Turchet, Luca; Lagrange, Mathieu; Rottondi, Cristina; Fazekas, György; Peters, Nils; Østergaard, Jan; Font, Frederic; Bäckström, Tom; Fischione, Carlo

The Internet of Sounds: Convergent Trends, Insights and Future Directions

Published in:
IEEE Internet of Things Journal

DOI:
10.1109/JIOT.2023.3253602

Published: 01/07/2023

Document Version
Publisher's PDF, also known as Version of record

The Internet of Sounds: Convergent Trends, Insights, and Future Directions

Luca Turchet, *Senior Member, IEEE*, Mathieu Lagrange, Cristina Rottondi, *Senior Member, IEEE*, György Fazekas, *Member, IEEE*, Nils Peters, *Senior Member, IEEE*, Jan Østergaard, *Senior Member, IEEE*, Frederic Font, Tom Bäckström, *Senior Member, IEEE*, and Carlo Fischione, *Senior Member, IEEE*

***Abstract*—Current sound-based practices and systems developed in both academia and industry point to convergent research trends that bring together the field of sound and music Computing with that of the Internet of Things. This article proposes a vision for the emerging field of the Internet of Sounds (IoS), which stems from such disciplines. The IoS relates to the network of Sound Things, i.e., devices capable of sensing, acquiring, processing, actuating, and exchanging data serving the purpose of communicating sound-related information. In the IoS paradigm, which merges under a unique umbrella the emerging fields of the Internet of Musical Things and the Internet of Audio Things, heterogeneous devices dedicated to musical and nonmusical tasks can interact and cooperate with one another and with other things connected to the Internet to facilitate sound-based services and applications that are globally available to the users. We survey the state-of-the-art in this space, discuss the technological and nontechnological challenges ahead of us and propose a comprehensive research agenda for the field.**

***Index Terms*—Audio, embedded systems, Internet of Audio Things (IoAuT), Internet of Musical Things (IoMusT), Internet of Things (IoT), machine listening, music information retrieval (MIR), music, semantic audio.**

I. Introduction

RECENT developments in networking infrastructures, embedded systems,s and sound processing algorithms have opened opportunities for the integration of a variety of musical and nonmusical practices within the new contexts provided by computing and networking research. In the last two decades the Internet of Things (IoT) has made its inroads in the field of sound and music computing, leading to the emergence of novel paradigms, such as the Internet of Musical Things (IoMusT) [1] and the Internet of Audio Things (IoAuT) [2].

From a computer science perspective, IoMusT refers to the networks of computing devices embedded in physical objects (Musical Things) dedicated to the production and/or reception of musical content. Musical Things, such as smart musical instruments [3], [4] or wearables serving a musical purpose (such as smart watches, smart bracelets, or headsets for virtual/augmented reality) [5], [6], [7], are connected by an infrastructure that enables multidirectional communication, both locally and remotely. Similarly, 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. Networked Audio Things, such as nodes of wireless acoustic sensor networks (WASNs) [8] or networked systems for interactive sonification [9], can communicate both locally and remotely in order to, for example, gather meaningful information about the environment.

Both the IoMusT and IoAuT fields intersect the domains of embedded audio [10], networks [11], acoustic sensor networks [8], [12], ubiquitous and pervasive computing [13], [14], machine listening [15], and human–computer interaction [16]. Along the same lines, both fields lead to the emergence of novel ecosystems that form around sound-based technologies and that impact a large variety of stakeholders. In addition, IoMusT and IoAuT share common challenges related to privacy, security, interoperability, and standardization. Furthermore, several systems, discussed in this article, signal trends toward the convergence of approaches in system design, implementation and usage, in both the musical and nonmusical domains.

A parallel development is the rapid rise of voice-operated devices, such as smart speakers, televisions, mobile phones, and even microwave ovens. For example, in 2019, the sales of smart speakers rose to 146.9 Million.[1] Such devices offer hands-free operation of information services and physical actuators, such as home automation. Many smart speaker

Manuscript received 12 May 2022; revised 16 February 2023; accepted 4 March 2023. Date of publication 7 March 2023; date of current version 23 June 2023. *(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).

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

Cristina Rottondi is with the Department of Electronics and Telecommunications, Politecnico di Torino, 10129 Turin, Italy.

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

Nils Peters is with the International Audio Laboratories Erlangen, University of Erlangen-Nuremberg, 91054 Erlangen, Germany.

Jan Østergaard is with the Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark.

Frederic Font is with the Department of Information and Communication Technologies, Universitat Pompeu Fabra, 08002 Barcelona, Spain.

Tom Bäckström is with the Department of Signal Processing and Acoustics, Aalto University, 02150 Espoo, Finland.

Carlo Fischione is with the Department of Network and Systems Engineering, KTH Royal Institute of Technology, 114 28 Stockholm, Sweden.

Digital Object Identifier 10.1109/JIOT.2023.3253602

[1] https://www.forbes.com/sites/ilkerkoksal/2020/03/10/the-sales-of-smart-speakers-skyrocketed/

manufacturers already offer systems with a network of interacting devices, which could be called voice operated IoT devices. A recent project, the Open Voice Network, generalizes this concept and attempts to develop a standard for communicating between interoperable voice operated IoT devices.[2]

All the trends reported above are converging toward a novel research area that we coin as the Internet of Sounds (IoS). The emergence of the IoS field as a discipline in its own right is not only witnessed by an increasing corpus of literature on the several and diverse underlying topics, but also by the creation of a community of academics, practitioners, and industrial representatives who are gathering around a dedicated annual event, the "International Symposium on the IoS" (arrived in 2023 to the fourth edition)[3]. In November 2022 the IoS community, with the aim of sharing knowledge about and foster research on IoS topics, launched the "IoS Research Network" initiative,[4] which at the time of this writing gathers more than 90 partner institutions worldwide and more than 200 researchers have subscribed to the mailing list. In March 2023, the IEEE Communication Society accepted the proposal of the IoS community to form an "IEEE Emerging Technology Initiative on the IoS."[5] Moreover, the Journal of the Audio Engineering Society has promoted the field with a dedicated special issue published in 2021.[6] Furthermore, in recent years, funding bodies have supported a variety of IoS projects in both the musical[7] and nonmusical[8] domain, which indicates the timeliness and importance of this field. Following these trends, IoS courses at master degree and Ph.D. levels are taught in different universities.[9]

In this article, we present a survey of the IoS field aiming at organizing the knowledge accumulated in previous studies to build a foundation for future IoS works. To this end, we offer a review of studies related to the analysis, representation, and interconnection of sound-related information, in both musical and nonmusical domains. We also provide an overview of hardware and software-enabling technologies for the IoS, with a particular emphasis on system architecture paradigms and applications to realistic use cases. Based on the review of tools and applicative results, we identify and discuss open challenges of this field. Our aim is not only to bridge existing research areas and communities and foster cross-domain collaborations, but also to ensure that IoS-related challenges are tackled within a shared, pluralist, and system-level perspective.

We believe that the IoS has the potential to foster new opportunities for the IoT industry, paving the way to novel services and applications that are capable of exploiting the interconnection of the digital and physical realms, especially in the Smart Home and Smart City contexts. Nevertheless, for IoS technologies to emerge and be widely adopted by end users, a number of computational and human-related challenges need to be addressed.

While we are aware that in other contexts the term "audio" refers to a macro-category for musical, speech, and environmental sounds, it is worth noticing that in the specific context of this article, we highlight the difference between the terms "music," "audio," and "sound." With "music," we exclusively refer to musical data, with "audio" we refer solely to the domain of nonmusical auditory data, whereas with "sounds" we mean the union of both music and audio. Though speech, audio, sound, and music are all closely related, speech, and speech interfaces are such large areas of study on their own (e.g., [17], [18], [19], [20], and [21]) that we choose to mostly exclude them from this work and focus on music, audio, and sound.

The remainder of this article is organized as follows. Section II introduces the conceptual basis of the IoS field. Section III surveys works and technologies related to the IoS. Section IV discusses the main research challenges ahead of us on the IoS landscape. Finally, Section V outlines a research agenda for this area, while in Section VI we provide summarizing conclusions and final remarks.

II. Scientific Field of the Internet of Sounds

The IoS field addresses musical and nonmusical domains in networked contexts. We see the IoS as the union of the two paradigms of the IoMusT and the IoAuT. As described in [2], these two fields are intersecting. Both fields have been envisioned in [1] and [2] as subfields of the general IoT field. Therefore, the IoS is a specialization of the IoT, where one of the prime objectives is to enable processing and transmission of musical and nonmusical data and information.

In the proposed vision, the IoS enables the connection of digital and physical domains by means of appropriate information and communication technologies, fostering novel applications and services based on musical and nonmusical information. The IoS enables the integration and cooperation among heterogeneous devices with different sensing, computational, and communication capabilities and resources, in musical and nonmusical contexts as well as in co-located and remote settings.

The IoS has strong connections with and could be seen as a subfield of the Internet of Media Things (IoMT), which in turn is a subfield of the IoT. The IoMT is defined as a network of Things capable of sensing, acquiring, actuating, or processing media or metadata [22]. The IoS differentiates from the IoMT for its focus on sound-based applications, whereas the IoMT also deals with other multimedia aspects, such as video. Fig. 1 illustrates the positioning of the field IoS with respect to its composing subfields of the IoAuT and IoMusT, and to the IoT and IoMT fields.

The most critical difference between IoS and IoT is the nature of the acquired, processed, and transmitted information, which in the case of the IoS is sound and other sound-related content. Moreover, the IoS requires dedicated devices able to

[2]https://openvoicenetwork.org/

[3]https://internetofsounds.net/is2_2023/

[4]https://internetofsounds.net/

[5]https://www.comsoc.org/about/committees/emerging-technologies-initiatives/internet-of-sounds

[6]https://www.aes.org/journal/online/JAES_V69/10/JAES_V69_10_PG706.pdf

[7]E.g., https://cordis.europa.eu/project/id/749561

[8]E.g., https://cordis.europa.eu/project/id/956369

[9]E.g., https://www.audiolabs-erlangen.de/fau/professor/peters/teaching/2021s_AIoT

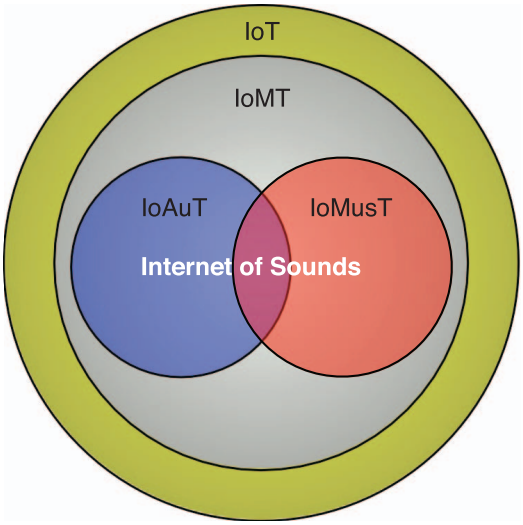

Fig. 1. Schematic representation of the relation between the IoS, the composing fields of IoAuT and IoMusT, and the parent fields of the IoMT and IoT.

sample, process, and/or synthesize sound (especially in real-time contexts) as well as specific network architectures (for instance to support a multidirectional stream of audio packets from/to nodes that is continuous and periodic). In contrast to IoT and IoMT, the IoS may pose stringent requirements and challenges related to the collection, analysis, and communication of sound-related information. For instance, to enable geographically distant musicians to play together in real time, the network infrastructure must ensure latencies at the order of milliseconds, to ensure no perceptible delays. Similarly, a distributed array of microphones in a WASN deployed in an open environment might need to be synchronized tightly with low-latency communications to detect audio events in real time. Current IoT hardware, protocols, and systems are insufficient to tackle those challenges. Moreover, the IoS demands a new set of analytic tools specific to the sound processing domain, which should be able to process large amounts of sound-related data and extract meaningful information given tight latency and energy constraints. This entails specific challenges in the areas of real-time signal processing and machine learning (ML). Furthermore, current data models (e.g., ontologies) devised for the representation of the IoT and IoMT domains are not adequate to describe the knowledge related to IoS ecosystems. This hampers the creation of ecosystems where heterogeneous IoS devices can communicate through a common, interoperable framework [23].

A. IoS Definitions

Following the definitions of the IoMusT and the IoAuT reported In [1] and [2], respectively, we provide a definition of a Sound Thing as "a networked computing device, equipped with sensors and/or actuators, with the capabilities to acquire, process, exchange, or generate sound or sound-related information". With "sound-related information" we refer to "data sensed and processed by a Sound Thing, and/or exchanged with a human or with another Sound Thing". We define the IoS as "the ensemble of Sound Things, network infrastructures, protocols, and representations of sound-related information that enable services and applications for the communication of sound-related information in physical and/or digital realms".

Furthermore, similarly to what the Web of Things[10] represents for the IoT [24], we use the term "Web of Sound Things" to refer to approaches taken to provide an Application Layer that supports the creation of IoS applications.

Just like the general IoT domain [25], [26], the IoS may be structured into ecosystems. An *IoS ecosystem* consists of IoS technologies (hardware and software platforms as well as standards) and communities of stakeholders utilizing them. From the technological perspective, the four core components of an IoS ecosystem are listed.

1) *Sound Things:* As defined above, Sound Things are networked devices utilized to control, generate, or track responses to sonic content. Sound Things are entities that can be used in a musical or nonmusical context to produce sound-related content or to observe phenomena associated to sound-based experiences. They can be connected to a local and/or remote network and act as sender and/or receiver. At the hardware level, they may be equipped with sensors (in particular microphones), actuators (in particular loudspeakers), and a variety of wireless and wired connectivity options. At the software level, they encompass programs that not only enable the collection, analysis, reception, and transmission of sound-related information, but also provide context-awareness and proactive capabilities. Key factors are interoperability and synchronization. The IoS vision predicts that in the future, several and new kind of musical and nonmusical devices will be connected to the Internet for the users to free themselves from several constraints, such as geographical locations and synchronous presence of every stakeholder. This could have a transformative effect on how humans conduct sound-based activities and interact with sound-based devices to better comprehend the environment or perform creative tasks in a more convenient way.
2) *Network:* The IoS network infrastructure supports multidirectional wireless or wired communication between Sound Things. The interconnection of Sound Things may happen over local and/or remote networks and is achieved by means of dedicated hardware and software technologies, as well as standards and protocols that regulate the communication. Some IoS applications focusing on real time, such as music live performance or synchronized WASNs, put particular constraints on communications. In such use cases, the connectivity infrastructure should ensure communications with low latency, high reliability, high perceptual quality, and tight synchronization between the nodes. Typically, these requirements are not present in the vast majority of IoT applications.

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

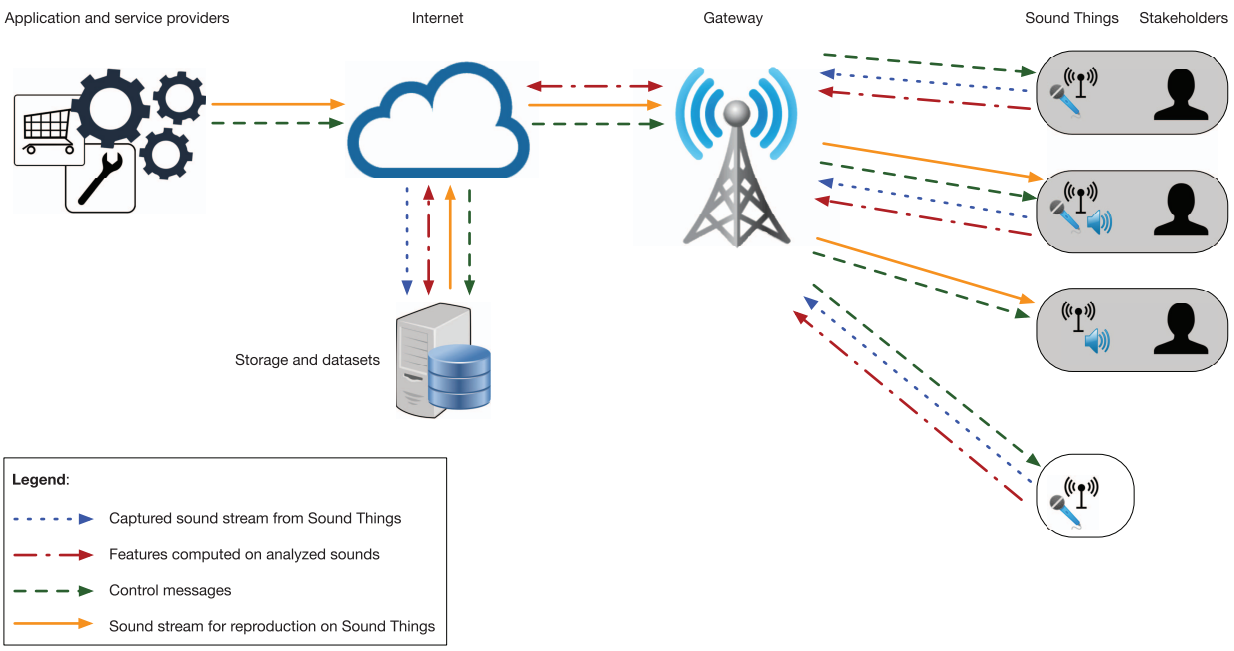

Fig. 2. Schematic representation of a general architecture supporting IoS ecosystems.

3) *Storage and Data Sets:* Most IoS ecosystems strongly rely on online storage of sound-related information, which is one of the main differentiating aspects with respect to IoT. This storage, centralized or distributed, can be of different types and refers to musical and nonmusical information. One type is represented by online sound repositories, which are entities in evolution: users can update the repository by changing its content, e.g., uploading new sounds or deleting existing records. Examples of this category are Freesound[11] or Jamendo.[12] A second type consists of sound corpora. These are essentially static databases that users can download and that typically contain audio files together with pre-extracted audio features and metadata annotations. This is the case of the MTG-Jamendo data set [27] for automatic music tagging, or FSD50K [28], a data set for tagging of sound events. In music data sets, however, it is very common that copyright limitations do not allow the audio to be shared as part of a data set. In that case, links to sound files are normally provided together with pre-extracted audio features. An example of this category is AcousticBrainz.[13] The third type is represented by test beds, i.e., platforms to facilitate diverse research through providing collections of data sets with meta-level annotations. The Open Multitrack Testbed[14] [29], for example, can be used to search multitrack recordings with a specific set of instrumentation or stems, and provides a browsing and searching interface with metadata filters related to specific devices used in music production.

4) *Applications and Services:* Different kinds of applications and services can be created on top of the connectivity infrastructure, which target a variety of stakeholders. Such applications and services may have an interactive or a noninteractive nature. To establish interactive applications, real-time computations have a particular importance. Analogously to the IoT field, the IoS can leverage Web application programming interfaces (APIs) and Web of Things architectures devised to serve sonic purposes. Services can be exposed by Sound Things via Web APIs. Applications are part of a higher layer in the Web of Sound Things architecture letting users interact with content or Sound Things directly.

Fig. 2 depicts the main technological components of an architecture supporting a generic IoS ecosystem. The data flow exchanged via wireless and/or wired links between such components can be grouped into the following.

