audio content; 5) edge computing; 6) synchronization; 7) privacy and security; and 8) Audio Things design.

A. Connectivity

Communication based on audio-related information may pose stringent requirements and challenges, which is why many of the general-purpose protocols designed for the IoT may not be appropriate or feasible for the IoAuT. Distributed audio sensors may require low-delay communications for real-time monitoring and processing [120]. This is the case of event detection, such as crashes and accidents, which could be monitored by distributed microphones that could also contribute to the control of traffic lights and car speed in the neighboring of the event. Moreover, in addition to low latency, the communication network may have to support high data rates, such as when the signal-to-noise ratios are low and signals will have to be quantized with high resolution in order to extract the desired information. In this regard, the use of mmWave wireless communications could be an enabling technology for IoAuT, because they potentially enable ultralow latency and massive bandwidths at the physical layer of the wireless communications [121]. Audio applications will experience latency at all layers of the protocol stack. Hence, many aspects of communication systems will need to be reconsidered and customized for audio transmission purposes.

IoAuT applications will likely generate large data sets, which we will have to analyze in real time. Reliable automatic speech recognition can now be performed [122] even in a noisy environment. To achieve such impressive results, machine learning needs big data sets and very large computational and communication resources, especially for the training tasks [123]. However, in IoAuT applications, data sets of any size will be distributed among several nodes (people, devices, objects, or machines) that might not be able to time share data due to bandwidth or privacy constraints, or may not have enough computational resources to run the machine learning training tasks. Existing machine learning methods and related algorithms are mostly intended for proprietary or high performing networks (e.g., in data centers), and would greatly stress public communication networks, such as IoT and 5-6G wireless networks [122], [124]. We expect that the research community will have to address several fundamental advancements within machine learning over networks, which will likely use ideas from active learning and distributed optimizations over networks.

One major issue to apply machine learning over communication networks for the IoAuT is the fundamental bandwidth limitations of the channels. The huge number of nodes and their data sets' transmissions may congest the practically available bandwidth. The emerging technology of extremely low-latency communications will rely on short packets that carry a few bits [120]. The nodes generating audio data may not have enough communication bandwidth to transmit data to the place where it has to be analyzed, or simply not enough computational power to perform local training and data analysis. A further problem is that the privacy and security are key societal concerns. A malicious observer could reconstruct a node's (such as a person's) private audio information, or misuse the analysis of data belonging to others.

Finally, developing efficient communication protocols and shared conceptualization of the information being distributed is also important. For example, communication bandwidth may be saved if IoAuT devices are able to communicate using short and universally accepted identifiers to signal certain conditions instead of complex (e.g., XML) data structures. This will be discussed in the following sections in more detail.

Thus, we suggest that the design and deployment of alternative communication techniques and protocols together with the audio machine learning tasks are necessary to target better performances for the support of communication of audio-related information over the IoAuT infrastructure.

B. Interoperability and Standardization

What emerges from the survey of the literature presented in Section III is a picture of the IoAuT as a field rather fragmented, where various authors have focused on single technologies or single application domains. Such a fragmentation hinders the development and successful adoption of the IoAuT technologies. Standardization activities represent a central pillar for the IoAuT realization as the success of the IoAuT depends strongly on them. Indeed, standardization provides interoperability, compatibility, reliability, and effective operations on both local and global scales. However, much of this article remains unrealized. Whereas various *ad hoc* solutions exist, their adoption is still low due to the issues of fragmentation and weak interoperability. More standardized formats, protocols, and interfaces need to be built in the IoAuT to provide more interoperable systems. This issue is also common to the more general IoT field [125].

Within the IoAuT, different types of devices are used to generate, detect, or analyze audio content, and need to be able to dynamically discover and spontaneously interact with heterogeneous computing, physical resources, as well as digital data. Their interconnection poses specific challenges, which include the need for *ad hoc* protocols and interchange formats for auditory-related information that have to be common to the different Audio Things, as well as the definition of common APIs specifically designed for IoAuT applications. Semantic technologies, such as semantic Web [126] and knowledge representation [127] can be envisioned as a viable solution to enable interoperability across heterogeneous Audio Things. However, to date, an ontology for the representation of the knowledge related to IoAuT ecosystems does not exist.

A common operating system for Audio Things can be considered as a starting point for achieving interoperability. Recent technological advances in the field of the music technology have led to the creation of platforms for embedded

audio that are suitable for IoAuT applications. To date, the most advanced platform for embedded audio is arguably the Elk Audio OS developed by Elk.[8] Elk Audio OS is an embedded operating system based on Linux. It uses the Xenomai real-time kernel extensions to achieve latencies below 1 ms, which makes it suitable for the most demanding of low-latency audio tasks. It is highly optimized not only for low-latency and high-performance audio processing but also for handling wireless connectivity to local and remote networks using the most widespread communication protocols as well as *ad hoc* ones. Recently, the operating system has integrated support for 5G connectivity. Elk Audio OS is a platform independent, supporting various kinds of Intel and ARM CPUs. Thanks to these features, Elk has the potential to become a standard for operating systems running on various kinds of embedded hardware for the nodes of the IoAuT.

C. Machine Analysis of Audio Content

Traditionally described as acoustic pressure levels computed over long time scales, audio is now considered in much more detail in order to gather rich information of the sound environment. While in this section, the focus will be put on urban areas monitoring, it is worth noticing that the growing field of ecoacoustics has also many challenges and potential applications [106].

The recent availability of large amounts of recordings has fueled research on the use of machine learning methods to gather high-level information about the sound environment, particularly in urban areas [25]. A scientific community emerged in 2010 to address this topic and the first Detection and Classification of Acoustic Scene and Events challenge was launched in 2013 [14], sponsored by the IEEE Acoustics, Speech, and Signal Processing Society. As the name of the challenge states, two levels of information are considered. One at the time scale of the event, where precise timing detection is required and the other at a longer time scale, where an abstract description of the audio has to be predicted. The typology of the predicted events and scene types is task dependent.

The acoustic scene classification task was originally tackled by considering probabilistic classification techniques based on explicitly designed audio features [128]. Those approaches have now been replaced by end-to-end deep learning methods [129], that tend to perform better and better as the volume of available training data increases.

Nonnegative matrix factorization techniques are well suited for the acoustic event detection task and methods based on these techniques perform well [130]. With special care, deep learning techniques also achieve state of the art results [131]. Due to the scarcity of training data for the acoustic event detection task, considering data augmentation techniques often mandatory [132].

New analytic tools are needed to make the most of the IoAuT. Such tools should be able to process large amounts of audio-related data and extract meaningful information given tight temporal constraints. Deep learning [133] offers encouraging ways to obtain high-level features that could capture the nature of the event that generated the auditory content.

In this context, a substantial challenge is learning from noisy [104], [134] and weakly labeled [135] data sets, which are much more readily available. To this end, the development of appropriate neural network architectures is ongoing work, where the use of attention mechanisms [135], [136] provides a promising direction.

In the envisioned IoAuT ecosystem, an Audio Thing may possess multiple spatially distributed sensors which poses another challenge. While deep learning applied to audio provides state of the art performance in many tasks and has become a mature field of research, there is currently very little attention to problems involving multiple audio sensors while multisensor data processing and integration using deep learning are in its infancy. This usually involves the use of case-specific tricks or data fusion techniques [137], while the system may also need to deal with imperfect time synchronization in light of the issues discussed in Section IV-F. There are network architectures capable of comparing audio signals or processing them in a sequence (see [138]) but real-time multisensor processing remains a challenge.

D. Data Collection and Representation of Audio Content

Several common challenges exist across the different audio analysis methodologies mentioned in Section IV-C. These include the problem that machine learning-based techniques require large amounts of accurate training data that covers most or all relevant use cases. This is a substantial problem owing to both the expense and difficulty of collecting data, as well as the difficulty of accurately annotating data.