1) *Captured Sound Streams:* Some Sound Things are equipped with microphones to collect data from the user (e.g., a networked music performance (NMP) system [30]) or from the environment (e.g., nodes of WASNs [31]).
2) *Features Computed on Analyzed Sounds:* Some Sound Things have the capability of performing real-time or offline analysis of the sounds captured from microphones (e.g., a smart musical instrument [32]), while online storage and data sets can provide the features upon queries.
3) *Control Messages:* Beyond sound or sound-related information, the components of an IoS ecosystem may exchange messages to control the behaviors of each other (e.g., data sent to control sonic shoes used in rehabilitation applications [9] or used to query online sound repositories [3]).

[11] https://freesound.org/
[12] https://www.jamendo.com/
[13] https://acousticbrainz.org/
[14] http://multitrack.eecs.qmul.ac.uk/

TABLE I
POTENTIAL BENEFITS FOR DIFFERENT STAKEHOLDERS BROUGHT ABOUT THE IMPLEMENTATION OF THE IOS PARADIGM

Stakeholder	Benefit
Amateur and professional musicians (performers, composers, students and teachers, sound producers, live sound engineers)	• Novel kinds of individual and collaborative musical activities via networked musical devices: new forms of musical expression, pedagogy, interaction with musical content, collaboration, remote participation, and remote control. • Novel forms of monetization via blockchain technology: new forms of revenue streams for artists, music distribution without intermediaries.
Audiences	• Enjoyment of new artistic forms based on networked interactions between performers, between performers and audience members, as well as between audience members. • New forms of revenue streams specific for audiences via blockchain technology.
Doctors, patients and the healthcare sector	• Novel forms of sound-based therapies in networked contexts (e.g., tele-rehabilitation).
Citizens at large	• Possibility for policy makers of defining measures to reduce acoustic pollution in cities via the monitoring of the urban soundscape through WASNs. • Real-time monitoring of dangerous situations in cities (e.g., gunshots) and remote surveillance. • Real-time synchronized acoustic information provided to users at large indoor or outdoor spaces (such as hospitals or stadiums) via a distributed network of loudspeakers. • Low-latency and highly reliable voice communications (via phone or videoconferencing systems).
Wildlife and environment	• Possibility for policy makers of defining measures to safeguard wildlife via ecoacoustics monitoring methods leveraging WASNs. • Possibility for policy makers of defining measures to improve the quality of the urban environment by considering monitoring methods leveraging WASNs. • Environmental sustainability: reduced pollution by zeroing the travels of musicians in need to meet for collaborative musical activities, as well as the travels of patients in need to go to hospitals for the sound-based therapies. • Real-time monitoring and predictions of volcanoes evolution based on vibration and sound analysis to alert citizens.
Manufacturing industry and machine operators	• Sound-based predictive maintenance and automatic sound-based anomaly detection in manufacturing machines or other systems handled by human operators (e.g., large motors of ships).
Audio service providers	• Availability of novel technologies enabling the creation of new types of sound-based services, such as context-aware recommendation of musical content.

4) *Sound Streams for Reproduction:* Some Sound Things encompass loudspeakers which enable to reproduce sound that is synthesized (e.g., by smart speakers [33] or smart musical instruments [34]), played back (e.g., from online sound repositories [35]), or received from other Sound Things as a real-time flow (e.g., in NMPs [36]).

B. IoS Stakeholders

The IoS field impacts a large variety of stakeholders, which has important implications at societal and economical levels. As far as the IoMusT is concerned, the stakeholders include amateur and professional musicians, performers, composers, conductors, studio producers, live sound engineers, audience members, students, teachers, schools of music, record labels, publishers, musical instruments manufacturers, concert venues, and musical services providers. Regarding the IoAuT, stakeholders may include all actors involved in ecoacoustics [37], smart city [31], or smart homes contexts who are using audio-based services. In addition, relevant stakeholders are Audio Things manufactures and audio-based service providers. Furthermore, common to both fields there are stakeholders, such as telecommunication companies, institutions, and regulatory bodies.

All such stakeholders will form ecosystems around IoS technologies, which will allow them to interact and avail themselves of innovative services. Sound Things will be able to support the activities of such stakeholders, thanks to their embedded intelligent capabilities (such as context-awareness and proactivity [38], [39]). Table I summarizes the potential benefits for different stakeholders brought about the implementation of the IoS paradigm.

C. Types of Interactions

Different kinds of interaction can occur within IoS ecosystems, which depend on the following aspects.

Entities: Sounds and sound-related information can be exchanged between: 1) machines (e.g., nodes of WASNs); 2) humans (e.g., collaborative music making using an NMP system); and 3) machines and humans (e.g., queries from a smart musical instrument to an online sound repository).

Temporal Aspects: Interactions between the entities above can be: 1) real time (e.g., via NMP systems or via smart

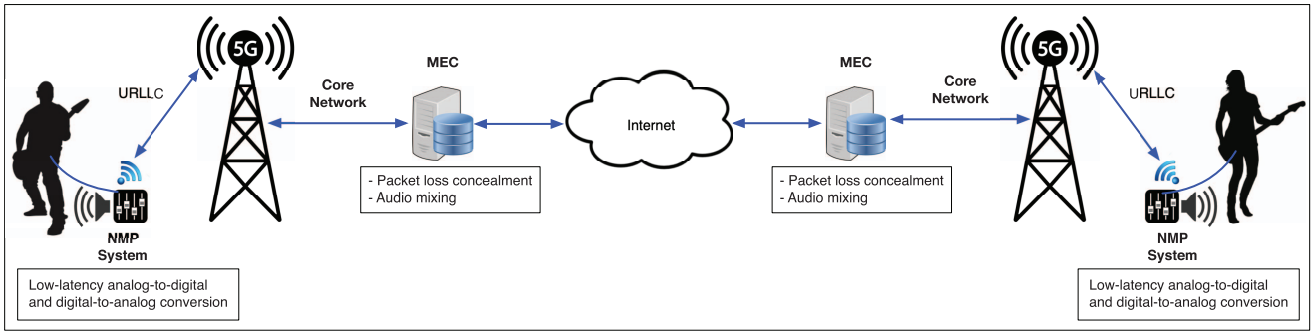

Fig. 3. Diagram of a 5G architecture supporting low-latency, highly-reliable musical interactions between two geographically displaced musicians.

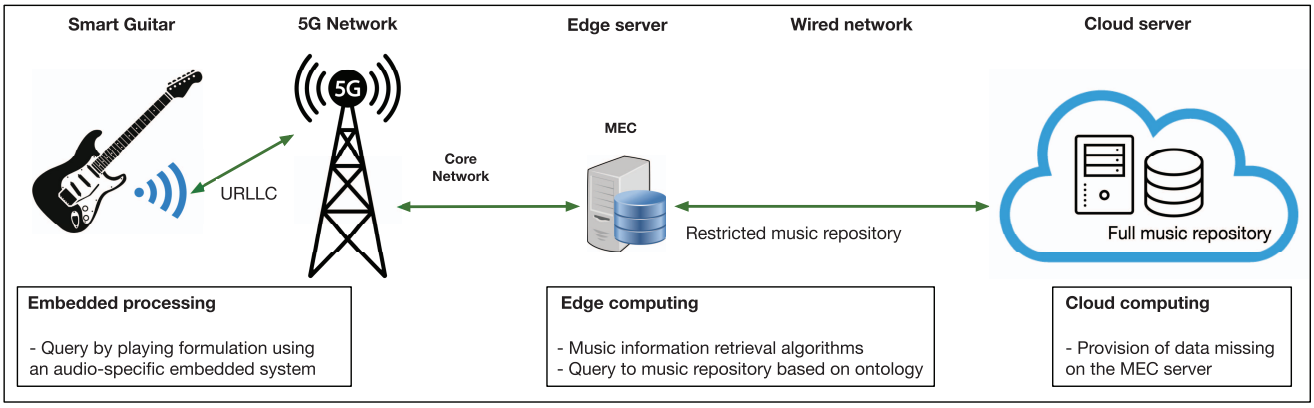

Fig. 4. Diagram of a 5G architecture supporting the use case of a smart guitar performing queries by playing to a server on the edge-cloud continuum.

instruments that repurpose the results of the analysis of sounds at the moment in which they are generated) or nonreal time (e.g., offline analysis of sounds in online repositories) and 2) synchronized (e.g., nodes of a WASNs sharing the same clock or musical devices sharing the same tempo) or asynchronized (e.g., musical performances not needed of tight synchronizations).

Spatial Aspects: Interactions may be: 1) co-located, when entities share the same physical location or 2) remote, when entities are geographically displaced.

Content: Sound and sound-related information in the IoS may be: 1) musical or 2) nonmusical, which lead to radically different interactions between the entities.

Directionality: Interactions between entities may be: 1) one-to-one; 2) one-to-many; 3) many-to-one; or 4) many-to-many.

III. Related Work

In this section, we review the state-of-the-art in relation to both the musical and nonmusical domains, with a focus on the most recent studies. The review is not meant to be exhaustive, rather we aim to describe the results of various application domains that lead to the emergence of the IoS field. We first review studies specific to the IoMusT, second those specific to the IoAuT, and then those that are relevant to both fields. We survey both hardware and software systems, along with application and services. In Figs. 3–7, we provide schematic diagrams of IoS architectures supporting various applications and services, detailing the main software and hardware components and their integration.

A. Relevant Works in the IoMusT

The IoMusT research field originates from the integration of many lines of existing research, including NMP systems [36], [40], ubiquitous music [14], new interfaces for musical expression [41], music information retrieval (MIR) [42], human–computer interaction [43], Musical XR [44], participatory art [45], IoT [46], and aspects of semantic audio [47] combining Web and audio technologies [48], [49]. In the following we survey the essential IoMusT components and underlying technologies.

1) Networked Music Performances: In essence, NMPs systems are real-time audio/video streaming applications aimed at supporting remote musical interactions among performers placed in different physical locations. Nowadays, several experimental and commercial NMP solutions are available (see [30], [50], [51], [52], [53], [54], [55]). The ongoing Sars-CoV-2 pandemic has significantly fostered their development, due to the sudden need to support the daily activities of musicians and music schools while adhering to social distancing rules enforced during lockdowns. Literature from the past two decades also reports a wide range of demonstrations of networked musical performances (some recent examples include [56], [57], [58]).

To achieve performative conditions as similar as possible to those experienced by musicians in traditional in-presence settings, the mouth-to-ear latency perceived by the performers must be extremely low: according to several studies appeared in the last decade [59], [60], [61], [62], [63], [64] it should not exceed 20–30 ms. This is the amount of time

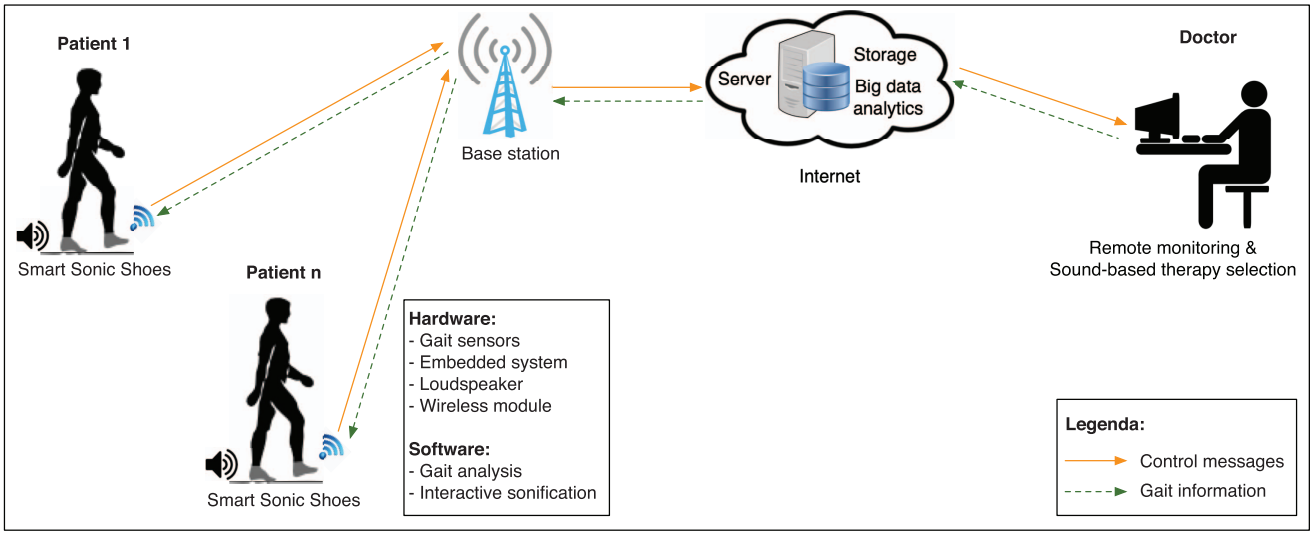

Fig. 5. Diagram of an architecture supporting remotely controlled sound-based therapies for gait rehabilitation via smart sonic shoes providing interactive sonification of the gait.

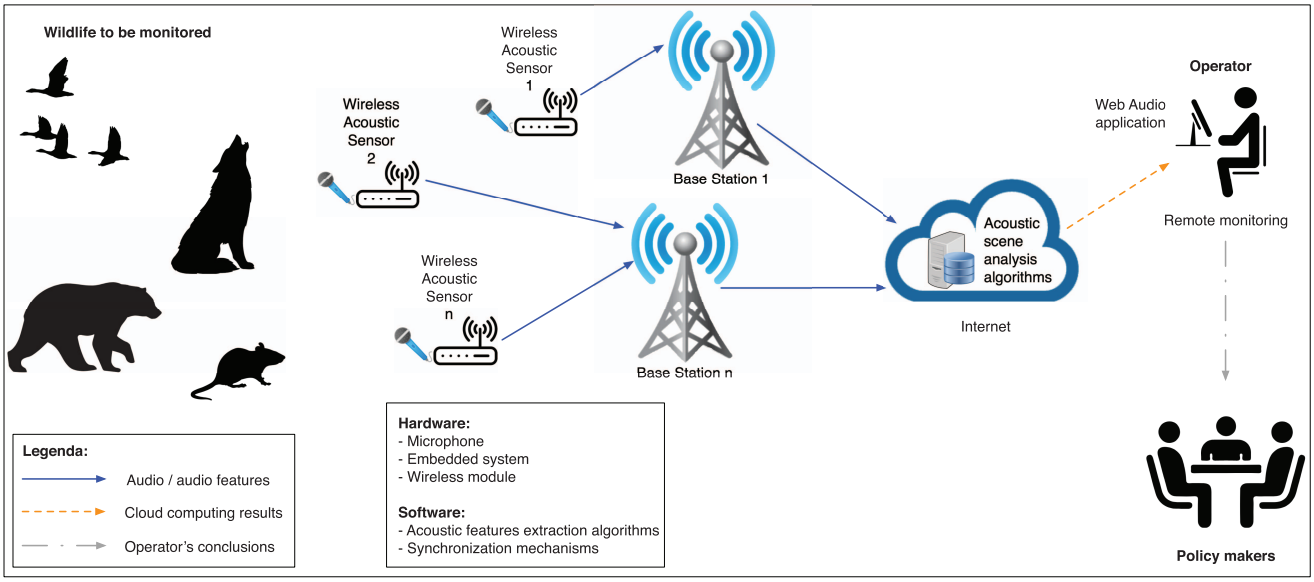

Fig. 6. Diagram of a WASN for wildlife monitoring.

taken by a sound wave to propagate in air 8–10 m (which is usually considered the maximum physical displacement tolerated by a musical ensemble to synchronously perform together in absence of additional reference tempo cues). Above 30 ms, latency becomes clearly perceivable and typically has a negative impact on the performance quality, as it makes the maintenance of tempo stability and of the desired musical interplay progressively harder, due to the fact that perceiving the counterpart as "late" generates a tendency toward tempo deceleration. For this reason, research efforts have recently been devoted to the integration of artificial metronomes capable of providing adaptive audio cues to the remote musicians [65], [66].

The overall mouth-to-ear latency includes multiple contributions introduced by different stages of the audio acquisition, processing, transmission, and reproduction processes. Due to the packet-switching approach adopted in Internet-based networks, audio data are subdivided in chunks and packetized, then each packet is routed individually from source to destination through the telecommunication network infrastructure. It follows that consecutive packets may experience different delays, as some latency components (e.g., the queueing time in intermediate routers) exhibits variations along time. The variation of the interarrival times between consecutive packets is named "jitter." Jitter excursions and packet losses may cause audio portions to arrive too late (or not arrive at all) to be reproduced at the receiver side, thus causing audio artifacts (note that conventional packet retransmission mechanisms cannot be adopted in NMP applications, as they would further increase the mouth-to-ear latency). By introducing a delay, the variance of the jitter can be reduced through a combination of jitter buffers and forward error correction codes [67].

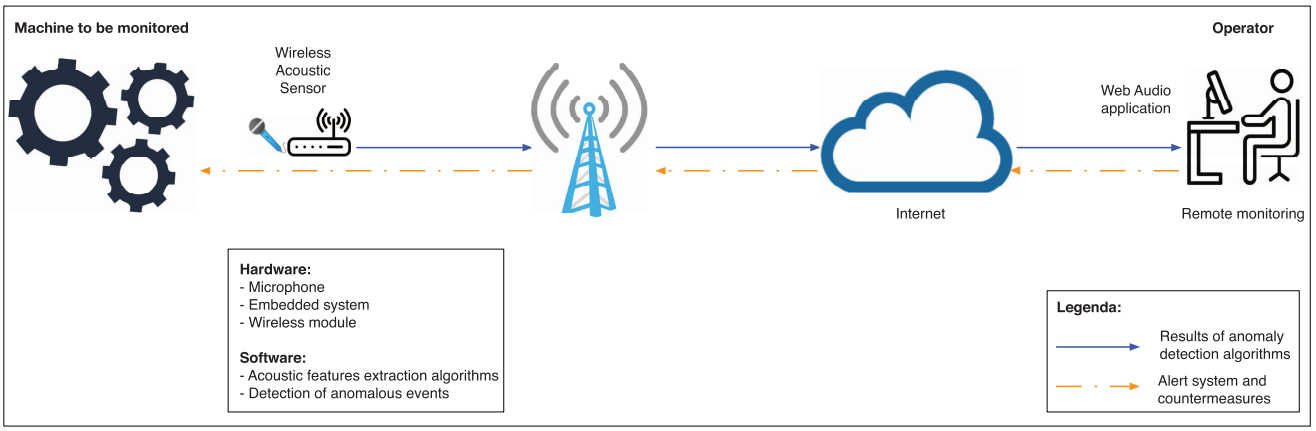

Fig. 7. Diagram of a networked, sound-based anomaly detection system for the monitoring of manufacturing machines.

Another source of audio impairments is the drifting effect caused by the imperfect synchronization of local clock oscillators, which leads to buffer over/underruns due to the deviation between the number of audio samples generated by the sender and the number of those reproduced by the receiver during a given time period [68], [69]. To mitigate the impact of such artifacts without introducing a delay, robust audio codecs designed for ultralow-latency streaming applications over imperfect networks can be used. These include audio coders where packet loss concealment is integrated (such as OPUS [70], [71]) or multiple-description audio coders, where audio frames are encoded into several redundant packets which individually yield a desired quality, and where any combination of packets improves the quality [72], [73], [74], [75], [76], [77].

2) Musical Things: Various types of Musical Things have been proposed by the IoMusT research community (e.g., [3], [6], [34], [78], [79], and [324]), along with frameworks to connect them (e.g., [80], [81], and [82]).

The so-called Smart Musical Instruments are one of the most prominent instances of Musical Things [4]. This is an emerging class of musical instruments characterized by sensors, actuators, wireless connectivity, and embedded intelligence. Such instruments are the result of the integration of various technologies, including IoT, sensor- and actuator-based augmented instruments, embedded acoustic and electronic instruments, NMP systems, as well as methods for sensor fusion, audio pattern recognition, and semantic audio. Smart musical instruments are envisioned to have advanced intelligence capabilities, such as context-awareness, adaptation, and proactivity. Nevertheless, up to now, only a few musical instruments exist in both industry and academia which encompass such features. Examples from industrial research are Smart Acoustic Guitar by HyVibe and the Sensus Smart Guitar developed by Elk [79]. Examples in academic research are the Smart Cajón reported in [32], the SOURCE sampler detailed in [3], or the Smart Mandolin described in [83].

A number of innovative applications associated to such instruments are also emerging. Turchet and Barthet [84] proposed a smart guitar system that uses the instrument as a hub for collaborative music making over a local wireless network. Thanks to this system, performers using musical apps on mobile devices produce sounds by wirelessly controlling the instrument's sound engine. At the same time, the player can not only actually play and control other parts of the instrument's sound engine, but can also send messages to the connected mobile devices changing the configuration of the app. In a different vein, the system described in [85] reports a smart guitar application enabling the use of distributed intelligence, via cloud computing, and edge computing paradigms, for recreational music making and learning contexts. Thanks to direct Internet connectivity and embedded processing, the instrument sends requests of desired musical pieces to online music repositories and reproduces the downloaded music. In particular, the search is performed using musical features, such as tempo and chords, which are extracted by the instrument's capabilities, rather than utilizing conventional text-based search, such as the title of the song or the artist's name.

Wearable devices utilized for musical purposes are another category of Musical Things [6], [35]. For instance, headsets for virtual or augmented reality can be considered as Musical Things if used in networked musical applications and in conjunction with other Musical Things (see [79], [86]). However, this line of research has thus far received remarkably little attention, despite its great potential for performance, composition, and education [44], [87]. On the one hand, Musical Haptic Wearables, are an emerging class of wearable devices embedding haptic stimulation, tracking of gestures and/or physiological parameters, and wireless connectivity features. On the other hand, such devices were devised to enhance communication between performers, as well as between performers and audience members by leveraging the sense of touch in both co-located and remote settings [88], [89]. They were devised to enrich musical experiences of audiences of music performances by integrating haptic stimulations, as well as provide novel capabilities for creative participation thanks to embedded sensor interfaces [7].

3) IoMusT Connectivity Systems: Recent endeavors in IoMusT research have explored the creation of dedicated systems to interconnect Musical Things, paving the way to the creation of IoMusT ecosystems. Some scholars have

proposed frameworks that allow one to exchange information within artistic installations [81], [90] or performance settings leveraging the real-time dimension [91], [92], [93].

Other authors have proposed preliminary architectures based on Semantic Web technologies to foster interoperability across heterogeneous Musical Things. The semantically senriched IoMusT architecture reported in [94] relies on a semantic audio server, embedded audio systems, and edge computing techniques. In particular, the SPARQL[15] (SPARQL Protocol and RDF Query Language) Event Processing Architecture described in [95] was used as an interoperability enabler allowing multiple prototypes of Musical Things to cooperate. However, Semantic Web technologies are typically not suitable for dealing with real-time aspects of IoMusT applications, as the Semantic Web stack is oriented toward scenarios where information evolves at a low rate. To cope with this issue, Viola et al. [96] improved the architecture reported in [94] by using CoAp, a lightweight IoT protocol for machine-to-machine communication.

Such architecture has been further improved and extended, leading to the musical semantic event processing architecture (MUSEPA), a semantically based architecture designed to meet the IoMusT requirements of low-latency communication, discoverability, interoperability, and automatic inference [97]. MUSEPA uses at its core the IoMusT Ontology, an ontology dedicated to the representation of knowledge related to the IoMusT domain [98].

4) Music Information Retrieval: The wider fields of Music Informatics and Semantic Audio are both concerned in part with the extraction of information from audio signals, signifying the relevance of these fields to both IoAuT and IoMusT and the broader IoS and IoMT domains. Semantic Audio goes beyond typical concerns of Music Informatics in its use of explicitly structured data and applications outside the music domain as discussed in Section IV-C. As a field related to both, MIR has seen an increase in its multidisciplinarity, and it is now commonly thought to cover research and techniques for music understanding and modeling that use information processing methodologies [99]. This includes music generation as well as cognitive or musicological analyses of different representations of music.

Core research in MIR, however, still focuses on tasks, such as key and chord recognition [100], [101], tempo and beat tracking [102], the detection of musical note onsets [103], [104], automatic music transcription (AMT) [105], classification [106], and description (also known as captioning) [107], [108] as well as music emotion recognition (MER) [109], [110], [111]. A large body of research considers musical audio in these tasks to support search, retrieval, and interaction use cases. This makes MIR especially relevant for the increasing number of interconnected IoS devices, such as smart speakers, networked media players, set-top boxes, general purpose intelligent assistants [112], as well as more specific devices such as smart instruments [4].

Conventional key and chord estimation techniques typically share a similar signal-processing pipeline. A time-frequency representation, such as a spectrogram or constant-q-transform is obtained first from the audio followed by aggregation steps, e.g., a pitch class profile or chroma feature calculation for every time frame. The objective is to fold energies from different octaves into a feature that is robust to timbre variations. The features are finally pooled over time and compared with templates for each chord or key. The best matching template determines the label. ML techniques, such as a hidden Markov model [113] may be used to smooth transitions or impose a musicological model to improve accuracy over raw labels extracted from the signal directly. More recently, end-to-end feature learning approaches have been proposed for these tasks [114], [115], where typically a convolutional neural network frontend is applied to process a time-frequency signal representation into increasingly higher level features [116]. This is followed by dense layers and softmax classification to determine the label. A comprehensive survey of the development of chord recognition is provided in [117].

Typical approaches to determine the onset of musical notes involve the calculation of an onset detection function (ODF), e.g., using time-domain energy fluctuation or zero crossing rate and frequency-domain techniques, such as spectral difference or phase deviation [103]. The ODF is then smoothed and its peaks are picked to find note onsets. The time-frequency representation of the signal may be preprocessed to reduce the effects of instrument playing techniques such as vibrato [118], while different acoustical features may be fused in order to emphasize less salient notes in the signal [119]. Neural networks have also been employed in this task [120], where a convolutional frontend followed by dense layers is used to compute an ODF. While deep learning represents the state-of-the-art in terms of accuracy, in IoS applications with hard real time and energy constraints, signal processing methods are still relevant [104].

AMT is concerned with analyzing an acoustic signal to extract parameters of the sound corresponding to traditional music notation or other symbolic representation that allows reproducing the piece using musical instruments or synthesizers [105]. This is relevant to IoS as this allows advanced search, for instance, by matching melodic patterns, or advanced interaction through altered resynthesis. AMT is often divided into subproblems, such as note onset detection and multipitch estimation, however, nonnegative matrix factorization (NMF) and Neural Networks have been predominant in the field in recent years [121]. These aim to jointly estimate the spectro-temporal characteristics of musical notes. NMF considers a nonnegative time-frequency signal representation as the product of a dictionary and an activation matrix which are iteratively updated to minimize the divergence between, e.g., a spectrogram and the matrix product. Neural networks for AMT use one or more time-frequency representations as input and aim to estimate note onsets, pitch, and duration. A notable example is Google Brain's Onsets and Frames Network [122]. State-of-the-art approaches aim to capture short and long-term characteristics of the signal, e.g., the temporal evolution of note spectra as well as dependencies between musical notes. For example, the model proposed in [123] represents audio features and language-like dependencies jointly.

[15]http://www.w3.org/TR/sparql11-query/

The association of semantic labels, such as genre or mood with audio signals is also very prominent in MIR. Instrument, genre, or emotion recognition may be considered separate tasks, or part of a multilabel classification problem known as automatic music tagging. In the IoS domain, these labels facilitate retrieval. MER is conventionally treated as a supervised ML problem. Acoustic features corresponding to cognitive or psychological factors may be selected, then models are trained to predict either categorical mood labels (e.g., happy versus sad) or a continuous representation of moods (e.g., arousal and valence). Features may also be selected adaptively [124] or using a data-driven approach [125], [126]. A comprehensive survey of MER can be found in [109]. Similar pipelines have been dominant in instrument and genre recognition too [106], but state-of-the-art results are now obtained using end-to-end deep learning models. These tasks are increasingly considered part of a broader music tagging problem [127], [128]. Convolutional and recurrent networks are combined in [129] for predicting labels corresponding to genre, mood, instrument, and musical era. A musically motivated front-end focusing on timbral and temporal features simultaneously is proposed in [130], while features extracted from a convolutional network trained on tags are shown to perform well across a range of MIR tasks in [127]. Generating full-sentence descriptions of a music piece may be considered an extension of the tagging problem. This involves the use of an acoustic model and a large language model [108].

In the broader context of interconnected IoS devices, MIR can facilitate new mechanisms for search and retrieval as well as new interactions with large music collections on the Web. Processing audio queries can be handled for instance by extracting features locally in the IoS device using MIR techniques, and using these features in the retrieval process. A relevant application is demonstrated in [85] in the context of a smart guitar, comparing keyword search with both cloud and edge computing approach to retrieve content using tempo, chords, key, and tuning features. Earlier applications for hand-held IoS devices (e.g., tablets) demonstrate how MIR facilitates music learning in an IoS context [131], [132], while both retrieval and repurposing of sounds are exemplified in [133].

B. Relevant Works in the IoAuT

Similarly to the IoMusT, the IoAuT research field is positioned at the confluence of different disciplines. Beside the IoT and networking, these include WASNs, sound and music computing, sonic interaction design, semantic audio, artificial intelligence, and human–computer interaction.

1) Acoustic Sensor Networks: As the impact of human activities on the health of humans and the natural environment is getting more and more important, there is an urge to reliably gather information about our environment that allows us to take action. To tackle this issue, one approach is to foster on the advent of low-cost microphones based for example on the MEMS technology [134] to monitor our environment. In order to produce reliable, interpretative, and privacy-preserving information using this kind of technology, it has to be done at scale, on the edge, and in a efficient manner. The audio modality has many advantages. Its is contactless, noninvasive, omnidirectional, provides features of relatively low bandwidth and can operate without light. Those advantages lead to the widespread of application scenarios in the recent years, mostly for surveillance and monitoring.

Gathering information about our environment using the acoustic modality requires a complex processing chain. The pipeline usually consists in a microphone, a processing unit for computing low-rate features, and another processing unit that, from those features, predicts high-level attributes such as the presence of sources of interest. Finally, the audio, the features, and the high-level attributes may be consumed via gateways or stored using storage facilities. Some constraints have to be met when placing key components over the graph that constitutes the network of sensors and servers. For example, the microphone must obviously be located in the sensor. For the others, the design choices leading to the correct network architecture is task-specific [135] depending on the correct balance between privacy, energy harvesting, security, and reliability for the project at hand. Traditional acoustic array processing allows for spatio-temporal processing of the surrounding sound field, and is frequently used for detection and localization of sources, analyses of signals, enhancement, noise reduction, etc. In acoustical sensor networks (ASNs), the wireless microphone sensors can be placed at arbitrary locations and is not restricted to a fixed grid. This provides a high degree of flexibility but also introduces challenges, such as synchronization issues between the sensors, artefacts due to audio coding and transmission errors, potentially unknown and time-varying network topology and quality, and energy restrictions in the case that the sensors are battery driven.

Predicting high-level attributes from those sensors is done using machine-listening software components that detect sound events and classify them into application-specific ontologies. The now well-established DCASE challenge[16] provides a unique venue fostering academic and industrial efforts toward effective detection or classification models [136], [137] for standardized tasks that can serve as reference when considering real application scenarios. The challenge allows the community to compare different inference techniques in a standardized setting geared toward replicability and fairness of comparison. A companion workshop is held that allows the challenge participants to describe and discuss their design choices. Besides striving for more and more powerful inference models, there is a number of important issues to address in order to ensure that the society takes full profit of the technology at hand. This section describes some of these challenges and discusses recent approaches taken to improve them.

a) Urban audio: Urban environments are designed by humans for humans but their acoustic quality are most of the time neglected. IoS technologies are envisioned to play a role in improving the quality of the urban environment by many means [138]. Passive monitoring to allow better enaction is under research in many major cities in the world [31], [139].

[16]http://dcase.community

Active solutions where sound displays are played at specific places are also investigated [140]. At the scale of housing, similar potentials are also envisioned.

The urban environment, as designed by humans, should provide security and safety to its citizen. In order to do so, urban planners needs to balance many factors. Noise mitigation and urban soundscape quality improvement are now recognized as important, but citizens and policy-makers currently lack quantitative and qualitative information about it. European regulations request that large cities publicly display a noise emission map to allow citizens to have quantitative information about their exposure to noise. As of today, these maps are built using sound propagation techniques similar to ray-tracing in image synthesis and are only based on some estimates of the number and the speed of emitting sources (e.g., cars, trains, and planes) [141]. While those noise maps have their interests, they typically lack time resolution, display only sound pressure levels, and are by essence, only predictive and not tied to direct measurements. In order to better describe the acoustic environment and its potential impact on the quality of life of the citizens, it is necessary to characterize it in terms of presence of sources.

For those reasons, in several countries around the world, innovative projects, such as SONYC in New York City [142], DYNAMAP in Rome [143], SONORUS in Antwerp [144], StadtLärm in Jena [145], or CENSE in Lorient [139] have deployed acoustic sensor networks with various designs. Measuring the sound environment in urban areas is far from trivial because of the diversity of sound sources but the use of sensors opens a wide range of improvements. First, continuous monitoring allows citizens to interact with regulations administration more efficiently [31]. Second, noise maps produced by sound propagation techniques and ASNs have different biases. Considering both source of information helps reduce estimation biases in order to produce more reliable estimates [146]. Third, information about sources of interest, vehicle, humans, and animals are very useful for prediction high-level attributes of the sound environment such as its pleasantness [147].

Despite the potential of acoustic sensor networks for urban sound monitoring, these technologies currently face multiple challenges. In particular, meeting citizen's privacy constraints [135], designing reliable networks, efficiently adapting source detectors to new monitoring areas without requiring a significant amount of human labor [148], [149], all those challenges are now under active research.

b) Ecoacoustics: Along the study of vocalizing animals at an individual level, a field of research called bioacoustics, a new field have recently emerged that is termed ecoacoustics [37]. The purpose of ecoacoustics is to gather high-level information about given species living in an ecological niche through the use of passive network of microphones. The quantities of interest here are typically the number of animal calls in order to monitor for example the impact of human activities or climate change on the behavior of a specific species.

Most of those applications require the acquisition of some metrics at an hour rate, often for very long periods of time, ideally up to several years. For example, the study of the impact of light pollution on bird vocalizations [150] requires a high level of precision in time. Those requirements, along the fact that most ecosystem under study are rather remote with low access to power and wireless networking, brings strong constraints on the design of the network of sensors. Among the most drastic is the autonomy of the sensors, autonomy meaning here low dependence to energy harvesting and maintenance [151].

In more densely populated areas with cell phone coverage, citizen science approaches, such as Merlin BirdId[17] Warblr[18] or birdnet[19] allows, based on advanced machine listening techniques [152], [153], the citizen to easily geotag bird calls. Indeed, the cell phones are equipped with the necessary hardware needed to record, identify, and send bird species identification reports. While this kind of approach has numerous benefits socially speaking [154], it has strong limitations for scientific investigations due to the sampling bias induced by considering only voluntary contributions. As such, remote places such as off-shore underwater areas require specific, reliable dedicated hardware [155] to perform long-term studies [156].

2) Audio Things: Various IoT devices designed to provide or collect sonic information to/from users have been proposed while others are currently under study [151]. An illustrative example is represented by sonic shoes devised for clinical or sport training applications [9], [157]. Another example of Audio Things concerns the analysis of environmental sounds. Various embedded systems have been utilized for this purpose, using both real-time and nonreal-time architectures. Relevant examples concern environmental monitoring [158], ecoacoustics (e.g., birds monitoring) [159], [160], [161] and urban sounds [31]. Smartphones are also increasingly used for similar purposes [162] along with drones [163], [164].

In addition, Audio Things with real-time sound analysis capabilities have been employed for remote anomaly detection in manufacturing or machine operation contexts [165], [166]. Moreover, voice-based interfaces such as smart speakers have been utilized to access cloud-based repositories of nonmusical sounds [33].

Hearing aids are another prominent category of Audio Things. Next-generation AI-powered wireless hearings aids will be connected to computing IoS devices and thereby enabling functionalities, such as audio streaming from your smart TV, hands-free communications via the smart phone, and real-time translation of the received speech signal into any desired language [167]. In addition, hearing aids will be equipped with nonacoustic sensors, such as cameras [168], EEG sensors [169], bone conduction and skin vibration sensors [170], etc., in order to obtain enhanced speech quality and intelligibility in noisy environments [171].

C. Relevant Works Common to Both IoMusT and IoAuT

The IoAuT and IoMusT may share some technologies enabling their applications. The following surveys some among the most prominent of them.

[17] https://merlin.allaboutbirds.org
[18] https://www.warblr.co.uk
[19] https://birdnet.cornell.edu

1) Embedded Systems for Sound-Based Applications: Researchers have proposed a number of operating systems for the IoT (such as RIOT [172]). However, these are not specifically conceived bottom-up to handle sound and are insufficient for most IoS applications, which usually have requirements in terms of low-latency of sensor and sound signal processing.

Finally, there has been an increasing attention toward the development of embedded systems dedicated to digital sound processing, where a variety of audio software runs on single-board computers, such as the Raspberry Pi or the Beaglebone [173]. This endeavor is usually referred to as "embedded audio." Notable examples, conceived for the makers community, are Axoloti[20] and Prynth [174]. According to the results reported in [175], the state-of-the-art in this space is represented by Bela [176]. This platform is based on the BeagleBone Black single-board computer, which is extended with a custom expansion board featuring inputs and outputs for audio, sensors, and actuators.

In industrial contexts, the state-of-the-art in embedded audio today is arguably represented by the Elk Audio OS, a Linux-based, open-source operating system developed by the company Elk [50]. Elk Audio OS is optimized for ultralow-latency and high-performance sound and sensor processing on embedded hardware, as well as for handling wireless connectivity to local and remote networks. Differently from other systems, it supports a variety of single-board computers. To achieve latencies below 1-ms Elk Audio OS (as well as Bela) uses the Xenomai real-time kernel extension, which according to different studies [10], [177] is the best-performing of the hard real-time Linux environments.