For specific domains, such as an office environment, manual data collection is feasible [14], [128]. This approach does not necessarily scale however. The problem can be addressed using crowdsourcing both content and annotation, as is the cases of Freesound.org, which provides community created data sets [139]. These are increasingly annotated within semantic hierarchies [140] such as those provided by the AudioSet ontology [141]. However, an accurate taxonomy, let alone more complex multihierarchical relationships between sound events are difficult to represent and to agree upon by multiple annotators. This is a challenge in part because many existing representations follow a single hierarchical tree structure, while in the real world, graph-structured complex relationships are much more common and potentially more useful. A comprehensive ontology that addresses this issue is yet to be developed.

E. Edge Computing

State of the art deep learning models achieve remarkable performance and are being widely used in multimedia systems [131], [142]–[145]. Many of these models can have tens of millions of parameters (e.g., AlexNet [146]), to achieve such high performance. However, the realization of the IoAuT demands applying these heavy models to cheap sensor devices. Limited computational and energy resources prohibit the use

[8]https://www.elk.audio

of heavy training and/or inference algorithms [147], short in-device storage challenges the deployment of heavy pretrained models [148], and low bandwidth links and real-time nature of the audio signals hinders the use of traditional cloud-based inference [149], [150]. Much fundamental research is still needed to properly address the urgent multidisciplinary research problem of edge computing for the IoAuT.

There are multiple existing solutions to support the AI interface at the edge. Examples include hardware accelerators, such as Intel's Neural Compute Stick 2 (NCS2)[9] or Google's Edge Tensor Processing Units (TPU).[10] These are compatible with common single-board computers. However, these solutions are suitable only for simple visual and audio recognition tasks, with no guarantees on real-time processing or model compression.

A series of recent works focused on compressing big neural networks to save storage, energy, communication, and computational resources at the edge nodes. The proposed approaches for solving this problem could be broadly classified into two categories. The first class includes methods that reduce the number of parameters in the model [151], [152]. The second class includes methods to reduce the quantization precision for storing and processing model parameters [153]. Iandola *et al.* [151] proposed smaller modules as building blocks for emulating AlexNet. With their approach, the authors designed an architecture that has $50\times$ fewer parameters than AlexNet [146] with almost no loss in the inference accuracy. However, this approach is specifically designed for AlexNet, and it is not easily applicable to compress other big models. Simpler approaches include pruning, deleting the connections of the trained model with small values, and quantization, reducing the number of bits needed to store a parameter. Han *et al.* [154] proposed the deep compression algorithm that combines both pruning and quantization, leading to $35\times$ compression of AlexNet. These solutions often need the availability of the original data set to retrain the new (small) model, which is not available in many use cases due to privacy or intellectual property protections. Krishnamoorthi [155] proposed a quantization-aware training, in which the author added artificial quantization noise to the parameters during the training phase to make it more robust to potential future quantization. However, this approach suffers an inherent tradeoff that adding more quantization noise to the training pipeline may lead to a very bad solution for the original less-noisy problem. Moreover, in the literature, the model compression techniques have been applied mostly to natural language processing and image classification, whose signal statistics and machine learning methods are very different from real-time audio processing.

In many scenarios, e.g., WASNs, edge computing may face a massive connectivity challenge where many edge devices may need to coordinate and send some locally processed information to a central coordinator [156]. Sharma and Wang [157] proposed a framework to exploit the network-wide knowledge at the cloud center to guide edge computing at local IoT devices. However, it cannot address the problem of massive connectivity and the resulting significant performance drop of wireless networks. Device-to-device communications and local collaborations among the Audio Things are essential, yet the area is very open in the literature. Such collaboration can also improve the robustness of the decision making and real-time data analytics to potential outlier and/or straggler devices and compensate for per-device performance reduction due to the use of compressed models and lower precision.

F. Synchronization

Distributed computational resources need to be synchronized in time, though the degree of precision to which this synchronization shall be is application dependent.

In order to maintain a good level of synchronization between nodes of a processing graph, two quantities shall be controlled: 1) the *local time* of each node and 2) the *delay*, i.e., the amount of time needed by the node to record or playback and audio signal once the request to do so have been received. Quality of service is ensured by minimizing the following quantities: the variance of the difference between the local time of each node σ_t and the variance of the difference between the delays of each node σ_d. In order to better grasp the importance of these quantities, three use cases are now described, with growing requirements in terms of synchronization accuracy.

1) In WASN, the data have to be synchronized in order to be able to interpret some behaviors happening across different nodes. In this case, σ_t and σ_d shall remain below the second.
2) On the contrary, distributed playback systems that operate over the Internet protocol (IP) [158], like RAVENNA [159] or Dante [160], reducing σ_t and σ_d below the millisecond is critical as the human auditory system is highly sensitive to phase delays. In this case, σ_d is not a strong issue as the nodes are simple playback systems that are not in charge of audio processing or synthesis and in most commercial systems of very similar hardware specifications.
3) Laptop [161], [162] or smartphone [163] orchestras are much more challenging as they have the same requirements as distributed playback systems but have to face much more stress on σ_d as the nodes of the network have to process and synthesize audio before rendering using a wide diversity of the hardware platform. The latter calls for software-based solutions [164] that are inherently limited in terms of precision.

Time synchronization issues are ubiquitous in distributed computing, therefore many tools are available to minimize σ_d. It has been tackled for standard usage by the network time protocol (NTP) proposed in [165]. This protocol stands out by the virtue of its scalability, self-configuration in large multihop networks, robustness to failures and sabotage, and ubiquitous deployment. NTP allows the construction of a hierarchy of time servers, multiply rooted at canonical sources of external time.

Despite being in use in many sensor networks, it may face issues with this specific application. The first is that NTP

[9]https://software.intel.com/en-us/neural-compute-stick
[10]https://coral.ai/

assumes that computational and network resources are cheap and available. While this may hold for traditional networks of workstations, it may not be the case for low consumption sensor networks. Furthermore, the dynamic topology of the network can influence the degree of precision to which a recently disconnected node is synchronized. Fortunately, NTP operates well over multihop networks. If those matters are of importance for the considered use case, other approaches, such as the ones researched in [166] and the ones based on flooding proposed in [167] and [168] may be considered.

When there is a need for very precise synchronization, the precision time protocol (PTP) can be considered. Indeed, NTP targets millisecond-level synchronization, whereas the PTP targets nanosecond-level synchronization. This can only be achieved by considering dedicated hardware at least for the masters responsible for broadcasting the trusted time.

Tackling the issue of minimizing the delay for laptop or smartphone orchestra can only be achieved for most applications by considering calibration in order to estimate the maximal delay achieved by the nodes. Mostly based on standard software tools such as Web Audio, the proposed solutions will improve as the software tools improve over those matters. Still, the results presented in [113] are already quite satisfying, as they report σ_d of 0.2–5 ms for a wide range of devices. If the use of hardware is possible, one can consider low-cost alternatives to the PTP hardware that broadcast GPS reference time over the network [169].

G. Privacy and Security Challenges

The IoAuT paradigm brings challenges related to personal data, such as privacy and security, since some Audio Things have the ability to automatically collect, analyze, and exchange data related to their users.

Given the pervasive presence of the IoAuT, transparent privacy mechanisms need to be implemented on a diverse range of Audio Things. It is necessary to address issues of data ownership in order to ensure that Audio Things users feel comfortable when participating in IoAuT-enabled activities. IoAuT users must be assured that their data will not be used without their consent. Concerning the IoT field, the Weber recently highlighted the growing need for technical and regulatory actions capable of bridging the gap between the automatic data collection by IoT devices and the rights of their users, who are often unaware of the potential privacy risk to which they are exposed [170], [171]. Examples include data leaks and unauthorized collection of personal information [172], [173]. Necessarily, the same holds for the IoAuT. The definition of privacy policies is one approach to ensure the privacy of information. Audio Things can be equipped with machine-readable privacy policies, so that when they come into contact they can each check the other's privacy policy for compatibility before communicating [174]. Security risks also come from hardware hacking, which points toward the necessity of the hardware-level encryption to ensure privacy policies are adhered to. Thus, it is paramount that Audio Things designers and manufacturers adopt a "privacy by design approach" as well as incorporate privacy impact assessments into the design stage of Audio Things.

Since Audio Things are wireless devices, they are subject to the security risks of wireless communications. In today's Internet, encryption is a key aspect to ensure information security in the IoT. As a consequence, Audio Things should be designed to support robust encryption, which poses the challenge of making these devices powerful enough to support it. Nevertheless, enabling encryption on Audio Things requires algorithms more efficient and less energy consuming, along with the development of efficient key distribution schemes [175]. Importantly, a uniform security standard should be developed by the IoAuT research community and industry in order to ensure the safety of the data collected by Audio Things. This challenge is currently unsolved also in the IoT field [170].

WASNs can be very useful to gather rich information about different aspects of the quality of life in urban areas. Having precise knowledge about that is mandatory for effective actuation. This, in the end, will improve the quality of life of citizens. That being said, the deployment of WASNs shall be performed with a lot of care regarding the preservation of the privacy of citizens. Even if speech is a rather weak biometric indicator, the information gathered using WASNs must not contain any speech information that could be used by humans or computers to capture information about the location or spoken sentences of individuals. Following the different designs detailed in Section III-A, different means can be considered. If only the detection labels are propagated on the network, this privacy is guaranteed by design. If spectral features are sent, the frame rate must be sufficiently low to ensure that speech cannot be reproduced [26]. If the raw audio has to be transmitted, source separation techniques can be considered to remove speech before transmission [176].

Novel business models can emerge leveraging data arising from IoAuT technologies, for example, to provide services related to monitoring activities (such as ambient intelligence or surveillance). Ethical and responsible innovation are crucial aspects that need to be considered when designing such services to ensure that they are socially desirable and undertaken in the public interest. Ultimately, the key to the success of the IoAuT will be the users' confidence. Hardware and software manufacturers will need to convince consumers that the use of Audio Things is safe and secure and to do this, much work is still needed.

H. Audio Things Design

One of the most stringent design challenges for Audio Things relates to the limited energy resources available to most of them (e.g., the nodes of WASNs). Indeed, the battery life of the devices represents a constraint for communication and computational energy usage. Typically, besides a system for wireless communication Audio Things encompass microphones and a processing board, and in other cases also loudspeakers and various kinds of sensors. All these components require a substantial (and in most cases continuous) amount of energy. Solar panels have been utilized in various

systems to cope with this issue (see [34]) but advances in miniaturization and power of batteries are necessary. Another possibility would be to augment existing objects deployed in smart cities that are distributed and by default are connected to a power supply, such as smart street lights, as in the CENSE project [41].

Another design challenge relates to the creation of solutions able to provide high quality in recording and/or sound production, while still being cost effective. To date, cost-effective solutions that can be deployed on large scale are MEMS microphones, which on average, however, do not offer a wide frequency response (typically 100 Hz–10 KHz) and resolutions, which may translate into low analytics capabilities. In addition, miniaturization of the components of an Audio Thing (from the microphone to the computational unit) is also a desirable feature.

Furthermore, novel design paradigms should be devised for systems exploiting the yet unexplored opportunities offered by linking the IoT field with that of sonification or interactive sonification. The IoT has the potential to facilitate the emergence of novel forms of interactive sonification that are the result of shared control of the sonification system by both the user performing the gestures locally to the system itself, and one or more remote users. This can, for instance, impact therapies based on auditory feedback where the control of the sound generation is shared by the patient and the doctor (see the smart sonic shoes reported in [5]). The effect of such therapy can be remotely monitored and data from several patients performing such a sound-based therapy can be collected by means of big data analytics techniques.

V. Conclusion

This article introduced the IoAuT as a novel paradigm in which heterogeneous devices dedicated to audio-based tasks can interact and cooperate with one another and with other things connected to the Internet to facilitate audio-based services and applications that are globally available to the users. We presented a vision for this emerging research field, which stems from different lines of existing research including IoT, sound and music computing, semantic audio, artificial intelligence, and human–computer interaction. The IoAuT relates to wireless networks of smart devices dedicated to audio purposes, which allow for various forms of interconnection among different stakeholders, in both co-located and remote settings. The IoAuT vision offers many unprecedented opportunities but also poses both technological and nontechnological challenges that we expect will be addressed in upcoming years by both academic and industrial research.

This is arguably the first article to introduce the IoAuT paradigm and to identify its requirements and issues. We believe that substantial standardization efforts are needed to address the open issues in order to realize the true potential of the envisioned IoAuT. Just like for the general IoT field, the success of the IoAuT strongly relies on standardization requirements, which are currently unmet. The definition of standards for platforms, formats, protocols, and interfaces will allow for the achievement of interoperability between systems. Issues related to security and privacy of information, which are also common to the IoT, need to be addressed, especially for IoAuT systems deployed for the masses. In addition, research will need to address the challenge of how to design systems capable of supporting rich interaction paradigms that enable users to fully exploit the potentials and benefits of the IoAuT.

This article presented a vision for the IoAuT, highlighted its unique characteristics in contrast to the IoT, and identified the major challenges and requirements in order to realize it. The realization of the proposed IoAuT vision would ultimately benefit society, by providing a widespread use of ambient intelligence mechanisms involved to monitor environments in smart cities, as well as by offering new ways of interacting with sounds across the network (such as sound-based therapies involving remotely connected users).

We propose a roadmap for the implementation of the IoAuT vision.

1) To progress the design of Audio Things, with new solutions for the analysis of audio-related information based on the edge computing paradigm.
2) To advance the current connectivity infrastructure, with the implementation of novel interoperable protocols for the exchange of audio-related information.
3) To tackle the challenges of privacy and security of personal data, with a privacy by design approach.
4) To define standards and shared ontologies that will allow one to avoid fragmentation and facilitate interoperability among Audio Things as well as the services they offer.

It is hoped that the content of this article will stimulate discussions within the sound and music computing and IoT communities, so for the IoAuT to flourish.

References

[1] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey," *Comput. Netw.*, vol. 54, no. 15, pp. 2787–2805, 2010. [Online]. Available: https://doi.org/10.1016/j.comnet.2010.05.010

[2] D. Miorandi, S. Sicari, F. De Pellegrini, and I. Chlamtac, "Internet of Things: Vision, applications and research challenges," *Ad Hoc Netw.*, vol. 10, no. 7, pp. 1497–1516, 2012.

[3] E. Borgia, "The Internet of Things vision: Key features, applications and open issues," *Comput. Commun.*, vol. 54, pp. 1–31, Dec. 2014.

[4] J. Ardouin *et al.*, "An innovative low cost sensor for urban sound monitoring," in *Proc. INTER-NOISE NOISE-CON Congr. Conf.*, vol. 258, 2018, pp. 2226–2237.

[5] L. Turchet, "Interactive sonification and the IoT: The case of smart sonic shoes for clinical applications," in *Proc. Audio Mostly Conf.*, 2019, pp. 252–255.

[6] J. P. Bello *et al.*, "SONYC: A system for monitoring, analyzing, and mitigating urban noise pollution," *Commun. ACM*, vol. 62, no. 2, pp. 68–77, 2019. [Online]. Available: http://doi.acm.org/10.1145/3224204

[7] L. Turchet, C. Fischione, G. Essl, D. Keller, and M. Barthet, "Internet of musical things: Vision and challenges," *IEEE Access*, vol. 6, pp. 61994–62017, 2018. [Online]. Available: https://doi.org/10.1109/ACCESS.2018.2872625

[8] R. Harper, *Inside the Smart Home*. New York, NY, USA: Springer, 2006.

[9] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for smart cities," *IEEE Internet Things J.*, vol. 1, no. 1, pp. 22–32, Feb. 2014.

[10] G. Fazekas, Y. Raimond, K. Jakobson, and M. Sandler, "An overview of semantic Web activities in the OMRAS2 project," *J. New Music Res. Music Inf. OMRAS2 Project*, vol. 39, no. 4, pp. 295–311, 2011.

[11] G. Fazekas and T. Wilmering, "Semantic Web and semantic audio technologies," presented at the 132nd Convent. Audio Eng. Soc., Budapest, Hungary, 2012.