2) Semantic Audio: Semantic Audio is an interdisciplinary field providing techniques to extract meaningful and structured information from audio. The field is situated in the intersection of semantic technologies concerned with helping machines understand data through structured knowledge representations that facilitate automated information processing [178], and information extraction from audio through the use of signal processing and ML techniques. The objective of semantic audio is often to facilitate interaction with audio in human terms. This is achieved by providing high-level meaningful control in complex scenarios such as audio production [47].

Semantic audio methods find application in several IoS-related scenarios, from intelligent audio and music processing and production [39], [179], [180], [181], [182], [183], [184], to online music distribution [48], [49], to auditory scenes classification tasks [136], to ecoacoustics [185] as well as speech technology, including speaker, gender, or language identification [186], [187].

An important endeavor in the Semantic Audio field has been that of defining a variety of ontologies to represent knowledge related to the musical and nonmusical domains. Regarding musical contexts, examples include the music ontology[21] (MO) [188], [189], a high-level ontology for representing the music domain, in particular modeling the music value-chain from production to consumption [49]. Since MO is arguably the most comprehensive ontology frameworks for the music domain [190], several extensions have been proposed for specific subdomains. These include the Studio Ontology [191], a framework comprising of a set of modular ontologies that represent artefacts and technical workflows in music production. The Audio Effect Ontology[22] represents audio effects [192] and their applications [193] in music production workflows. The Audio Features Ontology[23] is concerned with providing a structures schema for information extracted from audio signals, i.e., descriptors representing specific characteristics of sound signals [194] linked to information extraction tools. Broader extensions of these ontologies encompass music theoretical domains [195], ethno-musicology [196], musical archives [197], [198], as well as musical instruments [199], [200] and mobile devices for music making [201]. This and the following extensions are particularly relevant in the IoS context. The IoMusT Ontology[24] represents knowledge related to IoMusT ecosystems, including network connectivity and types of Musical Things [98], while the Smart Musical Instruments Ontology represents knowledge associated to the family of smart musical instruments [202]. Various applications have been devised which leverage these ontologies at their core (see [203], [204], [205], [206]).

Concerning the nonmusical domain, relevant ontologies include the Audio Set Ontology, which represents knowledge related to general auditory events [207]. The Audio Commons Ontology[25] is a high-level ontology binding several audio-related ontologies together, which was designed to facilitate the integration of audio content repositories on the Web as well as content consumption by software agents [208]. Ontologies in the broader sound-related domains include EBU Core[26] [209] and the MPEG-21 media contract ontology [210], while specifically for the sound domain, an early ontology has been proposed in [211], with similar structures incorporated into later works [207], [208], [212]. Ontology-aware models for sound classification have been proposed in [213] and [214], while an instrument recognition model using hierarchical structures is presented in [215].

3) Sound-Based Repositories: Several online repositories exist which feature APIs that allow the integration of their content with IoAuT and IoMusT applications. Specially relevant are those that feature content under open licenses that allow their reuse. On the audio side, Freesound [216] is the biggest repository with Creative Commons licensed audio. On the music side, Jamendo and the Free Music Archive[27] represent the most important sources containing Creative Commons music tracks. Other repositories with a broader purpose but also incorporating significant amounts of open sound content accessible through an API include the Internet Archive[28] and Europeana.[29] Of relevance is also the recent Audio Commons

[20] http://www.axoloti.com/
[21] http://musicontology.com/
[22] https://w3id.org/aufx/ontology/1.0
[23] https://w3id.org/afo/onto/1.1
[24] http://purl.org/ontology/iomust/internet_of_things/0.1
[25] https://w3id.org/ac-ontology/aco
[26] https://www.ebu.ch/metadata/ontologies/ebucore/
[27] https://freemusicarchive.org
[28] https://archive.org
[29] https://www.europeana.eu

Initiative [217], which focused on development of technologies and tools to improve reusability of Creative Commons sound content, including sound and music analyzers and ontologies for a uniform conceptualization of sound repositories.

A slightly different type of sound repositories which are also relevant for IoAuT and IoMusT applications are those featuring *static* content that can be downloaded and used offline. These include scientific research data sets, such as the MTG-Jamendo and FSD50k mentioned above [27], [28], and other collections of reusable sound content available online. While these resources do not allow online interaction such as that offered by online APIs, they are very relevant for research purposes while developing applications in the IoS domain. For example, the MIR community (see Section III-A4) maintains a broad range of public data sets[30,31] that come with audio.

4) Web-Based Digital Sound Applications: One of the most recent and widely adopted among the technologies for musical and nonmusical sound applications on the Web is represented by the Web Audio API [218]. Such a technology enables real-time sound synthesis and processing on Web browsers simply by writing JavaScript code. It is a World Wide Web Consortium (W3C) proposed standard[32] and represents a promising basis for the creation of distributed audio applications such as those envisioned in the IoS.

Several projects have demonstrated how sound-based applications can be integrated into the Web browser via the Web Audio API. A large amount of these projects have a musical nature (e.g., [219], [220], [221], and [222]), even including the implementation of MIR audio analysis techniques in the Web browser [223], and a full *DAW*.[33] Another strand of projects have focused on both musical and nonmusical sounds, such as the Freesound Explorer.[34] Conversely, other projects have only focused on nonmusical sounds, such as the online real-time sound effects synthesis platform described in [224].

In recent years, Web Audio technologies have been employed in embedded systems, thus bridging the realm of audio applications leveraging the Web with that of smart objects. For instance, Matuszewski and Bevilacqua [225] described a system comprising Raspberry Pi platforms, each running a Web Audio application. Such an application could exploit various libraries previously built for mobile-based applications (e.g., for synchronization purposes [226]), with the purpose of implementing a distributed architecture for musical performances.

Another example of system integrating Web-based digital audio technologies and embedded audio was proposed in [35] for body-centric sonic performances. The system consisted of sensor- and actuator-equipped jacket and trousers enabling the interactive manipulation of musical and nonmusical sounds retrieved from online sound repositories.

5) Networking: In Section II, we have seen that the real-time communication of audio-related information poses stringent requirements in terms of latency (delay from the transmission to the reception of packets), jitter (sudden variations of the latency) and reliability (probability of successful packet reception) [227], [228], [229]. Moreover, in some use cases the data rate is low, such as to detect crashes from accidents or simple acoustic signals, whereas in other use cases, the data rates are very high, such as when the acoustic signals carry complex information and need to be quantized at very high sampling rates. Communication protocols, such as those used in cellular wireless communications, local area network communications (e.g., Wi-Fi), and Internet connectivity have not been traditionally designed or intended for the use cases of Internet in IoS. However, in the recent year, there have been significant research efforts to design and introduce general-purpose protocols that can potentially support these "low-latency and high-reliability applications." These protocols work both at the network core level (Internet connectivity), and at the wireless access level (wireless local area and cellular wireless networks), as we briefly survey below. For a deeper discussion, see [230].

At the network core level, the Internet protocol (IP) includes lower layers protocols which greatly determine the delay, jitter, and reliability performance. One of the most prominent set of layers in this regard is Ethernet, and within Ethernet, the time-sensitive networking (TSN) [231]. TSN is a set of protocols that has been introduced to support applications, such as professional audio and video. TSN aims at supporting packet (level 2 frames) loss ratios from 10^{-9} to 10^{-12}. TSN divides the traffic in two main categories: 1) Class A and 2) Class B. In Class A, the end-to-end latency that can be supported is up to 2 ms, but for packets with limited bandwidth and very few hops in the path source-to-destination. In Class B, the end-to-end latency that can be supported is up to 2 ms, but for packets with relatively limited bandwidth and up to seven hops in the path source-to-destination. TSN requests to set up a contract between the source of traffic and the network operator, which ensures zero queueing loss and high packet synchronization. One issue with the adoption of TSN is that it is a service that will have to be required by the packet source when installing or maintaining the Internet service, usually at additional costs on top of the Internet connectivity. When the TSN cannot be used, we can resort to fall-back options, such as using user datagram protocol (UDP) protocols without bothering which lower lever protocol is implemented. However, this fall back options greatly limits the bit rates, and introduces higher packet loss probabilities, and higher jitter compared to TSNs.

At the wireless access level, we have two main technologies to meet the requirements of the IoS: 1) wireless local area networks (WLANs) and 2) cellular networks. For WLANs, we can use TSN to achieve the communication requirements that are suitable for audio services. Here, TSN offers the same pros and cons as in Ethernet for Internet. For the cellular networks technologies, in the recent years there have been major efforts to define the 5G standards. 5G is already being implemented and offers very good communication services for audio applications. This is possible, among others, thanks to three features: 1) at the physical layer; 2) the network architectural level; and 3) at the network management level.

[30] https://www.ismir.net/resources/datasets/
[31] http://www.audiocontentanalysis.org/data-sets/
[32] https://www.w3.org/TR/webaudio/
[33] https://www.soundtrap.com
[34] http://labs.freesound.org/fse/

1) At the physical layer, one of the major novelty is the introduction of the millimeter waves communications [232], which make it possible to transit very high data rates from the audio source at some wireless devices up to the base station and back, with rates of the order of Giga bits per second and latency below 1 ms.
2) At the network architectural level, one major novelty is the concept of edge computing [233], where the idea is to place computational resources near the wireless access points (base stations) so that wireless devices that need to perform complex computations can devolve them to near located edge computing nodes [multiaccess edge computing (MEC)], with major delay gains as compared to sending these computations to the cloud. Moreover, the edge nodes can store content near the wireless devices, so that the delay to fetch such content may be limited.
3) Finally, the third major novelty of 5G at the network management level is the concept of network slicing [234], which allows to transmit simultaneously on the physical wireless network several classes of traffic each having quite different requirements in terms of latency and reliability. In practice, network slicing allows to use sets of communication protocols that are tailored to the specific uses cases, such as audio or other real-time services.

The future composition of TSN and 5G protocols (when the source is a wireless device that needs to be connected to the Internet) will therefore allow to arguably meet most of the communication requirements of the IoS. Recent research in this space has proposed dedicated architectures to interconnect distributed musicians over wireless links along with analysis of simulations under different conditions as well as real-world applications of such architectures [82], [229], [235], [236].

6) Synchronization: In several IoS applications the distributed nodes need to be synchronized in time. Nevertheless, the accuracy of such a synchronization shall depend on the application at hand. To keep a satisfactory level of synchronization between the nodes, developers focus shall be devoted to the control of two quantities: 1) the *local time* of each node and 2) the *delay*, i.e., the amount of time needed by the node to record, synthesize or playback an audio signal once the request to do so have been received. A complicating factor is that IoS nodes could slightly differ even in presence of the same hardware and software, and even minimal differences in parameters, such as sampling rates in the long run can result in clock drifts and therefore cause synchronization issues [237]. Moreover, even if different devices in the network would initially share the same clock, they need a resynchronization procedure from time to time.

In general, Quality of Service (QoS) is ensured minimizing the two quantities: 1) σ_t the variance of the difference between the local time of each node and 2) σ_d the variance of the difference between the delays of each node. To show the importance of such quantities, in the following we describe three use cases with growing requirements in terms of synchronization accuracy.

1) In WASN, synchronization is typically needed to interpret some high-level behaviors happening across different nodes. In this case, σ_t and σ_d shall remain below the second.
2) On contrary, distributed playback systems that operate over IP [238], such as RAVENNA [239] or Dante [240], reducing σ_t 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 since the nodes are simple playback systems that are not in charge of audio processing or synthesis.
3) Smartphone [241] orchestras, laptop [242], [243] orchestras, or any NMP systems involving synchronization [5] represent a much more challenging case. 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 hardware platform. The latter calls for software-based solutions [244] which, however, are known to be inherently limited in terms of precision [245].

Different systems and protocols have been developed by the IoT community to minimize σ_d as well as to address the issue of establishing an accurate, network-wide notion of time, which is crucial for scenarios demanding precise temporal coordination [246], [247]. Various systems have attempted to improve the weakness of the Network Time Protocol, which is widely used in sensor networks [248]. A particular focus has been placed on minimizing synchronization errors between nodes of wireless sensor networks (see the Flooding Time-Synchronization Protocol [249], PulseSync [250], and variations of it [251]).

Synchronization issues are particularly relevant to NMPs occurring in both wide and local area networks and in both wired and wireless networks [36], [40], [62]. Several NMP systems have been devised for WLANs, typically leveraging Wi-Fi and using protocols to exchange musical messages between devices, such as MIDI or Open Sound Control [98], [252], [253], [254]. Synchronization aspects in WLANs have been addressed by various studies in [1], [36], [255], [256], and [257]. An example is represented by the approach based on HTML5 proposed in [226] to synchronize mobile-based applications leveraging the Web Audio framework. To date, the most widely adopted synchronization protocol for musical applications within Wi-Fi-based WLANs is Link, a de-facto standard developed by the company Ableton [258]. However, recent research has assessed the limits of such a protocol in supporting a large number of nodes [245].

Moreover, a significant body of research has investigated at the technical and perceptual level the use of metronomes in NMPs distributed over the Internet. For instance, dedicated hardware that broadcast GPS reference time over the network has been proposed in [66], [259], and [260], while adaptive metronomes have been proposed in [65] and [261].

IV. Challenges

In this section, we discuss open questions that currently hinder the development of the IoS. In particular, we describe the challenges that are common to both the IoMusT and the IoAuT, as well as those that are specific to such fields.

A. Embedded Audio

As the IoS emerges, sound-specific operating systems are required on embedded hardware to ease development and portability of IoS applications. Most of current embedded systems specific to sound processing offer a little range of connectivity options and scarce hardware-software methods supporting advanced ML algorithms. In the IoS 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.

Current challenges in this area include: 1) platform independence (i.e., the operating system should support a number of hardware platforms as great as possible, to foster software portability); 2) low-latency audio and sensor processing (especially for IoMusT applications); 3) support wireless connectivity options, in particular the latest generation of cellular network; 4) development of methods able to optimize the consumption of resources (e.g., memory, processing, and power); and 5) enabling ML (primarily inference, but also adaptation procedures or retraining) on edge devices by dedicated hardware accelerators.

B. Machine Listening

The recent availability of large amounts of audio recordings has fostered research on the use of ML methods to gather both low and high-level information about various aspects of musical and nonmusical contexts. These endeavors fall within the remits of the so-called field of "Machine Listening."

Concerning musical contexts, the field of MIR investigates computational methods to extract information from musical content [42], [262]. One of the major branches of this area deals with the analysis of audio signals captured by microphones, from recordings of single musical instruments to recordings of large ensembles. One of the challenges concerning MIR in the context of the IoS paradigm regards the real-time aspect and the use of embedded systems to perform the computations. To date, the majority of MIR research has focused on offline methods analyzing large data sets of audio recordings [263]. The availability of good MIR techniques for real-time contexts, especially on embedded systems, is scarce. This is due not only to the stringent requirement on the processing time to report the wanted information, but also to the fact that pre- and post-processing of the audio signal, typically involved in offline methods, cannot be performed. An area of application of embedded real-time MIR is interactive systems for musical performance [264] such as smart musical instruments [4], where the extracted information is repurposed, with unperceivable latency for the human player, into digital sounds. Current real-time MIR methods and related data sets are severely affected by temporal inaccuracies. Therefore, novel temporally accurate data sets are needed to advance the state-of-the-art in this space, along with innovative information retrieval algorithms conceived for real-time usage.

Along the same lines, the development of ML approaches to predict the future evolution of an audio signal in real-time, based on past audio samples, would find application in NMP scenarios as an alternative approach for packet loss concealment [265]. Nowadays, recovery mechanisms included in audio codec implementations are adopted when a packet carrying audio data gets lost or is received too late to be reproduced in due time, but they introduce additional processing delays which add up to the overall perceived mouth-to-ear latency. Moreover, in a long term and more "visionary" scenario, ML algorithms could even become capable of learning to anticipate what the musician will play ahead of time, thus being able to reproduce a predicted version of the audio stream *before* the real audio data generated by a remote performer is received. Such advanced artificial intelligence capabilities would definitively eliminate any delay-induced usability limitation.

Regarding nonmusical contexts, as described in Section III-B1 several efforts have been conducted to extract meaning from soundscapes (i.e., sonic environments), particularly in urban areas [173] and wildlife monitoring. While those two application areas are now well established, one recent and promising field of application for ASNs is the surveillance of industrial processes for security and quality control. The main scientific challenge here is that, due to the complexity of the production chain, all possible abnormal behaviors can hardly be modeled. Alternatively, one has to numerically define what is a normal behavior and raise an issue when the recorded behavior significantly departs from it. Quantifying and thresholding this level of difference between the observed data the model of normality the system have in memory is, at this time, a very open research question.

New tasks have been proposed in the DCASE challenge to foster research on this topic [266] and the body of knowledge is gaining momentum, as demonstrated by a systematic research study done by considering seven journals and 16 conferences relevant to the field [267]. Within those publication venues, the study shows that the number of publications on this topic is approximately doubling each year since 2015.

C. Semantic Audio

With the proliferation of "smart" devices featuring advanced information interaction capabilities and access to large amounts of data, there is increasing demand for such features in the context of IoS. This calls for the use of state-of-the-art audio and music analysis and processing techniques in IoS devices provided by Semantic Audio and MIR (see Sections III-A4 and III-C2). These fields, however, are increasingly dominated by deep learning solutions [116], [268], [269].

Although the available computing power is fast increasing in IoS and edge computing devices, with more and more innovative AI accelerators proposed [270], decoding large neural networks still presents a problem in certain applications, particularly where computational, bandwidth, or energy resources are limited, or where hard real-time criteria need to be observed, e.g., for efficient interaction between human and machine [104]. There are currently no clear guidelines and established methodologies for optimizing applications

given specific technical criteria and user requirements. For example, in [85] edge and cloud computing solutions for semantic audio analysis are painstakingly compared with traditional keyword-based search to determine the tradeoff between these techniques with respect to user requirements. Similar experiments would be needed to establish the optimal solution for most applications involving MIR or Semantic Audio techniques in IoS applications.

Test-beds providing diverse annotated data and platforms to easily test algorithms in different application contexts would help breaking the barriers in applying advanced audio analysis and processing techniques in IoS. Advanced methodologies for optimizing neural networks for specific hardware would also be beneficial to overcome these barriers. Particularly in the context of IoAuT, model compression [271], and quantization [272] have been noted as a promising way forward in [2]. Further challenges in edge computing in IoS are discussed in Section IV-D.

Data interoperability is another substantial barrier in creating complex IoS applications. The ontologies discussed in Section III-C2 provide a mechanism for shared conceptualizations that alleviate the problem of matching the "meaning" of data and metadata items in complex networked applications with multiple stakeholders and solution providers. There are ontologies ripe for use in the audio and music domains, and their utilities have been demonstrated across diverse IoS relevant applications [48], [49], [198], [203], [204], [205], [206], [273]. However, it can be observed that no ontology is optimal for any given application, since tradeoffs are necessary for viable shared conceptualization [190]. As a solution to this problem, the development of flexible ontology frameworks that allow solution-specific components to be plugged into higher-level models that include only foundational classes and relations is proposed in [190], while a layered approach to information modeling is proposed in [197] in the music and audio domains.

A further related challenge is the efficiency of ontology-based solutions. This can be divided into efficiency challenges related to data exchange, and challenges related to storage, querying, and inference. In the context of data exchange, i.e., the syntax used for representing semantically enhanced information, verbose XML-based formats such as RDF/XML[35] have been criticized, with Turtle[36] and later JSON-LD[37] emerging as broadly supported solutions. In the context of IoS, a further consideration is the optimal granularity at which semantics should be embedded in an application. While binary and very concise exchange formats such as Open Sound Control [274] with weak underlying semantic models are popular and necessary in some IoS contexts, data-efficient semantic information exchange formats are also emerging. This includes binary serialization formats for semantically enhanced data such as Header Dictionary Triples (HDT/RDF)[38] [275].

[35]http://www.w3.org/TR/rdf-syntax-grammar/
[36]https://www.w3.org/TR/turtle/
[37]https://json-ld.org/
[38]https://www.rdfhdt.org/

Regarding querying and complex machine-to-machine communication, the Constrained Application Protocol [276] have been proposed as a lightweight IoT protocol using semantic technologies [95]. This has been demonstrated to cope with efficiency constraints in a musical context in [96] and [97].

Inference remains an ongoing challenge since reasoning with very large knowledge bases is known to be computationally expensive. Solutions to this problem include restricting the expressivity of knowledge representation, e.g., through the use of a specific profile of an ontology language such as OWL EL[39] which allows reasoning in polynomial time. Deep learning-based solutions for reasoning with ontologies are also emerging [277]. The granularity at which semantic representation is used is another important consideration in the IoS. For example, using an ontology to annotate large blobs of homogeneous data, instead of individual data items, may be a good tradeoff in some IoS applications. Further semantic audio-related challenges in the context of interoperability and storage are discussed in Sections IV-I and IV-H.

D. Audio Detection on the Edge

Edge computing is a distributed computing paradigm that brings computation and data storage closer to the sources of data. In ASNs, there are numerous advantages in concentrating most of the processing on the sensors. Computational inference of quantities of interest on the edge reduces the bandwidth required to transmit the data, and most of the time reduces privacy issues. The latter arises because a greater level of control can be performed since the data transferred is usually more compact.

To do so, the sensor has to embed computational components which must be carefully chosen to wisely balance energy consumption and processing capabilities. While sensing and features computation are relatively low consumption processes, powerful state-of-the-art machine listening detectors based on deep learning models have to be simplified to fit memory and energy consumption requirements [278]. Even with such compact embedded software systems, sensors are heavily dependent on a reliable and continuous source of energy. That means that the sensor has to be wired or has to carry powerful batteries to overcome the nonavoidable down times of energy harvesting components like solar panels.

Resorting to embedded power storage reduces the ease of maintenance and increases the dependency on polluting materials. Those issues call for innovative solutions such as batteryless designs [151]. While this kind of hardware potentially allows simpler and more reliable networks designs, it also opens an interesting challenges in sparse network management to be able to produce reliable estimates of quantities of interest from sources of information that are only intermittently able to record, process, and transmit.

E. Networking Architectures

To date, actual IoS deployments over 5G networks remain scarce, and very limited statistical results are available on the

[39]http://www.w3.org/TR/owl-profiles

actual latency and reliability of 5G networks for IoS scenarios [82], [236]. There is the need to design, implement, and evaluate novel networking architectures exploiting the new capabilities offered by 5G to support IoS services and applications. Research should address fundamental questions, such as how to design 5G slices dedicated to low-latency and high-reliable transmissions, which computations can be offloaded from local devices to MEC servers, what is the interplay between MEC and cloud servers, and how to optimize the position of MEC servers even in a dynamic fashion.

Furthermore, IoS research could address emerging computing paradigms such as in-network computing. To date only a handful of IoS studies have investigated such avenue [279].

F. Networked Music Performances

For what concerns NMP solutions, the adoption of 5G wireless transmission technologies would ensure access delay to the backbone telecommunication infrastructure below 10 ms. Such figure is compliant to the typical latency requirements of NMP while fostering flexibility and portability of NMP setups, which are nowadays constrained by the need of leveraging cabled Ethernet connections for local area network connectivity (note that Wi-Fi wireless connections are typically avoided for NMP purposes, as they introduce too high jitter). As further step, involving Internet service providers and network operators in the development of commercial NMP services would allow the adoption of prioritization criteria for NMP-generated audio/video data (e.g., in the context of software-defined networking design approaches), thus providing to the users adequate QoS guarantees in terms of latency and jitter.

Video conferencing has advanced dramatically with the introduction of ultrareliable low-latency communication (URLLC) in 5G and Nvidia's radical new idea, where only key points in the video are transmitted and then deep neural networks (DNNs) reconstruct artificial video representations, which have very high similarity with the original video [280]. The challenge would be to exploit similar tricks for the audio content, where key points in the acoustic scenes are transmitted, and then DNNs would be constructing artificial representations of the significant parts of the original acoustic scenes.

A challenge is to change the current paradigm where users interact with their Sound Things via individual interfaces and networks, into a more sound service-oriented and network-driven paradigm, where any sound technology can be seamlessly accessed via a common IoS software layer. In principle, users only see the sound services and not the underlying Sound Things that automatically organize themselves, analyze the acoustic scenes, and exchange information with each other.

G. Distributed Machine Learning Over Networks

ML is a research area that is increasingly gaining attention to design, optimize, and manage sound-related systems. Although ML is well established within computer vision and speech and text analysis, ML faces new major challenges when it comes sounds and Internet. In an IoS ecosystem, all the involved units generate distributed, heterogeneous, and imbalanced data that cannot be easily collected timely due to the bandwidth of the communication protocols, the economic unsustainability to deploy high bandwidth communications, and privacy concerns.

Unfortunately, to achieve the highly desirable ML-based targets for IoS services, we cannot simply apply existing ML. The prominent successes of ML are largely in the domains of images, speech, where the availability of large data sets and popular platforms that can provide vast amounts of dedicated computational and communication resources in centralized setups such as data centers. Such assumptions challenge the distributed, networked, and real-time nature of the IoS [281]. On the one side, ML methods and algorithms are not yet mature for being used in the IoS domain. On the other side, it is not clear up to which extent the ML predictions can be realistically used in an IoS data communication architecture. Specifically, in the IoS, we face two prominent challenges: 1) unbalanced data sets and 2) distributed data sets, as we survey below.

1) The first challenge for ML-based IoS systems is that the predictions have to be done at locations or times from which noisy or sparse/partial measurements are available. ML achieves impressive performance mostly on the data that contains no missing values and have balanced classes. In IoS data sets (for both time-series data or statistical data), missing data commonly exists due to the distributed devices connected over communication networks. IoS data suffer from a number of unpredictable impairments since the hardware and software of the involved units will be produced by different vendors with different standards and different costs, memory, and computation [282]. Data sets may also be imbalanced, which may increase the risk of having the so-called bias in the training pipeline. Despite recent advancements in connectivity solutions for IoT, the existing methods cannot properly address how to handle heterogeneous types of data. A recent survey [283] showed that deep learning methods achieve good performance, among other alternative approaches, for most time-series classification tasks with missing data. Unfortunately, the existing methods are usually limited to nonsparse and balanced data sets, which we may rarely see in cyber–physical systems such as the IoS.
2) The second challenge for ML in IoS is that the data sets are distributed and must be connected over a public communication network, including via Edge computing [281], [284]. In these solutions, ML services will have to be solved by distributed algorithms, where the heavy coordination and computation procedures of ML are much hindered by bandwidth limitations, latency, and message loss. The computational capability and storage of the IoS devices challenges the use of heavy pretrained models. The communication networks enforce inference with partial knowledge. The challenge here is to support self-adaptive ML-based applications composed of edge and cloud modules. A promising direction to address these issue is the design of novel wireless and

wired communication protocols tailored for distributed ML [285].

H. Storage

Storage is an important component of the IoS ecosystem with relevant open challenges. Fast access from embedded devices to content stored in cloud repositories that can provide satisfactory user experiences when working with audio in real-time and consuming audio from the cloud, is one of the basic challenges that can be addressed with the adoption of faster network technologies such as 5G. Moreover, interoperability between different cloud content providers is also an open challenge. In a rich IoS ecosystem, different processes running in different devices and under different platforms should be able to interact with storage solutions using common protocols and APIs.

In addition to that, the traceability and authentication of audio content distributed in the IoS ecosystem is a substantial challenge that needs to be addressed for successful deployments of IoS applications. To that end, *blockchain*-based technologies have already been envisioned to verify the integrity of original audio content and to enable its secure distribution, as well as have been implemented as a proof-of-concept within the paradigm of IoAuT [286]. However, challenges remain open, including the traceability of audio content usage in the IoS ecosystem.

I. Interoperability and Standardization

The result of our survey of the IoS field reveals that, to date, active research on IoS-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. However, 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 successful adoption of IoS technologies, a recurring issue within the more general IoT field [46].

Within the IoS, different types of musical and nonmusical devices targeting a variety of stakeholders are utilized to generate, detect, or analyze sonic content. Those devices need to be able to dynamically discover and spontaneously interact with heterogeneous computing, physical resources, as well as digital data. Challenges related to their interconnection include the need for ad-hoc protocols and interchange formats for sound-related information that have to be common to the different Sound Things, as well as the definition of common APIs specifically designed for IoS applications. Semantic technologies, such as Semantic Web [287] and knowledge representation [288] have arguably the potential to become a viable solution to enable interoperability across heterogeneous Sound Things. However, to date, only a few ontologies exist for the representation of the knowledge related to IoS ecosystems.

The Moving Picture Experts Group was a working group of ISO/IEC and has since 1988 been instrumental in driving the development and standardization of audio compression, synchronization, processing, transmission, and file formats for a wealth of applications [289]. Recently new organizations, such as the Moving Picture, Audio and Data Coding by Artificial Intelligence has been formed in order to further push the technology toward artificial intelligence data coded standards [290]. Standards targeting the combination of communications and audio for personal area networks include Bluetooth LE [291], [292]. A challenge would be to further develop the standardizations and interoperability within the intersection of communications and audio beyond personal area networks.

A common operating system for Sound Things can be considered as a starting point for achieving interoperability. As surveyed in Section III-C1, recent technological advances in the embedded audio field have led to the creation of platforms that are suitable for IoS applications. Their wide adoption and deployment are therefore expected to enable advanced IoS ecosystems and has the potential to foster interoperability within and even between the musical and nonmusical sides of the IoS.

J. Quality of Service

The desire to distribute audio and audio-related data across IoS devices imposes application-specific service requirements of the IoS architecture.

In general, the reliability of the data delivery is defined under the umbrella term QoS. For telecommunication application, the telecommunication experts of the International Telecommunication Union defines QoS as the "[t]otality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service" [293]. This definition also holds for the IoS in which users could be humans or machines, e.g., in form of sound creators or listeners. For instance, for a telematic musical event with co-located musicians and audiences, the need to exchange audio signals or parametric control data over the network without perceivable artifacts is crucial. Perceivable artifacts can be caused by packet loss, transmission delay, and limited bandwidth. Packet loss may occur when utilizing an unreliable data transport protocol, like the UDP, or when the audio stream via a reliable transport mechanism, like the transmission control protocol (TCP), gets interrupted. Without proper mitigation or error concealment, audible artifacts, such as spectral distortion, drop-outs, pops, or crackles reduce the audio quality and consequently affect the listening experience. An end-to-end low-latency transmission chain enables remote musicians to perform as if they were in close proximity. It has been shown that a round trip delay as low as 40 ms already starts affecting the ability to perform together (see also Section III-A1) and a latency of above 150 ms will degrade a communication scenarios [294]. Especially in wireless networks, transmission latency is often time variant which needs to be compensated by additional audio buffers. This adds delay for real-time audio applications. The bandwidth limitations of many IoS network protocols, such as Bluetooth or Wi-Fi demand perceptual audio signal compression. The processing latency caused by the signal encoding and signal decoding contributes to the overall

transmission latency. Low-latency audio codecs have a coding delay as low as 5 ms [295], [296]. Generally, parametric control data (e.g., MIDI and Open Sound Control) require relatively little bandwidth compared to audio data. When transmitted together, the QoS has to ensure that control data and audio data are kept synchronized across the IoS network.

Many IoS devices are hardware-constrained and have limited capabilities to connect. Further, mobile IoS devices need to be resourceful with their limited battery power which may lead to throttled data processing and transmission. At the same time, network conditions may change over time and can become unstable. Supporting the high QoS requirements across time-varying network conditions and across devices that may have different processing and network capabilities poses a cross-disciplinary challenge for the IoS.

K. Ethics, Privacy, Security, and Sustainability

Notwithstanding the numerous societal, economic, and artistic benefits that the IoS promises, the ubiquitous nature, and increased autonomy of Sound Things raise concerns about the ethical compliance of the associated services. We need to incorporate ethics in the IoS so that the services provided do not infringe on the ethical rights of its beneficiaries.

Nevertheless, the study of ethics in music and audio technology is not yet well established, with only a few authors dealing in recent years with the importance of related ethical dilemmas. Researchers in the musical domain have questioned the practices of music streaming services in monitoring users and inducing behaviors, warning about the associated risks [297], [298]. Other authors have examined the ethical dimensions of the field of MIR, arguing that such technology is not value-neutral but is influenced by design choices, and so has ethically relevant implications [299]. In the field of New Interfaces for Musical Expression, some authors have discussed political issues inherent in new musical instruments [300]. Similar statements can be made for the audio-technology domain.

All these research strands are also relevant to the IoS field, although to the authors' best knowledge no investigations have been conducted yet on such topics at the scale of IoS. Ethical research in IoT has identified major issues, such as privacy, security, transparency, trust, social equity, social equality, and responsibility according to law. Research has also identified the factors that can increase the acceptance of IoT and proposed a set of guidelines to interact with IoT from a social perspective [301]. There is a strong need to follow some policies that support social issues also in the IoS in order for it to be socially acceptable and undertaken in the public interest. Nevertheless, the IoS is distinguished from the general IoT field for its sound-specific focus. While some technical and societal challenges are common to the two fields, IoS poses music- and audio-specific challenges that cannot be simply addressed by using the same methods and tools of IoT.

Sound Things have the ability to automatically collect, analyze, and exchange personal data related to their users, and consequently, they can expose users to breaches of security and privacy, which are common to other IoT subfields [20], [302]. Research in this domain is already emerging, in particular in the area of speech, where recent works have addressed the interpretation of legal frameworks [303], the challenge of quantifying privacy [304], and by organizing challenges, such as VoicePrivacy [305], to research and develop solutions to particular problems.

It is important to consider that users have an intuition and can be assumed to be competent to gauge their level of privacy and security only when it relates to human capabilities [306], [307]. However, where computer systems go beyond human, to super-human capabilities, users generally cannot be expected to have an intuition with respect to their security and privacy when they interact with the system. As an analog, it is important that passengers of an airplane *feel safe* (i.e., that they have an intuition of safety), but it is *more* important that they actually *are safe*, though we cannot reasonably expect passengers to be competent in evaluating whether they are safe or not. Similarly, it is thus important that regulations and guidelines are developed to govern the security and privacy of IoT. Such guidelines are already available or in development for voice interfaces, e.g., [308] and [309].

The range of problems related to, and approaches for preserving privacy in IoS and IoT is vast. For example, data intentionally shared can unintentionally *leak* related information or, as a form of *function creep*, data can be used in unexpected ways. Moreover, users' ignorance can lead to breaches, they can be tricked or coerced to share information, or their hardware or service providers can be hacked [310]. As a general policy, it is therefore important to develop products and services using the privacy-by-design approach, where the design and development process includes, from the beginning, impact assessment of security and privacy.

Application of privacy and security standards thus has to be included in all parts of systems design, including the acoustic environment, acoustic and hardware design of device, edge software, communication links, device interaction, cloud services, as well as user interface and service design. One approach to ensure the privacy of information consists of the definition of privacy policies. Sound Things could be equipped with machine-readable privacy policies, so that when they come into contact they can negotiate privacy policies for before communicating [311], [312].

Taking advantage of distributed computational resources in an IoS context seems to bear the risk of giving away control over shared data. For instance, transmitting audio signals to a cloud service for acoustic signal enhancement exposes the audio signals to the cloud provider and to the algorithm developer and could potentially enable unwanted data usage. There are well-established data encryption methods for protecting data from unwanted access while stored and while transmitted over the Internet (e.g., HTTPS and MQTT). However, protecting data during the actual processing remains challenging. Ensuring privacy during data processing is the field of secure signal processing (SSP) and privacy-preserving ML, respectively. SSP allows mathematical operations to be executed on encrypted data without the need for prior decryption. The foundation of SSP is homomorphic encryption, which is a special class of data encryption methods. Here,

one or more specific mathematical operations performed on the encrypted data corresponds to the mathematical operation performed on the original data. The result of the operation remains encrypted and cannot be decrypted without proper credentials.

In the audio context, homomorphic encryption has been used for audio analytics (e.g., privacy-preserving speaker verification and identification [313]) as well as signal processing tasks (e.g., convolution [314]). Besides the audio input data, it has been shown that even the actual algorithm can remain encrypted, which protects both the data provider and the algorithm developer from unwanted information sharing with the platform provider. Homomorphic encryption is also applied to ML methods, such as logistic regression [315], XGBoost [316], or even DNNs [317]. The main challenge of SSP using homomorphic encryption is the computational complexity. For instance, the inference on the privacy-preserving ML model in [316] is 400 times slower than its conventional version.

In a different vein, it is paramount to be aware of the adverse impact of the current IoS technology on the environment in terms of greenhouse gases emissions, pollution, and soil consumption. To date, research on sustainability aspects of the IoS is limited. The study reported in [318] has recently provided a survey of the environmental issues produced by current information and communication technology and related these to the use cases that the IoS envisions. On the basis of this survey, the authors identified some key aspects to reduce the footprint of IoS services and products and then provided suggestions to make advancements in IoS environment aware.

L. Blockchain Technologies

Blockchain is an emerging technology that is impacting several industries, including the creative industries and those operating in the IoT [319], [320], [321], [322]. The IoS vision requires, above all, IoT features, such as decentralization, seamless authentication, transparency, data integrity and privacy, and self-maintenance, as well as the musical domain's feature, such as efficient handling of copyrights and speed of royalties payment. Such features can be brought by blockchain, but its integration in the IoS has not been investigated thus far.

Recently the integration of IoMusT and blockchain technology has been proposed in [323], where several examples of use cases have been provided. For instance, in the "cover song identification" scenario, authors envisioned a Sound Thing, equipped with microphones and machine listening algorithms dedicated to the identification of the cover of songs, which is used by an inspector of a National Rights Society at a music venue during a live concert. The Sound Thing automatically verifies whether the played songs match the titles declared by musicians in the list of music pieced to be sent to the National Rights Society. As soon as the match is verified the composers of those music pieces are immediately rewarded thanks to the use of a smart contract deployed on the blockchain underlying the system.