[12] Y. Rogers, H. Sharp, and J. Preece, *Interaction Design: Beyond Human–Computer Interaction*. New York, NY, USA: Wiley, 2011.

[13] C. Rowland, E. Goodman, M. Charlier, A. Light, and A. Lui, *Designing Connected Products: UX for the Consumer Internet of Things*. Sebastopol, U.K.: O'Reilly Media, 2015.

[14] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Trans. Multimedia*, vol. 17, no. 10, pp. 1733–1746, Oct. 2015.

[15] H. Boley and E. Chang, "Digital ecosystems: Principles and semantics," in *Proc. IEEE Int. Conf. Digit. Ecosyst. Technol.*, 2007, pp. 398–403.

[16] O. Mazhelis, E. Luoma, and H. Warma, "Defining an Internet-of-Things ecosystem," in *Internet of Things, Smart Spaces, and Next Generation Networking*. Berlin, Germany: Springer, 2012, pp. 1–14.

[17] D. Guinard, V. Trifa, F. Mattern, and E. Wilde, "From the Internet of Things to the Web of Things: Resource-oriented architecture and best practices," in *Architecting the Internet of Things*. Berlin, Germany: Springer, 2011, pp. 97–129.

[18] S. A. Alvi, B. Afzal, G. A. Shah, L. Atzori, and W. Mahmood, "Internet of multimedia things: Vision and challenges," *Ad Hoc Netw.*, vol. 33, pp. 87–111, Oct. 2015.

[19] S. Skach, A. Xambó, L. Turchet, A. Stolfi, R. Stewart, and M. Barthet, "Embodied interactions with e-textiles and the Internet of sounds for performing arts," in *Proc. ACM Int. Conf. Tangible Embedded Embodied Interact.*, 2018, pp. 80–87.

[20] T. Hermann, A. Hunt, and J. G. Neuhoff, *The Sonification Handbook*. Berlin, Germany: Logos Verlag, 2011.

[21] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, "Context aware computing for the Internet of Things: A survey," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 414–454, 1st Quart., 2014.

[22] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *Proc. 18th IEEE Symp. Commun. Veh. Technol. Benelux (SCVT)*, 2011, pp. 1–6.

[23] E. T. Nykaza *et al.*, "A framework for providing real-time feedback of environmental noise levels over large areas," *J. Acoust. Soc. America*, vol. 140, no. 4, p. 3193, 2016.

[24] A. McPherson and V. Zappi, "An environment for submillisecond-latency audio and sensor processing on BeagleBone black," in *Proc. 138th Audio Eng. Soc. Convent.*, 2015, pp. 1–7. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=17755

[25] B. da Silva, A. W. Happi, A. Braeken, and A. Touhafi, "Evaluation of classical machine learning techniques towards urban sound recognition on embedded systems," *Appl. Sci.*, vol. 9, no. 18, p. 3885, 2019.

[26] F. Gontier, M. Lagrange, P. Aumond, A. Can, and C. Lavandier, "An efficient audio coding scheme for quantitative and qualitative large scale acoustic monitoring using the sensor grid approach," *Sensors*, vol. 17, no. 12, p. 2758, 2017.

[27] A. F. Smeaton and M. McHugh, "Towards event detection in an audio-based sensor network," in *Proc. 3rd ACM Int. Workshop Video Surveillance Sensor Netw.*, 2005, pp. 87–94.

[28] B. Malhotra, I. Nikolaidis, and J. Harms, "Distributed classification of acoustic targets in wireless audio-sensor networks," *Comput. Netw.*, vol. 52, no. 13, pp. 2582–2593, 2008.

[29] Á. Lédeczi, T. Hay, P. Volgyesi, D. R. Hay, A. Nádas, and S. Jayaraman, "Wireless acoustic emission sensor network for structural monitoring," *IEEE Sens. J.*, vol. 9, no. 11, pp. 1370–1377, Nov. 2009.

[30] J. P. Bello, C. Mydlarz, and J. Salamon, "Sound analysis in smart cities," in *Computational Analysis of Sound Scenes and Events*. Cham, Switzerland: Springer, 2018, pp. 373–397.

[31] J. Segura-Garcia, S. Felici-Castell, J. J. Perez-Solano, M. Cobos, and J. M. Navarro, "Low-cost alternatives for urban noise nuisance monitoring using wireless sensor networks," *IEEE Sens. J.*, vol. 15, no. 2, pp. 836–844, Feb. 2015.

[32] G. Kokkonis, K. E. Psannis, M. Roumeliotis, and D. Schonfeld, "Real-time wireless multisensory smart surveillance with 3D-HEVC streams for Internet-of-Things (IoT)," *J. Supercomput.*, vol. 73, no. 3, pp. 1044–1062, 2017.

[33] M. Antonini, M. Vecchio, F. Antonelli, P. Ducange, and C. Perera, "Smart audio sensors in the Internet of Things edge for anomaly detection," *IEEE Access*, vol. 6, pp. 67594–67610, 2018.

[34] S. S. Sethi, R. M. Ewers, N. S. Jones, C. D. L. Orme, and L. Picinali, "Robust, real-time and autonomous monitoring of ecosystems with an open, low-cost, networked device," *Methods Ecol. Evol.*, vol. 9, no. 12, pp. 2383–2387, 2018.

[35] J. Sueur and A. Farina, "Ecoacoustics: The ecological investigation and interpretation of environmental sound," *Biosemiotics*, vol. 8, no. 3, pp. 493–502, 2015.

[36] M. Towsey, J. Wimmer, I. Williamson, and P. Roe, "The use of acoustic indices to determine avian species richness in audio-recordings of the environment," *Ecol. Informat.*, vol. 21, pp. 110–119, May 2014.

[37] C. Mydlarz, C. Shamoon, and J. P. Bello, "Noise monitoring and enforcement in New York city using a remote acoustic sensor network," in *Proc. INTER-NOISE NOISE-CON Congr. Conf.*, vol. 255, 2017, pp. 5509–5520.

[38] A. Jakob *et al.*, "A distributed sensor network for monitoring noise level and noise sources in urban environments," in *Proc. IEEE 6th Int. Conf. Future Internet Things Cloud (FiCloud)*, 2018, pp. 318–324.

[39] J. Abeßer *et al.*, "Urban noise monitoring in the Stadtlärm project—A field report," in *Proc. Workshop Detect. Classification Acoust. Scenes Events (DCASE)*, 2019, pp. 1–4.

[40] P. Bellucci, L. Peruzzi, and G. Zambon, "LIFE DYNAMAP project: The case study of Rome," *Appl. Acoust.*, vol. 117, pp. 193–206, Feb. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X1630113X

[41] J. Picaut *et al.*, "Cense project: Characterization of urban sound environments using a comprehensive approach combining open data, measurements and modeling," in *Proc. 8th Forum Acusticum 173rd Meeting Acoust. Soc. America*, 2017, pp. 1–12.

[42] P. Luquet, "Method for the objective description of an acoustic environment based on short LEQ values," *Appl. Acoust.*, vol. 15, no. 2, pp. 147–156, 1982.

[43] C. Mydlarz, S. Nacach, E. Rosenthal, M. Temple, T. H. Park, and A. Roginska, "The implementation of MEMS microphones for urban sound sensing," in *Proc. 137th Audio Eng. Soc. Convent.*, Los Angeles, CA, USA, 2014, pp. 740–748.