V. IoS: Research Agenda

The above sections have shed some light on the current challenges that prevent the IoS field to flourish. In this section, we collect together the different aspects and open questions that need to be answered in order to create, design, use and finally evaluate IoS systems. This may be seen as a roadmap that we hope would be addressed in the ongoing research of the emerging IoS community.

1) To progress the design of embedded platforms specific for the IoS, including the integration of methods for low-latency processing, and the support to the latest generation of cellular networks.
2) To progress the design of reliable, autonomous sensing devices to better monitor and understand our environment through the audio modality.
3) To progress the design of Sound Things, with new solutions for the analysis of sound-related information based on the edge computing paradigm and the most advanced machine listening approaches.
4) To advance the current connectivity infrastructure, with the implementation of novel interoperable protocols for the exchange of sound-related information.
5) To define standards (e.g., for protocols, shared ontologies, formats, and APIs) that will allow one to reduce fragmentation and facilitate interoperability among Sound Things as well as the services they offer including audio processing and storage; such endeavors could entail the creation of dedicated Web of Sound Things architectures.
6) To create advanced IoS ecosystems, in both the musical and audio domain, which include enabling technologies and communities of users interacting with them.
7) To investigate ethical concerns and define appropriate measures to address them; this includes the definition of principles for an Ethical IoS that can inform design, development, and evaluation of IoS ecosystems, their hardware and software components, and the interactions of stakeholders; moreover, this entails tackling the challenges of: a) privacy and security of personal data, with a "privacy by design" approach; b) sustainability aspects at all levels, from production to distribution; and c) inclusiveness and accessibility, conceiving systems for various categories of users.
8) To explore the integration of haptic feedback and motion tracking mechanisms in those IoS scenarios where gestural data play a fundamental role (e.g., in the context of remote music teaching) or when enhanced immersion in artificial soundscapes is required.
9) To devise methods capable of minimizing the need for the user to configure the Sound Things: the Sound Things should self-configure and automatically adapt to changing environments and use cases.
10) To ease the adoption by the industry of IoS-related technologies by providing end-to-end systems that are easy to deploy and to adapt to new needs.
11) To investigate novel network communication protocols for the IoS (e.g., TSN over 5G) that will be compatible with the network slicing concept introduced in 5G.

12) To establish distributed ML tasks (training and inference) that are robust to the data missing and data imbalance problems and that can run over optimized or ad-hoc communication protocols so that the computations will be reliable and with limited latencies.
13) To develop cognitive processing methods to fuse acoustic signals from ad-hoc microphone arrays into a single high-quality audio stream.
14) To design processing frameworks that allows dynamic distribution of audio processes within the network with service guarantees e.g., to take advantage of both low-latency edge computing and scalable cloud processing.
15) To devise new temporally accurate data sets and methods for the real-time analysis and classification of musical and nonmusical sonic content on embedded systems.
16) To integrate the blockchain into the IoS, and create effective and efficient applications based on such integration, such as secure audio distribution and traceability.

VI. Conclusion

In this article we introduced the paradigm of the IoS, highlighting its unique characteristics that differentiate it from to the IoT, and identifying the major challenges and requirements that need to be addressed for reaching its full potential. We introduced a definition for such a new subfield of the IoT, and placed it in the context of neighboring fields. In the IoS paradigm, which merges under a unique umbrella the emerging fields of the IoMusT and the IoAuT, heterogeneous devices dedicated to musical and nonmusical tasks can interact and cooperate with one another and with other things connected to the Internet to facilitate sound-based services and applications that are globally available to the users. We presented a vision for this emerging research field, which is rooted in different lines of existing research, including sound and music computing, IoT, machine listening, semantic audio, artificial intelligence, and human–computer interaction. The IoS relates to wireless networks of intelligent devices dedicated to musical and nonmusical purposes, which enable, in both co-located and remote scenarios, various forms of interconnection among different stakeholders.

The IoS vision not only offers several unprecedented opportunities, but also brings technological and nontechnological challenges that both academic and industrial research will need to address in upcoming years. The realization of the proposed IoS vision would ultimately benefit society, by providing several novel possibilities, which include musical interactions between geographically displaced performers and audiences, a widespread use of ambient intelligence systems employed to monitor environments in smart cities, and in general novel ways of interacting with sounds across the network (such as sound-based therapies involving remotely connected users).

We recognize that substantial standardization efforts are also necessary. Just like for the general IoT field, the success of the IoS strongly relies on standardization requirements, which are currently unmet. The definition of standards specific to the IoS (for platforms, formats, protocols, and interfaces) will allow to achieve interoperability between heterogeneous systems. Issues related to security and privacy of information, which are also common to the IoT, need to be addressed, especially for IoS systems deployed for the masses. Moreover, 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 IoS.

References

[1] 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.
[2] L. Turchet, G. Fazekas, M. Lagrange, H. S. Ghadikolaei, and C. Fischione, "The Internet of Audio Things: State-of-the-art, vision, and challenges," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 10233–10249, Jan. 2020.
[3] F. Font, "SOURCE: A Freesound community music sampler," in *Proc. Audio Mostly Conf.*, 2021, pp. 182–187.
[4] L. Turchet, "Smart musical instruments: Vision, design principles, and future directions," *IEEE Access*, vol. 7, pp. 8944–8963, 2019.
[5] A. Migicovsky, J. Scheinerman, and G. Essl, "MoveOSC—Smart watches in mobile music performance," in *Proc. Joint Int. Comput. Music Conf. Sound Music Comput. Conf.*, 2014, pp. 692–696.
[6] S. Thorn, "Telematic wearable music: Remote ensembles and inclusive embodied education," in *Proc. Audio Mostly*, 2021, pp. 188–195.
[7] L. Turchet, T. West, and M. M. Wanderley, "Touching the audience: Musical haptic wearables for augmented and participatory live music performances," *Pers. Ubiquitous Comput.*, vol. 25, no. 4, pp. 749–769, 2021.
[8] 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.
[9] 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.
[10] L. Vignati, S. Zambon, and L. Turchet, "A comparison of real-time Linux-based architectures for embedded musical applications," *J. Audio Eng. Soc.*, to be published.
[11] M. Maier, M. Chowdhury, B. Rimal, and D. Van, "The tactile Internet: Vision, recent progress, and open challenges," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 138–145, May 2016.
[12] Z. Sheng, S. Pfersich, A. Eldridge, J. Zhou, D. Tian, and V. C. Leung, "Wireless acoustic sensor networks and edge computing for rapid acoustic monitoring," *IEEE/CAA J. Automatica Sinica*, vol. 6, no. 1, pp. 64–74, Jan. 2019.
[13] M. Satyanarayanan, "Pervasive computing: Vision and challenges," *IEEE Pers. Commun.*, vol. 8, no. 4, pp. 10–17, Aug. 2001.
[14] V. Lazzarini, D. Keller, N. Otero, and L. Turchet, *Ubiquitous Music Ecologies*. London, U.K.: Routledge, 2020.
[15] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational Analysis of Sound Scenes and Events*. Heidelberg, Germany: Springer, 2018.
[16] Y. Rogers, H. Sharp, and J. Preece, *Interaction Design: Beyond Human–Computer Interaction*. Hoboken, NJ, USA: Wiley, 2011.
[17] S. Furui, *Digital Speech Processing, Synthesis, and Recognition: Synthesis, and Recognition*. New York, NY, USA: Marcel Dekker, 2018.
[18] E. Rubio-Drosdov, D. Díaz-Sánchez, F. Almenárez, P. Arias-Cabarcos, and A. Marín, "Seamless human-device interaction in the Internet of Things," *IEEE Trans. Consum. Electron.*, vol. 63, no. 4, pp. 490–498, Nov. 2017.
[19] Y. Moon, K. J. Kim, and D.-H. Shin, "Voices of the Internet of Things: An exploration of multiple voice effects in smart homes," in *Proc. Int. Conf. Distrib. Ambient Pervasive Interact.*, 2016, pp. 270–278.
[20] T. Bäckström, "Speech coding, speech interfaces and IoT-opportunities and challenges," in *Proc. IEEE 52nd Asilomar Conf. Signals Syst. Comput.*, 2018, pp. 1931–1935.
[21] M. Bunz and G. Meikle, *The Internet of Things*. Hoboken, NJ, USA: Wiley, 2017.
[22] 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.
[23] M. Noura, M. Atiquzzaman, and M. Gaedke, "Interoperability in Internet of Things: Taxonomies and open challenges," *Mobile Netw. Appl.*, vol. 24, no. 3, pp. 796–809, 2019.

[24] 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*. Heidelberg, Germany: Springer, 2011, pp. 97–129.

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

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

[27] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, "The MTG-Jamendo dataset for automatic music tagging," in *Proc. Mach. Learn. Music Disc. Workshop Int. Conf. Mach. Learn. (ICML)*, Long Beach, CA, USA, 2019, p. 19. [Online]. Available: http://hdl.handle.net/10230/42015

[28] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An open dataset of human-labeled sound events," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 829–852, 2021.

[29] B. De Man, M. Mora-Mcginity, G. Fazekas, and J. D. Reiss, "The open Multitrack Testbed," in *Proc. 137th Conv. Audio Eng. Soc.*, Oct. 2014, p. 14.

[30] J. Cáceres and C. Chafe, "JackTrip: Under the hood of an engine for network audio," *J. New Music Res.*, vol. 39, no. 3, pp. 183–187, 2010.

[31] 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.

[32] L. Turchet, A. McPherson, and M. Barthet, "Real-time hit classification in a smart Cajón," *Front. ICT*, vol. 5, p. 16, Jul. 2018.

[33] L. Turchet and A. Zanetti, "Voice-based interface for accessible soundscape composition: Composing soundscapes by vocally querying online sounds repositories," in *Proc. Audio Mostly Conf.*, 2020, pp. 160–167.

[34] L. Turchet, S. J. Willis, G. Andersson, A. Gianelli, and M. Benincaso, "On making physical the control of audio plugins: The case of the retrologue hardware synthesizer," in *Proc. Audio Mostly Conf.*, 2020, pp. 146–151.

[35] 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.

[36] C. Rottondi, C. Chafe, C. Allocchio, and A. Sarti, "An overview on networked music performance technologies," *IEEE Access*, vol. 4, pp. 8823–8843, 2016.

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

[38] 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.

[39] M. Lefford, G. Bromham, G. Fazekas, and D. Moffat, "Context aware intelligent mixing systems," *J. Audio Eng. Soc.*, vol. 69, no. 3, pp. 128–141, 2021.

[40] L. Gabrielli and S. Squartini, *Wireless Networked Music Performance*. Heidelberg, Germany: Springer, 2016.

[41] A. Jensenius and M. Lyons, *A NIME Reader: Fifteen Years of New Interfaces for Musical Expression*. Heidelberg, Germany: Springer, 2017.

[42] J. Burgoyne, I. Fujinaga, and J. Downie, "Music information retrieval," in *A New Companion to Digital Humanities*. Amsterdam, The Netherlands: Elsevier, 2016, pp. 213–228.

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

[44] L. Turchet, R. Hamilton, and A. Çamci, "Music in extended realities," *IEEE Access*, vol. 9, pp. 15810–15832, 2021.

[45] O. Hödl, G. Fitzpatrick, and F. Kayali, "Design implications for technology-mediated audience participation in live music," in *Proc. Sound Music Comput. Conf.*, 2017, pp. 28–34.

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

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

[48] M. Sandler, D. De Roure, S. Benford, and K. Page, "Semantic Web technology for new experiences throughout the music production-consumption chain," in *Proc. IEEE Int. Workshop Multilayer Music Rep. Process.*, 2019, pp. 49–55.

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

[50] L. Turchet and C. Fischione, "Elk audio OS: An open source operating system for the Internet of Musical Things," *ACM Trans. Internet Things*, vol. 2, no. 2, pp. 1–18, 2021.

[51] "Digital Stage." Accessed: Feb. 15, 2023. [Online]. Available: https://digital-stage.org/?lang=en.

[52] "Jamulus." Accessed: Feb. 15, 2023. [Online]. Available: https://jamulus.io/it/.

[53] C. Drioli, C. Allocchio, and N. Buso, "Networked performances and natural interaction via LOLA: Low latency high quality a/V streaming system," in *Proc. Int. Conf. Inf. Technol. Performing Arts Media Access Entertainment*, 2013, pp. 240–250.

[54] "JamKazam." Accessed: Feb. 15, 2023. [Online]. Available: https://jamkazam.com/

[55] A. Carôt and C. Werner, "Distributed network music workshop with soundjack," in *Proc. 25th Tonmeistertagung*, Leipzig, Germany, 2008, pp. 1–9.

[56] A. Olmos et al., "Exploring the role of latency and orchestra placement on the networked performance of a distributed opera," in *Proc. 12th Annu. Int. Workshop Presence*, 2009, pp. 1–9.

[57] L. Comanducci et al., "Investigating networked music performances in pedagogical scenarios for the intermusic project," in *Proc. IEEE 23rd Conf. Open Innov. Assoc. (FRUCT)*, 2018, pp. 119–127.

[58] M. Bosi, A. Servetti, C. Chafe, and C. Rottondi, "Unlocking remote music performance during the Lockdowns: A networked concert with Jacktrip," *J. Audio Eng. Soc.*, to be published.

[59] C. Chafe, J. Caceres, and M. Gurevich, "Effect of temporal separation on synchronization in rhythmic performance," *Perception*, vol. 39, no. 7, pp. 982–992, 2010.

[60] S. Farner, A. Solvang, A. Sæbo, and U. Svensson, "Ensemble hand-clapping experiments under the influence of delay and various acoustic environments," *J. Audio Eng. Soc.*, vol. 57, no. 12, pp. 1028–1041, 2009.

[61] P. Driessen, T. Darcie, and B. Pillay, "The effects of network delay on tempo in musical performance," *Comput. Music J.*, vol. 35, no. 1, pp. 76–89, 2011.

[62] C. Bartlette, D. Headlam, M. Bocko, and G. Velikic, "Effect of network latency on interactive musical performance," *Music Perception*, vol. 24, no. 1, pp. 49–62, 2006.

[63] A. Sawchuk, E. Chew, R. Zimmermann, C. Papadopoulos, and C. Kyriakakis, "From remote media immersion to distributed immersive performance," in *Proc. ACM SIGMM Workshop Exp. Telepresence*, 2003, pp. 110–120.

[64] C. Rottondi, M. Buccoli, M. Zanoni, D. Garao, G. Verticale, and A. Sarti, "Feature-based analysis of the effects of packet delay on networked musical interactions," *J. Audio Eng. Soc.*, vol. 63, no. 11, pp. 864–875, 2015.

[65] R. Battello et al., "An adaptive metronome technique for mitigating the impact of latency in networked music performances," in *Proc. 27th Conf. Open Innov. Assoc. (FRUCT)*, 2020, pp. 10–17.

[66] R. Hupke, L. Beyer, M. Nophut, S. Preihs, and J. Peissig, "Effect of a global metronome on ensemble accuracy in networked music performance," in *Proc. Audio Eng. Soc. Conv.*, 2019, p. 147.

[67] J. H. Sørensen, P. Popovski, and J. Østergaard, "Delay minimization in real-time communications with joint buffering and coding," *IEEE Commun. Lett.*, vol. 21, no. 1, pp. 52–55, Jan. 2017.

[68] C. Werner and R. Kraneis, "UNISON: A novel system for ultra-low latency audio streaming over the Internet," in *Proc. IEEE 18th Annu. Consum. Commun. Netw. Conf. (CCNC)*, 2021, pp. 1–4.

[69] P. Ferguson, C. Chafe, and S. Gapp, "Trans-Europe express audio: Testing 1000 mile low-latency uncompressed audio between Edinburgh and Berlin using GPS-derived word clock, first with Jacktrip then with dante," in *Proc. Audio Eng. Soc. Conv.*, 2020, p. 148.

[70] J. Valin, K. Vos, and T. Terriberry, *Definition of the OPUS Audio Codec*, IETF, Fremont, CA, USA, Sep. 2012.

[71] F. Pflug and T. Fingscheid, "Robust ultra-low latency soft-decision decoding of linear PCM audio," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 11, pp. 2324–2336, Nov. 2013.

[72] R. Arean, J. Kovacevic, and V. Goyal, "Multiple description perceptual audio coding with correlating transform," *IEEE Trans. Audio, Speech, Language Process.*, vol. 10, no. 2, pp. 140–145, Mar. 2000.

[73] G. Schuller, J. Kovacevic, F. Masson, and V. Goyal, "Robust low-delay audio coding using multiple descriptions," *IEEE Trans. Audio, Speech, Language Process.*, vol. 13, no. 5, pp. 1014–1024, Sep. 2005.

[74] J. Østergaard, O. Niamut, J. Jensen, and R. Heusdens, "Perceptual audio coding using n-channel lattice vector quantization," in *Proc. IEEE Int. Conf. Audio Speech Signal Process. (ICASSP)*, May 2006, pp. 197–200.

[75] J. Østergaard, D. E. Quevedo, and J. Jensen, "Real-time perceptual moving-horizon multiple-description audio coding," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4286–4299, Sep. 2011.

[76] J. Leegaard, J. Østergaard, J. Jensen, and R. Zamir, "Practical design of delta–sigma multiple description audio coding," *EURASIP J. Audio Speech Music Process.*, vol. 2014, p. 16, Apr. 2014.

[77] J. Østergaard, "Low delay robust audio coding by noise shaping, fractional sampling, and source prediction," in *Proc. Data Compression Conf. (DCC)*, 2021, pp. 273–282.

[78] D. Keller, C. Gomes, and L. Aliel, "The handy metaphor: Bimanual, touchless interaction for the Internet of Musical Things," *J. New Music Res.*, vol. 48, no. 4, pp. 385–396, 2019.

[79] L. Turchet, M. Benincaso, and C. Fischione, "Examples of use cases with smart instruments," in *Proc. Audio Mostly Conf.*, 2017, pp. 1–47.

[80] A. Fraietta, O. Bown, S. Ferguson, S. Gillespie, and L. Bray, "Rapid composition for networked devices: HappyBrackets," *Comput. Music J.*, vol. 43, no. 2, pp. 89–108, 2019.

[81] B. Matuszewski, "A Web-based framework for distributed music system research and creation," *J. Audio Eng. Soc.*, vol. 68, no. 10, pp. 717–726, 2020.

[82] M. Centenaro, P. Casari, and L. Turchet, "Towards a 5G communication architecture for the Internet of Musical Things," in *Proc. IEEE Conf. Open Innov. Assoc. (FRUCT)*, 2020, pp. 38–45.

[83] L. Turchet, "Smart mandolin: Autobiographical design, implementation, use cases, and lessons learned," in *Proc. Audio Mostly Conf.*, 2018, pp. 1–13.

[84] L. Turchet and M. Barthet, "An ubiquitous smart guitar system for collaborative musical practice," *J. New Music Res.*, vol. 48, no. 4, pp. 352–365, 2019.

[85] L. Turchet, J. Pauwels, C. Fischione, and G. Fazekas, "Cloud-smart musical instrument interactions: Querying a large music collection with a smart guitar," *ACM Trans. Internet Things*, vol. 1, no. 3, pp. 1–29, 2020.

[86] L. Men and N. Bryan-Kinns, "LeMo: Supporting collaborative music making in virtual reality," in *Proc. IEEE VR Workshop Sonic Interact. Virtual Environ.*, 2018, pp. 1–6.

[87] B. Loveridge, "Networked music performance in virtual reality: Current perspectives," *J. Netw. Music Arts*, vol. 2, no. 1, p. 2, 2020.

[88] L. Turchet and M. Barthet, "Co-design of musical haptic Wearables for electronic music performer's communication," *IEEE Trans. Human–Mach. Syst.*, vol. 49, no. 2, pp. 183–193, Apr. 2019.

[89] L. Turchet, D. Baker, and T. Stockman, "Musical haptic Wearables for synchronisation of visually-impaired performers: A co-design approach," in *Proc. ACM Int. Conf. Interactive Media Exp.*, 2021, pp. 20–27.

[90] O. Bown, S. Ferguson, A. D. P. Dos Santos, and K. Mikolajczyk, "Supporting creative practice in wireless distributed sound installations given technical constraints," *J. Audio Eng. Soc.*, vol. 69, no. 10, pp. 757–767, 2021.

[91] O. Bown, S. Ferguson, L. Bray, A. Fraietta, and L. Loke, "Facilitating creative exploratory search with multiple networked audio devices using HappyBrackets," in *Proc. Int. Conf. New Interfaces Musical Exp.*, 2019, pp. 286–291.

[92] A. Fraietta, O. Bown, and S. Ferguson, "Transparent communication within multiplicities," in *Proc. IEEE 27th Conf. Open Innov. Assoc. (FRUCT)*, 2020, pp. 61–72.

[93] R. Vieira and F. Schiavoni, "SunFlower: An environment for standardized communication of IoMusT," in *Proc. Audio Mostly*, 2021, pp. 175–181.

[94] L. Turchet, F. Viola, G. Fazekas, and M. Barthet, "Towards a semantic architecture for Internet of Musical Things applications," in *Proc. IEEE Conf. Open Innov. Assoc. (FRUCT)*, 2018, pp. 382–390.

[95] L. Roffia, P. Azzoni, C. Aguzzi, F. Viola, F. Antoniazzi, and T. S. Cinotti, "Dynamic linked data: A SPARQL event processing architecture," *Future Internet*, vol. 10, no. 4, p. 36, 2018.

[96] F. Viola, L. Turchet, 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.

[97] L. Turchet and F. Antoniazzi, "Semantic Web of Musical Things: Achieving interoperability in the Internet of Musical Things," *J. Web Semantics*, vol. 75, 2023, Art. no. 100758.

[98] L. Turchet, F. Antoniazzi, F. Viola, F. Giunchiglia, and G. Fazekas, "The Internet of Musical Things ontology," *J. Web Semantics*, vol. 60, Jan. 2020, Art. no. 100548.

[99] X. Serra et al., *Roadmap for Music Information ReSearch*, G. Peeters, Ed., 2013.

[100] S. Pauws, "Musical key extraction from audio," in *Proc. 5th Int. Soc. Music Inf. Retrieval Conf.*, 2004, pp. 1–4.

[101] T. Fujishima, "Realtime chord recognition of musical sound: A system using common LISP music," in *Proc. Int. Comput. Music Conf.*, 1999, pp. 464–467.

[102] M. E. Davies and M. D. Plumbley, "Casual tempo tracking of audio," in *Proc. 5th Int. Soc. Music Inf. Conf.*, Barcelona, Spain, Oct. 2004, pp. 1–2.

[103] J. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Trans. Audio, Speech, Language Process.*, vol. 13, no. 5, pp. 1035–1047, Sep. 2005.

[104] L. Turchet, "Hard real time onset detection for percussive sounds," in *Proc. Digit. Audio Effects Conf.*, 2018, pp. 349–356.

[105] A. Klapuri, "Signal processing methods for the automatic transcription of music," Ph.D. dissertation, Dept. Comput. Sci., Tampere Univ. Technol., Tampere, Finland, 2004.

[106] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Audio, Speech, Language Process.*, vol. 10, no. 5, pp. 293–302, Jul. 2002.

[107] K. Choi, G. Fazekas, B. McFee, K. Cho, and M. B. Sandler, "Towards music captioning: Generating music Playlist descriptions," in *Proc. Int. Soc. Music Inf.*, 2016, pp. 1–7.

[108] I. Manco, E. Benetos, E. Quinton, and G. Fazekas, "MusCaps: Generating captions for music audio," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2021, pp. 1–8.

[109] M. Barthet, G. Fazekas, and M. Sandler, *Music Emotion Recognition: From Content- to Context-Based Models*, (LNCS 7900). Heidelberg, Germany: Springer-Verlag, 2013.

[110] J. S. Gómez-Cañón et al., "Music emotion recognition: Toward new, robust standards in personalized and context-sensitive applications," *IEEE Signal Process. Mag.*, vol. 38, no. 6, pp. 106–114, Nov. 2021.

[111] L. Turchet and J. Pauwels, "Music emotion recognition: Intention of composers-performers versus perception of musicians, non-musicians, and listening machines," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 305–316, 2021.

[112] B. Azvine, D. Djian, K. Tsui, and W. Wobcke, "The intelligent assistant: An overview," *Intelligent Systems and Soft Computing* (LNCS 1804). Berlin, Germany: Springer, 2000.

[113] A. Sheh and D. P. W. Ellis, "Chord segmentation and recognition using EM-trained hidden Markov models," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2003, pp. 1–9.

[114] F. Korzeniowski and G. Widmer, "End-to-end musical key estimation using a convolutional neural network," in *Proc. 25th Eur. Signal Process. Conf.*, 2017, pp. 966–970.

[115] K. O'Hanlon and M. B. Sandler, "FifthNet: Structured compact neural networks for automatic chord recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 2671–2682, 2021.

[116] K. Choi, G. Fazekas, K. Cho, and M. Sandler. "A Tutorial on Deep Learning for Music Information Retrieval." 2017. [Online]. Available: https://arxiv.org/abs/1709.04396

[117] J. Pauwels, K. O'Hanlon, E. Gómez, and M. B. Sandler, "20 years of automatic chord recognition from audio," in *Proc. 20th Int. Soc. Music Inf. Retrieval Conf.*, 2019, pp. 54–63.

[118] S. Böck and G. Widmer, "Maximum filter vibrato suppression for onset detection," in *Proc. 16th Int. Conf. Digit. Audio Effects*, 2013, p. 148.

[119] M. Tian, G. Fazekas, D. A. A. Black, and M. Sandler, "Design and evaluation of onset detectors using different fusion policies," in *Proc. 15th Int. Soc. Music Inf. Retrieval Conf.*, 2014, pp. 631–636.

[120] J. Schlüter and S. Böck, "Improved musical onset detection with convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2014, pp. 6979–6983.

[121] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, "Automatic music transcription: An overview," *IEEE Signal Process. Mag.*, vol. 36, no. 1, pp. 20–30, Jan. 2019.

[122] C. Hawthorne et al., "Onsets and frames: Dual-objective piano transcription," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2018, pp. 50–57.

[123] C. Hawthorne, I. Simon, R. Swavely, E. Manilow, and J. Engel, "Sequence-to-sequence piano transcription with transformers," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2021, pp. 246–253.

[124] P. Saari, G. Fazekas, T. Eerola, M. Barthet, O. Lartillot, and M. Sandler, "Genre-adaptive semantic computing enhances audio-based music mood prediction," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 122–135, Apr./Jun. 2016.

[125] P. Saari, M. Barthet, G. Fazekas, T. Eerola, and M. Sandler, "Semantic models of musical mood: Comparison between crowd-sourced and curated editorial tags," in *Proc. IEEE Int. Conf. Multimedia Expo Workshop*, 2013, pp. 1–6.

[126] C. Baume, G. Fazekas, M. Barthet, D. Martson, and M. Sandler, "Selection of audio features for music emotion recognition using production music," in *Proc. AES 53rd Int. Conf. Semantic Audio*, Jan. 2014, pp. 1–4.

[127] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Transfer learning for music classification and regression tasks," in *Proc. 18th Int. Soc. Music Inf. Retrieval Conf.*, 2017, pp. 141–149.

[128] K. Choi, G. Fazekas, K. Cho, and M. Sandler, "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.

[129] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2392–2396.

[130] J. Pons, O. Nieto, M. Prockup, E. M. Schmidt, A. F. Ehmann, and X. Serra, "End-to-end learning for music audio tagging at scale," in *Proc. 19th Int. Soc. Music Inf. Retrieval Conf.*, Paris, France, Sep. 2018, pp. 637–644.

[131] M. Barthet, A. Anglade, G. Fazekas, S. Kolozali, and R. Macrae, "Music recommendation for music learning: Hotttabs a multimedia guitar tutor," in *Proc. 2nd Workshop Music Recommendation Disc. ACM Recommender Syst. Conf.*, 2011, pp. 1–9.

[132] J. Pauwels, A. Xambo, G. Roma, M. Barthet, and G. Fazekas, "Exploring real-time Visualisations to support chord learning with a large music collection," in *Proc. Web Audio Conf. (WAC)*, Sep. 2018, pp. 1–7.

[133] 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)*, Blacksburg, VA, USA, Jun. 2018, pp. 364–369.

[134] J. Picaut, A. Can, N. Fortin, J. Ardouin, and M. Lagrange, "Low-cost sensors for urban noise monitoring networks—A literature review," *Sensors*, vol. 20, no. 8, p. 2256, 2020.

[135] 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.

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

[137] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and evaluation of sound event localization and detection in DCASE 2019," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 684–698, 2020.

[138] D. Steele, J. Krijnders, and C. Guastavino, *The Sensor City Initiative: Cognitive Sensors for Soundscape Transformations*, GIS, Ostrava, Czech Republic, 2013.

[139] 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.

[140] L. Lavia et al., "Sounding brighton: Practical approaches towards better soundscapes," in *Proc. INTERNOISE NOISE CON Congr. Conf. Process.*, 2012, pp. 436–444.

[141] S. Kephalopoulos, M. Paviotti, and F. Anfosso-Lédée, *Common Noise Assessment Methods in Europe (CNOSSOS-EU)*, Publ. Office Eur. Union, Luxembourg City, Luxembourg, 2012.

[142] C. Mydlarz, M. Sharma, Y. Lockerman, B. Steers, C. Silva, and J. P. Bello, "The life of a New York City noise sensor network," *Sensors*, vol. 19, no. 6, p. 1415, 2019.

[143] P. Bellucci, L. Peruzzi, and G. Zambon, "LIFE DYNAMAP project: The case study of Rome," *Appl. Acoust.*, vol. 117, pp. 193–206, Feb. 2017.

[144] D. Botteldooren, L. Dekoninck, C. Meeussen, and T. Van Renterghem, "Early stage sound planning in urban re-development: The antwerp case study," in *Proc. Int. Congr. Expo. Noise Control Eng. (Inter-Noise)*, 2018, pp. 1–2.

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

[146] A. Lesieur, P. Aumond, V. Mallet, and A. Can, "Meta-modeling for urban noise mapping," *J. Acoust. Soc. America*, vol. 148, no. 6, pp. 3671–3681, 2020.

[147] F. Gontier, C. Lavandier, P. Aumond, M. Lagrange, and J.-F. Petiot, "Estimation of the perceived time of presence of sources in urban acoustic environments using deep learning techniques," *Acta Acustica United Acustica*, vol. 105, no. 6, pp. 1053–1066, 2019.

[148] M. Cartwright, J. Cramer, J. Salamon, and J. P. Bello, "TriCycle: Audio representation learning from sensor network data using self-supervision," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, 2019, pp. 278–282.

[149] F. Gontier, V. Lostanlen, M. Lagrange, N. Fortin, C. Lavandier, and J.-F. Petiot, "Polyphonic training set synthesis improves self-supervised urban sound classification," *J. Acoust. Soc. America*, vol. 149, no. 6, pp. 4309–4326, 2021.

[150] B. M. Van Doren, K. G. Horton, A. M. Dokter, H. Klinck, S. B. Elbin, and A. Farnsworth, "High-intensity urban light installation dramatically alters nocturnal bird migration," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 42, pp. 11175–11180, 2017.

[151] V. Lostanlen, A. Bernabeu, J.-L. Béchennec, M. Briday, S. Faucou, and M. Lagrange, "Energy efficiency is not enough: Towards a Batteryless Internet of sounds," in *Proc. Int. Workshop Internet Sounds*, 2021, pp. 147–155.

[152] D. Stowell and M. D. Plumbley, "Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning," *PeerJ*, vol. 2, p. e488, Jul. 2014.

[153] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "BirdNET: A deep learning solution for avian diversity monitoring," *Ecol. Inf.*, vol. 61, Mar. 2021, Art. no. 101236.

[154] E. Rovithis, N. Moustakas, K. Vogklis, K. Drossos, and A. Floros, "Design recommendations for a collaborative game of bird call recognition based on Internet of Sound practices," *J. Audio Eng. Soc.*, vol. 69, no. 12, pp. 956–966, 2021.

[155] G. E. Davis et al., "Long-term passive acoustic recordings track the changing distribution of north Atlantic right whales (Eubalaena Glacialis) from 2004 to 2014," *Sci. Rep.*, vol. 7, no. 1, pp. 1–12, 2017.

[156] M. F. Baumgartner et al., "Persistent near real-time passive acoustic monitoring for baleen whales from a moored buoy: System description and evaluation," *Methods Ecol. Evol.*, vol. 10, no. 9, pp. 1476–1489, 2019.

[157] M. Iber et al., "Mind the steps: Towards auditory feedback in tele-rehabilitation based on automated gait classification," in *Proc. Audio Mostly*, 2021, pp. 139–146.

[158] J. Vandendriessche, N. Wouters, B. da Silva, M. Lamrini, M. Y. Chkouri, and A. Touhafi, "Environmental sound recognition on embedded systems: From FPGAs to TPUs," *Electronics*, vol. 10, no. 21, p. 2622, 2021.

[159] 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.

[160] A.-M. Solomes and D. Stowell, "Efficient bird sound detection on the Bela embedded system," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 746–750.

[161] G. Kiarie and C. W. Maina, "Raspberry Pi based recording system for acoustic monitoring of bird species," in *Proc. IEEE IST Africa Conf.*, 2021, pp. 1–8.

[162] M. Kukushkin and S. Ntalampiras, "Automatic acoustic classification of feline sex," in *Proc. Audio Mostly*, 2021, pp. 156–160.

[163] 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.

[164] I. Zualkernan, J. Judas, T. Mahbub, A. Bhagwagar, and P. Chand, "An AIoT system for bat species classification," in *Proc. IEEE Int. Conf. Internet Things Intell. Syst. (IoTaIS)*, 2021, pp. 155–160.

[165] T. Michailidis, G. Meadow, C. Barlow, and E. Rajabally, "Implementing remote audio as a diagnostics tool for maritime autonomous surface ships," in *Proc. IEEE 27th Conf. Open Innov. Assoc. (FRUCT)*, 2020, pp. 157–163.

[166] 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.

[167] J. Hsu. "Starkey's AI Transforms Hearing Aids Into Smart Wearables." Accessed: Jul. 7, 2021. [Online]. Available: https://spectrum.ieee.org/the-human-os/biomedical/devices/starkeys-ai-transforms-hearing-aid-into-smart-wearables

[168] D. Michelsanti et al., "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 29, pp. 1368–1396, 2021.

[169] C. B. Christensen, R. K. Hietkamp, J. M. Harte, T. Lunner, and P. Kidmose, "Toward EEG-assisted hearing aids: Objective threshold estimation based on ear-EEG in subjects with sensorineural hearing loss," *Trends Hearing*, vol. 22, pp. 1–13, Dec. 2018.

[170] T. Quatieri et al., "Exploiting Nonacoustic sensors for speech encoding," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 2, pp. 533–544, Mar. 2006.

[171] X. Shen, W.-S. Gan, and D. Shi, "Multi-channel wireless hybrid active noise control with fixed-adaptive control selection," *J. Sound Vib.*, vol. 541, Dec. 2022, Art. no. 117300.

[172] E. Baccelli et al., "RIOT: An open source operating system for low-end embedded devices in the IoT," *IEEE Internet Things J.*, vol. 5, no. 6, pp. 4428–4440, Dec. 2018.

[173] 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.

[174] I. Franco and M. Wanderley, "Prynth: A framework for self-contained digital music instruments," in *Proc. Int. Symp. Comput. Music Multidiscipl. Res.*, 2016, pp. 357–370.

[175] E. Meneses, J. Wang, S. Freire, and M. M. Wanderley, "A comparison of open-source Linux frameworks for an augmented musical instrument implementation," in *Proc. Conf. New Interfaces Musical Exp.*, 2019, pp. 222–227.

[176] A. McPherson and V. Zappi, "An environment for Submillisecond-latency audio and sensor processing on BeagleBone black," in *Proc. Audio Eng. Soc. Conv.*, 2015, p. 138. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=17755

[177] J. Brown and B. Martin, "How fast is fast enough? Choosing between Xenomai and Linux for real-time applications," in *Proc. 12th Real Time Linux Workshop*, 2010, pp. 1–17.

[178] J. T. Pollock and R. Hodgson, *Adaptive Information: Improving Business Through Semantic Interoperability, Grid Computing, and Enterprise Integration (Wiley Series in Systems Engineering and Management)*. Hoboken, NJ, USA: Wiley Intersci., 2004.

[179] D. Moffat and M. B. Sandler, "Approaches in intelligent music production," *Arts*, vol. 8, no. 4, p. 125, 2019. [Online]. Available: https://www.mdpi.com/2076-0752/8/4/125

[180] R. Stables, B. De Man, S. Enderby, J. Reiss, G. Fazekas, and T. Wilmering, "Semantic description of timbral transformations in music production," in *Proc. ACM Multimedia*, Oct. 2016, pp. 337–341.

[181] S. Singh, G. Bromham, D. Sheng, and G. Fazekas, "Intelligent control method for the dynamic range compressor: A user study," *J. Audio Eng. Soc.*, vol. 69, nos. 7–8, pp. 576–585, 2021.

[182] M. B. Cartwright and B. Pardo, "Social-EQ: Crowdsourcing an equalization descriptor map," in *Proc. 14th Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, Nov. 2013, pp. 395–400.

[183] T. Wilmering, G. Fazekas, and M. Sandler, "High level semantic metadata for the control of multitrack adaptive audio effects," in *Proc. 133rd Conv. Audio Eng. Soc.*, San Francisco, CA, USA, 2012, pp. 1–8.

[184] G. Fazekas and M. Sandler, "Intelligent editing of studio recordings with the help of automatic music structure extraction," in *Proc. 122nd Conv. Audio Eng. Soc.*, Vienna, Austria, 2007, pp. 1–8.

[185] D. Stowell, M. 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.

[186] D. Doukhan, J. Carrive, F. Vallet, A. Larcher, and S. Meignier, "An open-source speaker gender detection framework for monitoring gender equality," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2018, pp. 5214–5218.

[187] A. Draghici, J. Abeßer, and H. Lukashevich, "A study on spoken language identification using deep neural networks," in *Proc. 15th Int. Conf. Audio Mostly*, 2020, pp. 253—-256.