[44] V. Risojeviä, R. Rozman, R. Pilipović, R. Češnovar, and P. Bulić, "Accurate indoor sound level measurement on a low-power and low-cost wireless sensor node," *Sensors*, vol. 18, no. 7, p. 2351, Jul. 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6068900/

[45] C. Mydlarz, J. Salamon, and J. P. Bello, "The implementation of low-cost urban acoustic monitoring devices," *Appl. Acoust.*, vol. 117, pp. 207–218, Feb. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X1630158X

[46] D. Offenhuber, S. Auinger, S. Seitinger, and R. Muijs, "Los Angeles noise array—Planning and design lessons from a noise sensing network," *Environ. Plan. B Urban Anal. City Sci.*, vol. 47, no. 4, pp. 609–625, Aug. 2018. [Online]. Available: https://doi.org/10.1177/2399808318792901

[47] R. M. Alsina-Pagès, U. Hernandez-Jayo, F. Alias, and I. Angulo, "Design of a mobile low-cost sensor network using urban buses for real-time ubiquitous noise monitoring," *Sensors*, vol. 17, no. 1, p. 57, Dec. 2016. [Online]. Available: http://www.mdpi.com/1424-8220/17/1/57

[48] J. Zuo, H. Xia, S. Liu, and Y. Qiao, "Mapping urban environmental noise using smartphones," *Sensors*, vol. 16, no. 10, p. 1692, Oct. 2016. [Online]. Available: http://www.mdpi.com/1424-8220/16/10/1692

[49] P. Duda, "Processing and unification of environmental noise data from road traffic with spatial dimension collected through mobile phones," *J. Geosci. Environ. Protect.*, vol. 4, no. 13, p. 1, Dec. 2016. [Online]. Available: http://www.scirp.org/journal/PaperInformation.aspx?PaperID=73038&#abstract

[50] P. Aumond *et al.*, "A study of the accuracy of mobile technology for measuring urban noise pollution in large scale participatory sensing campaigns," *Appl. Acoust.*, vol. 117, pp. 219–226, Feb. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X16302055

[51] E. Murphy and E. A. King, "Testing the accuracy of smartphones and sound level meter applications for measuring environmental noise," *Appl. Acoust.*, vol. 106, pp. 16–22, May 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0003682X15003667

[52] W. Zamora, C. T. Calafate, J.-C. Cano, and P. Manzoni, "Accurate ambient noise assessment using smartphones," *Sensors*, vol. 17, no. 4, p. 917, Apr. 2017. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5426841/

[53] E. D'Hondt, M. Stevens, and A. Jacobs, "Participatory noise mapping works! An evaluation of participatory sensing as an alternative to standard techniques for environmental monitoring," *Pervasive Mobile Comput.*, vol. 9, no. 5, pp. 681–694, Oct. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1574119212001137

[54] L. Miller *et al.*, "Environmental noise mapping with smartphone applications: A participatory noise map of West Hartford, CT," in *Proc. INTER-NOISE NOISE-CON Congr. Conf.*, vol. 252, Jun. 2016, pp. 445–451.

[55] A. Longo, M. Zappatore, M. Bochicchio, and S. B. Navathe, "Crowd-sourced data collection for urban monitoring via mobile sensors," *ACM Trans. Internet Technol.*, vol. 18, no. 1, pp. 1–21, Oct. 2017. [Online]. Available: http://doi.acm.org/10.1145/3093895

[56] G. Guillaume *et al.*, "Noise mapping based on participative measurements," *Noise Map.*, vol. 3, no. 1, pp. 140–156, 2016. [Online]. Available: http://www.degruyter.com/view/j/noise.2016.3.issue-1/noise-2016-0011/noise-2016-0011.xml?format=INT

[57] D. R. Nast, W. S. Speer, and C. G. L. Prell, "Sound level measurements using smartphone 'apps': Useful or inaccurate?" *Noise Health*, vol. 16, no. 72, p. 251, Jan. 2014. [Online]. Available: http://www.noiseandhealth.org/article.asp?issn=1463-1741;year=2014;volume=16;issue=72;spage=251;epage=256;aulast=Nast;type=0

[58] M. Celestina, J. Hrovat, and C. A. Kardous, "Smartphone-based sound level measurement apps: Evaluation of compliance with international sound level meter standards," *Appl. Acoust.*, vol. 139, pp. 119–128, Oct. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0003682X17309945

[59] J. Picaut, N. Fortin, E. Bocher, G. Petit, P. Aumond, and G. Guillaume, "An open-science crowdsourcing approach for producing community noise maps using smartphones," *Build. Environ.*, vol. 148, pp. 20–33, Jan. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360132318306747

[60] E. Bocher, G. Petit, J. Picaut, N. Fortin, and G. Guillaume, "Collaborative noise data collected from smartphones," *Data Brief*, vol. 14, pp. 498–503, Oct. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2352340917303414

[61] S. Grubesa, A. Petosic, M. Suhanek, and I. Durek, "Mobile crowd-sensing accuracy for noise mapping in smart cities," *Automatica*, vol. 59, nos. 3–4, pp. 286–293, Oct. 2018. [Online]. Available: https://doi.org/10.1080/00051144.2018.1534927

[62] A. Truskinger, H. Yang, J. Wimmer, J. Zhang, I. Williamson, and P. Roe, "Large scale participatory acoustic sensor data analysis: Tools and reputation models to enhance effectiveness," in *Proc. IEEE 7th Int. Conf. e-Sci.*, 2011, pp. 150–157.

[63] C. A. Kardous and P. B. Shaw, "Evaluation of smartphone sound measurement applications," *J. Acoust. Soc. America*, vol. 135, no. 4, pp. EL186–EL192, Mar. 2014. [Online]. Available: http://asa.scitation.org/doi/full/10.1121/1.4865269

[64] C. A. Kardous and P. B. Shaw, "Evaluation of smartphone sound measurement applications (apps) using external microphones—A follow-up study," *J. Acoust. Soc. America*, vol. 140, no. 4, pp. EL327–EL333, Oct. 2016. [Online]. Available: http://asa.scitation.org/doi/full/10.1121/1.4964639

[65] B. Roberts, C. Kardous, and R. Neitzel, "Improving the accuracy of smart devices to measure noise exposure," *J. Occupational Environ. Hygiene*, vol. 13, no. 11, pp. 840–846, Nov. 2016. [Online]. Available: http://oeh.tandfonline.com/doi/full/10.1080/15459624.2016.1183014

[66] R. Ventura, V. Mallet, V. Issarny, P.-G. Raverdy, and F. Rebhi, "Evaluation and calibration of mobile phones for noise monitoring application," *J. Acoust. Soc. America*, vol. 142, no. 5, pp. 3084–3093, Nov. 2017. [Online]. Available: https://asa.scitation.org/doi/abs/10.1121/1.5009448

[67] A. Deleforge, D. Di Carlo, M. Strauss, R. Serizel, and L. Marcenaro, "Audio-based search and rescue with a drone: Highlights from the ieee signal processing cup 2019 student competition [SP competitions]," *IEEE Signal Process. Mag.*, vol. 36, no. 5, pp. 138–144, Sep. 2019.

[68] Y. Fu, M. Kinniry, and L. N. Kloepper, "The chirocopter: A UAV for recording sound and video of bats at altitude," *Methods Ecol. Evol.*, vol. 9, no. 6, pp. 1531–1535, 2018.

[69] J. Socoró, F. Alías, and R. Alsina-Pagès, "An anomalous noise events detector for dynamic road traffic noise mapping in real-life urban and suburban environments," *Sensors*, vol. 17, no. 10, p. 2323, 2017.

[70] F. Miranda *et al.*, "Urban pulse: Capturing the rhythm of cities," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 1, pp. 791–800, Jan. 2017.

[71] J. M. Navarro, J. B. Tomas-Gabarron, and J. Escolano, "A big data framework for urban noise analysis and management in smart cities," *Acta Acustica United Acustica*, vol. 103, no. 4, pp. 552–560, Jul. 2017.

[72] M. Zappatore, A. Longo, and M. A. Bochicchio, "Crowd-sensing our smart cities: A platform for noise monitoring and acoustic urban planning," *J. Commun. Softw. Syst.*, vol. 13, no. 2, pp. 53–67, Jun. 2017. [Online]. Available: https://jcomss.fesb.unist.hr/index.php/jcomss/article/view/373

[73] B. Bach, P. Dragicevic, D. Archambault, C. Hurter, and S. Carpendale, "A review of temporal data visualizations based on space-time cube operations," in *Proc. Eurograph. Conf. Visual. (EuroVis)*, 2014, pp. 23–41.

[74] X. Sheng and Y.-H. Hu, "Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 53, no. 1, pp. 44–53, Jan. 2005.

[75] J. Lim, J. Lee, S. Hong, and P. Park, "Algorithm for detection with localization of multi-targets in wireless acoustic sensor networks," in *Proc. 18th IEEE Int. Conf. Tools Artif. Intell. (ICTAI)*, 2006, pp. 547–554.

[76] Y. Guo and M. Hazas, "Acoustic source localization of everyday sounds using wireless sensor networks," in *Proc. 12th ACM Int. Conf. Adjunct Papers Ubiquitous Comput.*, 2010, pp. 411–412.

[77] A. Griffin and A. Mouchtaris, "Localizing multiple audio sources from doa estimates in a wireless acoustic sensor network," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2013, pp. 1–4.

[78] M. Cobos, J. J. Perez-Solano, S. Felici-Castell, J. Segura, and J. M. Navarro, "Cumulative-sum-based localization of sound events in low-cost wireless acoustic sensor networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 1792–1802, Dec. 2014.

[79] J. Zhang, R. Heusdens, and R. C. Hendriks, "Relative acoustic transfer function estimation in wireless acoustic sensor networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 10, pp. 1507–1519, Oct. 2019.

[80] J. A. Belloch, J. M. Badía, F. D. Igual, and M. Cobos, "Practical considerations for acoustic source localization in the IoT era: Platforms, energy efficiency and performance," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 5068–5079, Jun. 2019.

[81] G. Kramer *et al.*, *The Sonification Report: Status of the Field and Research Agenda. Report Prepared for the National Science Foundation by Members of the International Community for Auditory Display*, Int. Community Auditory Display, Santa Fe, NM, USA 1999.

[82] D. Lockton, F. Bowden, C. Brass, and R. Gheerawo, "PowerChord: Towards ambient appliance-level electricity use feedback through real-time sonification," in *Proc. Int. Conf. Ubiquitous Comput. Ambient Intell.*, 2014, pp. 48–51.

[83] M. Iber, P. Lechner, C. Jandl, M. Mader, and M. Reichmann, "Auditory augmented reality for cyber physical production systems," in *Proc. Audio Mostly Conf.*, 2019, pp. 53–60.

[84] B. B. Bederson, "Audio augmented reality: A prototype automated tour guide," in *Proc. ACM Conf. Companion Human Factors Comput. Syst.*, 1995, pp. 210–211.

[85] J. Gómez, B. Oviedo, and E. Zhuma, "Patient monitoring system based on Internet of Things," *Procedia Comput. Sci.*, vol. 83, pp. 90–97, May2016.

[86] R. Altilio, L. Liparulo, M. Panella, A. Proietti, and M. Paoloni, "Multimedia and gaming technologies for telerehabilitation of motor disabilities [leading edge]," *IEEE Technol. Soc. Mag.*, vol. 34, no. 4, pp. 23–30, Dec. 2015.

[87] T. Hermann and A. Hunt, "Guest editors' introduction: An introduction to interactive sonification," *IEEE Multimedia*, vol. 12, no. 2, pp. 20–24, Apr.–Jun. 2005.

[88] L. Turchet, "Footstep sounds synthesis: Design, implementation, and evaluation of foot-floor interactions, surface materials, shoe types, and walkers' features," *Appl. Acoust.*, vol. 107, pp. 46–68, Jun. 2016.

[89] S. D. Bella *et al.*, "Gait improvement via rhythmic stimulation in parkinson's disease is linked to rhythmic skills," *Sci. Rep.*, vol. 7, Feb. 2017, Art. no. 42005.

[90] L. Turchet, S. Serafin, and P. Cesari, "Walking pace affected by interactive sounds simulating stepping on different terrains," *ACM Trans. Appl. Perception*, vol. 10, no. 4, pp. 1–14, 2013.

[91] M. Rodger, W. Young, and C. Craig, "Synthesis of walking sounds for alleviating gait disturbances in Parkinson's disease," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 22, no. 3, pp. 543–548, May 2013.

[92] A. Tajadura-Jiménez, M. Basia, O. Deroy, M. Fairhurst, N. Marquardt, and N. Bianchi-Berthouze, "As light as your footsteps: Altering walking sounds to change perceived body weight, emotional state and gait," in *Proc. 33rd Annu. ACM Conf. Human Factors Comput. Syst.*, 2015, pp. 2943–2952.

[93] K. P. Yeo, S. Nanayakkara, and S. Ransiri, "StickEar: Making everyday objects respond to sound," in *Proc. 26th Annu. ACM Symp. User Interface Softw. Technol.*, 2013, pp. 221–226.

[94] A. Mehrabi, A. Mazzoni, D. Jones, and A. Steed, "Evaluating the user experience of acoustic data transmission," *Pers. Ubiquitous Comput. J.*, to be published.

[95] G. Fazekas and M. Sandler, "Knowledge representation issues in audio-related metadata model design," in *Proc. 133rd Convention Audio Eng. Soc.*, San Francisco, CA, USA, 2012, pp. 1–12.

[96] I. Horrocks, "Ontologies and the semantic Web," *Commun. ACM*, vol. 51, no. 12, pp. 58–67, 2008.

[97] F. Font *et al.*, "Audio commons: Bringing creative commons audio content to the creative industries," in *Proc. 61st Int. Conf. Audio Games Audio Eng. Soc. Conf.*, 2016, pp. 1–7.

[98] A. Xambo, F. Font, G. Fazekas, and M. Barthet, "Leveraging online audio commons content for media production," *Foundations in Sound Design for Linear Media: An Interdisciplinary Approach*, M. Filimowicz, Ed. New York, NY, USA: Routledge, 2019, pp. 248–282.

[99] A. Xambo, G. Roma, A. Lerch, M. Barthet, and G. Fazekas, "Live repurposing of sounds: MIR explorations with personal and crowd-sourced databases," in *Proc. New Interfaces Musical Exp. (NIME)*, Jun. 2019, pp. 364–369.

[100] F. Viola, L. Turchet, and G. Antoniazzi, and F. Fazekas, "C Minor: A semantic publish/subscribe broker for the Internet of musical things," in *Proc. IEEE Conf. Open Innov. Assoc. (FRUCT)*, 2018, pp. 405–415. [Online]. Available: https://ieeexplore.ieee.org/document/8588087

[101] T. Wilmering, F. Thalmann, G. Fazekas, and M. Sandler, "Bridging fan communities and facilitating access to music archives through semantic audio applications," in *Proc. 43rd Convention Audio Eng. Soc.*, Oct. 2017, p. 387.

[102] F. Viola, A. Stolfi, A. Milo, M. Ceriani, M. Barthet, and G. Fazekas, "Playsound.space: Enhancing a live performance tool with semantic recommendations," in *Proc. ACM 1st SAAM Workshop*, 2018, pp. 46–53.

[103] M. Ceriani and G. Fazekas, "Audio commons ontology: A data model for an audio content ecosystem," in *Proc. Int. Semantic Web Conf.*, 2018, pp. 20–35.

[104] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "The effects of noisy labels on deep convolutional neural networks for music tagging," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 2, no. 2, pp. 139–149, Apr. 2018.

[105] Y. Hou, Q. Kong, J. Wang, and S. Li, "Polyphonic audio tagging with sequentially labelled data using CRNN with learnable gated linear units," in *Proc. Workshop Detect. Classification Acoust. Scenes Events (DCASE)*, 2018, pp. 1–5. [Online]. Available: http://epubs.surrey.ac.uk/849618/

[106] D. Stowell, M. D. Wood, H. Pamuła, Y. Stylianou, and H. Glotin, "Automatic acoustic detection of birds through deep learning: The first bird audio detection challenge," *Methods Ecol. Evol.*, vol. 10, no. 3, pp. 368–380, 2019.

[107] A. Allik, G. Fazekas, and M. Sandler, "An ontology for audio features," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2016, pp. 73–79.

[108] T. Wilmering, G. Fazekas, and M. Sandler, *AUFX-O: Novel Methods for the Representation of Audio Processing Workflows* (LNCS 9982). Cham, Switzerland: Springer, 2016, pp. 229–237.

[109] F. Thalmann, A. Carrillo, G. Fazekas, G. A. Wiggins, and M. Sandler, "The mobile audio ontology: Experiencing dynamic music objects on mobile devices," in *Proc. IEEE Int. Conf. Semantic Comput. (ICSC)*, Laguna Hills, CA, USA, Feb. 2016, pp. 47–54.

[110] G. Fazekas and M. Sandler, "The studio ontology framework," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2011, pp. 24–28.

[111] B. Smus, *Web Audio API: Advanced Sound for Games and Interactive Apps*. Sebastopol, U.K.: O'Reilly Media, 2013.

[112] B. Matuszewski and F. Bevilacqua, "Toward a Web of audio things," in *Proc. Sound Music Comput. Conf.*, 2018, pp. 1–8.

[113] J. Lambert, S. Robaszkiewicz, and N. Schnell, "Synchronisation for distributed audio rendering over heterogeneous devices, in HTML5," in *Proc. Web Audio Conf.*, 2016, pp. 1–6.

[114] C. Roberts and J. Kuchera-Morin, "GIBBER: Live coding audio in the browser," in *Proc. Int. Comput. Music Conf.*, 2012, pp. 64–69.

[115] C. Roberts, G. Wakefield, and M. Wright, "The Web browser as synthesizer and interface." in *Proc. Int. Conf. New Interfaces Musical Exp.*, 2013, pp. 313–318.

[116] C. B. Clark and A. Tindale, "Flocking: A framework for declarative music-making on the Web," in *Proc. Int. Comput. Music Conf.*, 2014, pp. 1550–1557.

[117] P. Bahadoran, A. Benito, T. Vassallo, and J. D. Reiss, "FXIVE: A Web platform for procedural sound synthesis," in *Proc. 144th Audio Eng. Soc. Convent.*, 2018, pp. 1–5.

[118] D. Rocchesso and F. Fontana, Eds., *The Sounding Object*. Florence, Italy: Edizioni Mondo Estremo, 2003.

[119] R. van Kranenburg and A. Bassi, "IoT challenges," *Commun. Mobile Comput.*, vol. 1, no. 1, p. 9, 2012.

[120] X. Jiang *et al.*, "Low-latency networking: Where latency lurks and how to tame it," *Proc. IEEE*, vol. 107, no. 2, pp. 280–306, Feb. 2019.

[121] H. Shokri-Ghadikolaei, C. Fischione, P. Popovski, and M. Zorzi, "Design aspects of short-range millimeter-wave networks: A MAC layer perspective," *IEEE Netw.*, vol. 30, no. 3, pp. 88–96, May 2016.

[122] J. Jagannath, N. Polosky, A. Jagannath, F. Restuccia, and T. Melodia. (2019). *Machine Learning for Wireless Communications in the Internet of Things: A Comprehensive Survey*. [Online]. Available: http://arxiv.org/abs/1901.07947

[123] A. Silver *et al.*, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[124] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan, and M. Zorzi, "Towards 6G networks: Use cases and technologies," Mar. 2019. [Online]. Available: arXiv:1903.12216.

[125] N. Shrestha, S. Kubler, and K. Främling, "Standardized framework for integrating domain-specific applications into the IoT," in *Proc. IEEE Int. Conf. Future Internet Things Cloud*, 2014, pp. 124–131.

[126] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic Web," *Sci. Amer.*, vol. 284, no. 5, pp. 34–43, 2001.

[127] J. Sowa, *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, vol. 13. Pacific Grove, CA, USA: Brooks/Cole, 2000.

[128] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 16–34, Jan. 2016.

[129] M. Valenti, A. Diment, G. Parascandolo, S. Squartini, and T. Virtanen, "DCASE 2016 acoustic scene classification using convolutional neural networks," in *Proc. Workshop Detection Classification Acoust. Scenes Events*, 2016, pp. 95–99.

[130] J. F. Gemmeke *et al.*, "An exemplar-based NMF approach to audio event detection," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2013, pp. 1–4.

[131] M. Espi, M. Fujimoto, K. Kinoshita, and T. Nakatani, "Exploiting spectro-temporal locality in deep learning based acoustic event detection," *EURASIP J. Audio Speech Music Process.*, vol. 2015, p. 26, 2015. [Online]. Available: https://doi.org/10.1186/s13636-015-0069-2

[132] N. Takahashi, M. Gygli, B. Pfister, and L. Van Gool, "Deep convolutional neural networks and data augmentation for acoustic event detection," 2016. [Online]. Available: arXiv:1604.07160.

[133] Q. Le *et al.*, "Building high-level features using large scale unsupervised learning," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2013, pp. 8595–8598.

[134] E. Fonseca, M. Plakal, D. P. Ellis, F. Font, X. Favory, and X. Serra, "Learning sound event classifiers from Web audio with noisy labels," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2019, pp. 21–25.

[135] Q. Kong, C. Yu, Y. Xu, T. Iqbal, W. Wang, and M. D. Plumbley, "Weakly labelled audioset tagging with attention neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process*, vol. 27, no. 11, pp. 1791–1802, Mar. 2019.

[136] D. Bahdanau, K. Cho, and Y. Bengio., "Neural machine translation by jointly learning to align and translate," 2014. [Online]. Available: arXiv:1409.0473.

[137] G. Psuj, "Multi-sensor data integration using deep learning for characterization of defects in steel elements," *Sensors*, vol. 18, no. 2, pp. 292–307, 2018.

[138] D. Sheng and G. Fazekas, "A feature learning Siamese model for intelligent control of the dynamic range compressor," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.

[139] E. Fonseca *et al.*, "Freesound datasets: A platform for the creation of open audio datasets," in *Proc. 18th Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, Suzhou, China, 2017, pp. 486–493.

[140] E. Fonseca *et al.*, "General-purpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline," in *Proc. Workshop Detect. Classification Acoust. Scenes Events (DCASE)*, 2018, pp. 1–5.

[141] J. F. Gemmeke *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2017, pp. 776–780.

[142] *Awesome Deep Vision*. Accessed: Jan. 14, 2020. [Online]. Available: https://github.com/kjw0612/awesome-deep-vision

[143] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[144] H. Soltau, H. Liao, and H. Sak, "Neural speech recognizer: Acoustic-to-word LSTM model for large vocabulary speech recognition," 2016. [Online]. Available: arXiv:1610.09975.
[145] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu, "Video paragraph captioning using hierarchical recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4584–4593.
[146] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017. [Online]. Available: http://doi.acm.org/10.1145/3065386
[147] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–13.
[148] J. Park, S. Samarakoon, M. Bennis, and M. Debbah. (2018). *Wireless Network Intelligence at the Edge*. [Online]. Available: http://arxiv.org/abs/1812.02858
[149] W. Zhu, C. Luo, J. Wang, and S. Li, "Multimedia cloud computing," *IEEE Signal Process. Mag.*, vol. 28, no. 3, pp. 59–69, May 2011.
[150] H. van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double qlearning," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 95–108.
[151] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50× fewer parameters and <0.5MB model size," 2016. [Online]. Available: arXiv:1602.07360.
[152] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1135–1143.
[153] R. Banner, Y. Nahshan, and D. Soudry, "Post training 4-bit quantization of convolutional networks for rapid-deployment," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 7948–7956.
[154] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016, pp. 1–14.
[155] R. Krishnamoorthi, "Quantizing deep convolutional networks for efficient inference: A whitepaper," 2018. [Online]. Available: arXiv:1806.08342.
[156] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet Things J.*, vol. 3, no. 5, pp. 637–646, Jun. 2016.
[157] S. K. Sharma and X. Wang, "Live data analytics with collaborative edge and cloud processing in wireless IoT networks," *IEEE Access*, vol. 5, pp. 4621–4635, 2017.
[158] A. Hildebrand, *AES Standard for Audio Applications of Networks-High-Performance Streaming Audio-Over-IP Interoperability*, AES Standard 67-2013, 2018.
[159] A. Holzinger and A. Hildebrand, "Realtime linear audio distribution over networks: A comparison of layer 2 and 3 solutions using the example of Ethernet AVB and RAVENNA," in *Proc. 44th Int. Conf. Audio Netw. Audio Eng. Soc. Conf.*, 2011.
[160] J.-S. Sheu, H.-N. Shou, and W.-J. Lin, "Realization of an Ethernet-based synchronous audio playback system," *Multimedia Tools Appl.*, vol. 75, no. 16, pp. 9797–9818, 2016.
[161] D. Trueman, P. Cook, S. Smallwood, and G. Wang, "PLORK: The Princeton laptop orchestra, year 1," in *Proc. Int. Comput. Music Conf.*, 2006, pp. 1–8.
[162] G. Wang, N. J. Bryan, J. Oh, and R. Hamilton, "Stanford laptop orchestra (SLORK)," in *Proc. ICMC*, 2009, pp. 1–4.
[163] J. J. Arango and D. M. Giraldo, "The smartphone ensemble. Exploring mobile computer mediation in collaborative musical performance," in *Proc. Int. Conf. New Interfaces Musical Exp.*, vol. 16, 2016, pp. 61–64.
[164] N. Schnell, V. Saiz, K. Barkati, and S. Goldszmidt, "Of time engines and masters an API for scheduling and synchronizing the generation and playback of event sequences and media streams for the Web audio API," in *Proc. Web Audio Conf.*, 2015, pp. 1–5.
[165] D. L. Mills, "Internet time synchronization: The network time protocol," *IEEE Trans. Commun.*, vol. 39, no. 10, pp. 1482–1493, Oct. 1991.
[166] J. E. Elson, *Time Synchronization in Wireless Sensor Networks*, Univ. California at Los Angeles, Los Angeles, CA, USA, 2003.
[167] K. S. Yildirim and A. Kantarci, "Time synchronization based on slow-flooding in wireless sensor networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 1, pp. 244–253, Jan. 2014.
[168] K. S. Yıldırım and Ö. Gürcan, "Efficient time synchronization in a wireless sensor network by adaptive value tracking," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3650–3664, Apr. 2014.
[169] R. Oda and R. Fiebrink, "The global metronome: Absolute tempo sync for networked musical performance," in *Proc. Conf. New Interfaces Musical Exp.*, 2016, pp. 26–31.
[170] R. Weber, "Internet of Things: Privacy issues revisited," *Comput. Law Security Rev.*, vol. 31, no. 5, pp. 618–627, 2015.
[171] A. Persson and I. Kavathatzopoulos. (2019). *How to Make Decisions With Algorithms: Ethical Decision-Making Using Algorithms Within Predictive Analytics*. [Online]. Available: https://doi.org/10.29297/orbit.v1i2.44
[172] Center for Data Ethics and Innovation. (2019). *Smart Speakers and Voice Assistants*. [Online]. Available: www.gov.uk/cdei
[173] B. Dainow. (2019). *Smart City Transcendent: Understanding the Smart City by Transcending Ontology*. [Online]. Available: https://doi.org/10.29297/orbit.v1i1.27
[174] R. Roman, P. Najera, and J. Lopez, "Securing the Internet of Things," *Computer*, vol. 44, no. 9, pp. 51–58, 2011.
[175] A. Whitmore, A. Agarwal, and L. Da Xu, "The Internet of Things—A survey of topics and trends," *Inf. Syst. Front.*, vol. 17, no. 2, pp. 261–274, 2015.
[176] A. Cohen-Hadria, M. Cartwright, B. McFee, and J. Bello, "Voice anonymization in urban sounds recordings," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, 2019, pp. 13–16.

Luca Turchet received the master's degrees (*summa cum laude*) in computer science from the University of Verona, Verona, Italy, in 2006, in classical guitar and composition from Music Conservatory of Verona, Verona, in 2007 and 2009, respectively, and in electronic music from the Royal College of Music of Stockholm, Stockholm, Sweden, in 2015, and the Ph.D. degree in media technology from Aalborg University Copenhagen, Copenhagen, Denmark, in 2013.

He is an Assistant Professor with the Department of Information Engineering and Computer Science, University of Trento, Trento, Italy. His scientific, artistic, and entrepreneurial research has been supported by numerous grants from different funding agencies, including the European Commission, the European Institute of Innovation and Technology, the Italian Minister of Foreign Affairs, and the Danish Research Council. He is the Co-Founder and the Head of Sound and Interaction Design at Elk. His main research interests are in music technology, Internet of Things, human–computer interaction, and multimodal perception.

György Fazekas (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering.

He is a Senior Lecturer (Assistant Professor) with the Center for Digital Music, Queen Mary University of London, London, U.K. He is an Investigator of UKRI's £6.5 million Centre for Doctoral Training in Artificial Intelligence and Music (AIM CDT) and he was QMUL's Principal Investigator on the H2020 funded Audio Commons project. He published over 130 papers in the fields of music information retrieval, semantic Web, deep learning, and semantic audio.

Dr. Fazekas received the Citation Award of the AES. He was the General Chair of ACM's Audio Mostly 2017 and papers Co-Chair of the AES 53rd International Conference on Semantic Audio.

Mathieu Lagrange (Member, IEEE) received the Ph.D. degree in computer science from the University of Bordeaux, Bordeaux, France, in 2004,

He is a CNRS Research Scientist with LS2N, Nantes, France, a French Laboratory dedicated to cybernetics. He visited several institutions, such as the University of Victoria, Victoria, BC, Canada; McGill University, Montreal, QC, Canada; Orange Labs, Paris, France; TELECOM ParisTech, Paris; and Ircam, Paris. He co-organized two editions of the Detection and Classification of Acoustic Scenes and Events Challenge with event detection tasks and is involved in the development of acoustic sensor networks for urban acoustic quality monitoring. His research focuses on machine listening algorithms applied to the analysis of musical and environmental audio.

Hossein S. Ghadikolaei (Member, IEEE) received the B.Sc. degree in electrical engineering from Iran University of Science and Technology, Tehran, Iran, in 2009, the M.Sc. degree in electrical engineering from the Sharif University of Technology, Tehran, in 2011, and the Ph.D. degree in electrical engineering and computer science from the KTH Royal Institute of Technology, Stockholm, Sweden, in 2018.

He is currently a Research Scientist with EPFL, Lausanne, Switzerland. His research interests include distributed optimization and machine learning, with applications in data science and networking.

Dr. Ghadikolaei was a recipient of the IEEE Communications Society Stephen O. Rice Prize in 2018, the Premium Award for the Best Paper in IET Communications in 2014, the Program of Excellence Award from KTH in 2013, and the Best Paper Award from the Iranian Student Conference of Electrical Engineering in 2011. He was selected as an Exemplary Reviewer for the IEEE TRANSACTIONS ON COMMUNICATIONS in 2017 and 2018.

Carlo Fischione (Senior Member, IEEE) received the Ph.D. degree in electrical and information engineering and the Laurea degree *(summa cum laude)* in electronic engineering from the University of L'Aquila, L'Aquila, Italy.

He is a Full Professor with the KTH Royal Institute of Technology, Stockholm, Sweden. He has held research positions with Massachusetts Institute of Technology, Cambridge, MA, USA (Visiting Professor); Harvard University, Cambridge (Associate Professor); and the University of California at Berkeley, Berkeley, CA, USA (Visiting Scholar and Research Associate). His research interests include optimization with applications to networks, and wireless and sensor networks.

Prof. Fischione received a number of awards, including the IEEE Communication Society "Stephen O. Rice" Award for the best IEEE TRANSACTION ON COMMUNICATIONS in 2015, the Best Paper Award from the IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, the Best Paper Awards at the IEEE International Conference on Mobile Ad-Hoc and Sensor System in 2005 and 2009, the Best Paper Award of the IEEE Sweden VT-COM-IT Chapter, the Best Business Idea Awards from VentureCup East Sweden and from Stockholm Innovation and Growth Life Science in Sweden, and the Junior Research Award from Swedish Research Council. He is an Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS and an Associated Editor of *IFAC Automatica*. He is the Co-Funder and the Scientific Director of MIND Music Labs.