[188] Y. Raimond, S. Abdallah, M. Sandler, and F. Giasson, "The music ontology," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2007, pp. 1–8.

[189] Y. Raimond, F. Giasson, K. Jacobson, G. Fazekas, T. Gangler, and S. Reinhardt. "The Music Ontology Specification." 2010. [Online]. Available: http://musicontology.com/

[190] G. Fazekas and M. Sandler, "Knowledge representation issues in audio-related metadata model design," in *Proc. 133rd Conv. Audio Eng. Soc.*, San Francisco, CA, USA, 2012, p. 8765.

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

[192] T. Wilmering, G. Fazekas, and M. Sandler, "The audio effects ontology," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2013, pp. 215–220.

[193] T. Wilmering, G. Fazekas, and M. Sandler, "AUFX-O: Novel methods for the representation of audio processing workflows," in *Proc. 15th Int. Semantic Web Conf.*, vol. 9982, 2016, pp. 229–237.

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

[195] S. M. Rashid, D. De Roure, and D. L. McGuinness, "A music theory ontology," in *Proc. 1st Int. Workshop Semantic Appl. Audio Music*, 2018, pp. 6–14.

[196] P. Proutskova, A. Volk, P. Heidarian, and G. Fazekas, "From music ontology towards ethno-music-ontology," in *Proc. 21st Int. Soc. Music Inf. Retrieval Conf.*, 2020, pp. 923–931.

[197] K. Page, S. Bechhofer, G. Fazekas, D. Weigl, and T. Wilmering, "Realising a layered digital library: Exploration and analysis of the live music archive through linked data," in *Proc. ACM/IEEE-CS Joint Conf. Digit. Libraries (JCDL)*, Toronto, ON, Canada, 2017, pp. 89–98.

[198] T. Wilmering, F. Thalmann, G. Fazekas, and M. B. Sandler, "Bridging fan communities and facilitating access to music archives through semantic audio applications," in *Proc. Audio Eng. Soc. Conv.*, 2017, p. 143.

[199] S. Kolozali, G. Fazekas, M. Barthet, and M. Sandler, "Knowledge representation issues in musical instrument ontology design," in *Proc. 12th Int. Soc. Music Inf. Retrieval (ISMIR) Conf.*, 2011, pp. 465–470.

[200] S. Kolozali, M. Barthet, G. Fazekas, and M. Sandler, "Automatic ontology generation for musical instruments based on audio analysis," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 10, pp. 2207–2220, Oct. 2013.

[201] 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.*, 2016, pp. 47–54.

[202] L. Turchet, P. Bouquet, A. Molinari, and G. Fazekas, "The smart musical instruments ontology," *J. Web Semantics*, vol. 72, Apr. 2021, Art. no. 100687.

[203] L. Turchet and D. Golishev, "SMIF: A format for the offline exchange of smart musical instruments configuration and data," *J. Audio Eng. Soc.*, vol. 69, no. 12, pp. 946–955, 2021.

[204] L. Turchet and P. Bouquet, "Smart musical instruments preset sharing: An ontology-based data access approach," in *Proc. IEEE World Forum Internet Things*, 2021, pp. 1–6.

[205] A. Allik, G. Fazekas, M. Barthet, and M. Sandler, "myMoodplay: An interactive mood-based music discovery app," in *Proc. 2nd Web Audio Conf. (WAC)*, Apr. 2016, pp. 1–5. [Online]. Available: http://hdl.handle.net/1853/54589

[206] F. Thalmann, A. P. Carillo, G. Fazekas, and M. Sandler, "The semantic music player: A smart mobile player based on ontological structures and analytical feature Metadata," in *Proc. Web Audio Conf. (WAC)*, Apr. 2016, pp. 1–7.

[207] 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.*, 2017, pp. 776–780.

[208] 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.

[209] "EBU core metadata set specification v1.10," Eur. Broadcast. Union, Geneva, Switzerland, Rep. 3293, 2020.

[210] V. Rodriguez-Doncel, J. Delgado, S. Llorente, E. Rodriguez, and L. Boch, "Overview of the MPEG-21 media contract ontology," *Semantic Web*, vol. 7, no. 3, pp. 311–332, 2016.

[211] T. Nakatani and H. G. Okuno, "Sound ontology for computational auditory scene analysis," in *Proc. AAAI/IAAI*, 1998, pp. 1004–1010.

[212] A. Lobanova, J. Spenader, and B. Valkenier, "Lexical and perceptual grounding of a sound ontology," in *Text, Speech Dialogue*. Berlin, Germany: Springer, 2007, pp. 180–187.

[213] A. Jiménez, B. Elizalde, and B. Raj, "Sound event classification using ontology-based neural networks," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2018, pp. 1–9.

[214] Y. Sun and S. Ghaffarzadegan, "An ontology-aware framework for audio event classification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2020, pp. 321–325.

[215] H. F. Garcia, A. Aguilar, E. Manilow, and B. Pardo, "Leveraging hierarchical structures for few-shot musical instrument recognition," in *Proc. 22nd Int. Soc. Music Inf. Retrieval Conf.*, 2021, pp. 1–7.

[216] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *Proc. ACM Int. Conf. Multimedia*, 2013, pp. 411–412.

[217] F. Font et al., "Audio commons: Bringing creative commons audio content to the creative industries," in *Proc. Audio Eng. Soc. Conf. 61st Int. Conf. Audio Games*, 2016, pp. 1–8.

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

[219] C. Roberts and J. Kuchera-Morin, "Gibber: Live coding audio in the browser," in *Proc. Int. Comput. Music Conf.*, 2012, pp. 1–9.

[220] 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.

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

[222] Q. Lan and A. Refsum, "Glicol: A graph-oriented live coding language developed with rust, WebAssembly and AudioWorklet," in *Proc. 6th Web Audio Conf. (WAC)*, 2021, pp. 1–6.

[223] A. A. Correya, J. Marcos Fernández, L. Joglar-Ongay, P. Alonso Jiménez, X. Serra, and D. Bogdanov, "Audio and music analysis on the Web using Essentia.js," *Trans. Int. Soc. Music Inf. Retrieval*, vol. 4, no. 1, pp. 167–181, 2021.

[224] P. Bahadoran, A. Benito, T. Vassallo, and J. D. Reiss, "FXive: A Web platform for procedural sound synthesis," in *Proc. Audio Eng. Soc. Conv.*, 2018, p. 144.

[225] B. Matuszewski and F. Bevilacqua, "Toward a Web of Audio Things," in *Proc. Sound Music Comput. Conf.*, 2018, pp. 1–7.

[226] J. Lambert, S. Robaszkiewicz, and N. Schnell, "Synchronisation for distributed audio rendering over heterogeneous devices, in HTML5," in *Proc. Web Audio Conf.*, 2016, pp. 1–8.

[227] A. Baratè, G. Haus, and L. A. Ludovico, "Advanced experience of music through 5G technologies," in *Proc. IOP Conf. Mater. Sci. Eng.*, vol. 364, Jun. 2018, Art. no. 12021.

[228] A. Baratè, G. Haus, L. A. Ludovico, E. Pagani, and N. Scarabottolo, "5G technology and its applications to music education," in *Proc. Multi Conf. Comput. Sci. Inf. Syst. Int. Conf. e-Learn. (MCCSIS)*, 2019, pp. 65–72.

[229] C. Rinaldi, F. Franchi, A. Marotta, F. Graziosi, and C. Centofanti, "On the exploitation of 5G multi-access edge computing for spatial audio in cultural heritage applications," *IEEE Access*, vol. 9, pp. 155197–155206, 2021.

[230] X. Jiang et al., "Low-latency networking: Where latency lurks and how to tame it," *Proc. IEEE*, vol. 107, no. 2, pp. 280–306, May 2018.

[231] N. Finn, "Introduction to time-sensitive networking," *IEEE Commun. Stand. Mag.*, vol. 2, no. 2, pp. 22–28, Jun. 2018.

[232] H. Shokri-Ghadikolaei, C. Fischione, G. Fodor, P. Popovski, and M. Zorzi, "Millimeter wave cellular networks: A MAC layer perspective," *IEEE Trans. Commun.*, vol. 63, no. 10, pp. 3437–3458, Oct. 2015.

[233] 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, Oct. 2016.

[234] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network slicing in 5G: Survey and challenges," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 94–100, May 2017.

[235] A. Carôt, M. Dohler, S. Saunders, F. Sardis, R. Cornock, and N. Uniyal, "The world's first interactive 5G music concert: Professional quality networked music over a commodity network infrastructure," in *Proc. 17th Sound Music Comput. Conf.*, 2020, pp. 407–412.

[236] J. Dürre et al., "In-depth latency and reliability analysis of a networked music performance over public 5G infrastructure," in *Proc. Audio Eng. Soc. Conv.*, 2022, Art. no. 10621.

[237] Y. Wu, Q. Chaudhari, and E. Serpedin, "Clock synchronization of wireless sensor networks," *IEEE Signal Process. Mag.*, vol. 28, no. 1, pp. 124–138, Jan. 2011.

[238] A. Hildebrand, *AES Standard for Audio Applications of Networks-High-Performance Streaming Audio-Over-IP Interoperability*, AES Standard AES67-2013, 2018.

[239] 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. Audio Eng. Soc. Conf. 44th Int. Conf. Audio Netw.*, 2011, pp. 1–9.

[240] 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.

[241] 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.

[242] D. Trueman, P. Cook, S. Smallwood, and G. Wang, "PLOrk: The Princeton laptop orchestra, year 1," in *Proc. Int. Comput. Music Conf.*, 2006, pp. 1–6.

[243] G. Wang, N. J. Bryan, J. Oh, and R. Hamilton, "Stanford laptop orchestra (SLORK)," in *Proc. ICMC*, 2009, pp. 1–8.

[244] 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.

[245] L. Turchet and E. Rinaldo, "Technical performance assessment of the Ableton link protocol," *J. Audio Eng. Soc.*, vol. 69, no. 10, pp. 748–756, 2021.

[246] Q. Li and D. Rus, "Global clock synchronization in sensor networks," *IEEE Trans. Comput.*, vol. 55, no. 2, pp. 214–226, Feb. 2006.

[247] F. Sivrikaya and B. Yener, "Time synchronization in sensor networks: A survey," *IEEE Netw.*, vol. 18, no. 4, pp. 45–50, Jul./Aug. 2004.

[248] D. L. Mills, "Internet time synchronization: The network time protocol," *IEEE Trans. Commun.*, vol. 39, no. 10, pp. 1482–1493, Oct. 1991.

[249] M. Maróti, B. Kusy, G. Simon, and Á. Lédeczi, "The flooding time synchronization protocol," in *Proc. 2nd Int. Conf. Embedded Netw. Sensor Syst.*, 2004, pp. 39–49.

[250] C. Lenzen, P. Sommer, and R. Wattenhofer, "Optimal clock synchronization in networks," in *Proc. 7th ACM Conf. Embedded Netw. Sensor Syst.*, 2009, pp. 225–238.

[251] 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.

[252] M. Wright, "Open sound control: An enabling technology for musical networking," *Org. Sound*, vol. 10, no. 3, pp. 193–200, 2005.

[253] T. Mitchell, S. Madgwick, S. Rankine, G. Hilton, A. Freed, and A. Nix, "Making the most of Wi-Fi: Optimisations for robust wireless live music performance," in *Proc. Conf. New Interfaces Musical Exp.*, 2014, pp. 251–256.

[254] J. Wang, E. Meneses, and M. Wanderley, "The scalability of WiFi for mobile embedded sensor interfaces," in *Proc. Conf. New Interfaces Musical Exp.*, 2020, pp. 73–76.

[255] E. Brandt and R. Dannenberg, "Time in distributed real-time systems," in *Proc. Int. Comput. Music Conf.*, 1999, p. 14.

[256] S. Madgwick, T. Mitchell, C. Barreto, and A. Freed, "Simple synchronisation for open sound control," in *Proc. Int. Comput. Music Conf.*, 2015, pp. 1–8.

[257] R. Dannenberg, "O2: A network protocol for music systems," *Wireless Commun. Mobile Comput.*, vol. 2019, May 2019, Art. no. 8424381.

[258] F. Goltz, "Ableton link—A technology to synchronize music software," in *Proc. Linux Audio Conf.*, 2018, pp. 39–42.

[259] R. Oda and R. Fiebrink, "The global metronome: Absolute tempo sync for networked musical performance," in *Proc. Conf. New Interfaces Musical Exp.*, 2016, pp. 1–7.

[260] R. Hupke, J. Peissig, A. Genovese, S. Sridhar, and A. Roginska, "Impact of source panning on a global metronome in rhythmic networked music performance," in *Proc. IEEE 27th Conf. Open Innov. Assoc. (FRUCT)*, 2020, pp. 73–83.

[261] R. Battello, L. Comanducci, F. Antonacci, G. Cospito, and A. Sarti, "Experimenting with adaptive metronomes in networked music performances!" *J. Audio Eng. Soc.*, vol. 69, no. 10, pp. 737–747, 2021.

[262] G. Widmer, "Getting closer to the essence of music: The con Espressione manifesto," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 2, p. 19, 2017.

[263] P. Alonso-Jiménez, D. Bogdanov, J. Pons, and X. Serra, "Tensorflow audio models in essentia," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 266–270.

[264] R. Fiebrink and P. R. Cook, "The Wekinator: A system for real-time, interactive machine learning in music," in *Proc. 11th Int. Soc. Music Inf. Retrieval Conf. (ISMIR)(Utrecht)*, vol. 3, 2010, pp. 1–8.

[265] P. Verma, A. I. Mezza, C. Chafe, and C. Rottondi, "A deep learning approach for low-latency packet loss concealment of audio signals in networked music performance applications," in *Proc. 27th Conf. Open Innov. Assoc. (FRUCT)*, 2020, pp. 268–275.

[266] Y. Kawaguchi et al., "Description and discussion on DCASE 2021 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring under domain shifted conditions," in *Proc. DCASE Workshop*, 2021, p. 9.

[267] E. C. Nunes, "Anomalous sound detection with machine learning: A systematic review," 2021, *arXiv:2102.07820*.

[268] L. Deng and D. Yu, "Deep learning: Methods and applications," *Found. Trends Signal Process.*, vol. 7, nos. 3–4, pp. 197–387, 2014.

[269] J. Kim, J. Urbano, C. C. S. Liem, and A. Hanjalic, "One deep music representation to rule them all? A comparative analysis of different representation learning strategies," *Neural Comput. Appl.*, vol. 32, no. 4, pp. 1067–1093, 2020.

[270] J. Feldmann et al., "Parallel convolutional processing using an integrated photonic tensor core," *Nature*, vol. 589, no. 7840, pp. 52–58, 2021.

[271] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," 2015, *arXiv:1510.00149*.

[272] R. Krishnamoorthi, "Quantizing deep convolutional networks for efficient inference: A whitepaper," 2018, *arXiv:1806.08342*.

[273] R. Stables, S. Enderby, G. De Man, B.and Fazekas, and J. D. Reiss, "SAFE: A system for extraction and retrieval of semantic audio descriptors," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2014, pp. 1–4.

[274] A. Freed and A. Schmeder, "Features and future of open sound control version 1.1 for NIME," in *Proc. New Interfaces Musical Exp.*, 2009, pp. 1–8.

[275] M. A. Martínez-Prieto, M. A. Gallego, and J. D. Fernández, "Exchange and consumption of huge RDF data," in *Proc. Extended Semantic Web Conf.*, 2012, pp. 437–452.

[276] Z. Shelby, K. Hartke, and C. Bormann, "The constrained application protocol (CoAP)," in *Proc. CoAP*, 2014, pp. 1–8.

[277] J. Liu, X. Zhang, Y. Li, J. Wang, and H.-J. Kim, "Deep learning-based reasoning with multi-ontology for IoT applications," *IEEE Access*, vol. 7, pp. 124688–124701, 2019.

[278] G. Cerutti, R. Prasad, A. Brutti, and E. Farella, "Compact recurrent neural networks for acoustic event detection on low-energy low-complexity platforms," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 4, pp. 654–664, May 2020.

[279] H. Wu, Y. Shen, X. Xiao, G. T. Nguyen, A. Hecker, and F. H. Fitzek, "Accelerating industrial IoT acoustic data separation with in-network computing," *IEEE Internet Things J.*, vol. 10, no. 5, pp. 3901–3916, Mar. 2023.

[280] "Maxine—Nvidia." Accessed: Jul. 7, 2021. [Online]. Available: https://blogs.nvidia.com/blog/2020/10/05/gan-video-conferencing-maxine

[281] H. Li, K. Ota, and M. Dong, "Learning IoT in edge: Deep learning for the Internet of Things with edge computing," *IEEE Netw.*, vol. 32, no. 1, pp. 96–101, Jan./Feb. 2018.

[282] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Inf. Sci.*, vol. 250, pp. 113–141, Nov. 2013.

[283] C. Huang, X. Wu, X. Zhang, S. Lin, and N. V. Chawla, "Deep prototypical networks for imbalanced time series classification under data scarcity," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manag.*, 2019, pp. 2141–2144.

[284] J. Park, S. Samarakoon, M. Bennis, and M. Debbah, "Wireless network intelligence at the edge," *Proc. IEEE*, vol. 107, no. 11, pp. 2204–2239, Nov. 2019.

[285] H. Hellström, J. M. B. d. Silva Jr, V. Fodor, and C. Fischione, "Wireless for machine learning," 2020, *arXiv:2008.13492*.

[286] "Storage and Authentication of Audio Footage for IoAuT Devices Using Distributed Ledger Technology." Accessed: Feb. 15, 2023. [Online]. Available: https://arxiv.org/abs/2110.08821

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

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

[289] "Moving Picture Experts Group (MPEG)." Accessed: Jul. 7, 2021. [Online]. Available: https://mpeg.chiariglione.org

[290] "Picture, Audio and Data Coding by Artificial Intelligence (MPAI)." Accessed: Jul. 7, 2021. [Online]. Available: https://mpai.community

[291] "Bluetooth LE." Accessed: Jul. 7, 2021. [Online]. Available: https://www.bluetooth.com/learn-about-bluetooth/recent-enhancements/le-audio/

[292] H. Cao, V. Leung, C. Chow, and H. Chan, "Enabling technologies for wireless body area networks: A survey and outlook," *IEEE Commun. Mag.*, vol. 47, no. 12, pp. 84–93, Dec. 2009.

[293] ITU-T, *E.800—Definitions of Terms Related to Quality of Service*. Geneva, Switzerland: Int. Telecommun. Union, 2008.

[294] ITU-T, *P.1305—Effect of Delays on Telemeeting Quality*. Geneva, Switzerland: Int. Telecommun. Union, 2016.

[295] M. Lutzky, M. Schnell, M. L. Valero, and J. Hilpert, "MPEG-4 AAC-ELD v2–the new state of the art in high quality communication audio coding," in *Microelectronic Systems*. Heidelberg, Germany: Springer, 2011, pp. 341–349.

[296] M. Schnell et al., "LC3 and LC3plus: The new audio transmission standards for wireless communication," in *Proc. 150th AES Conv.*, 2021, pp. 1–8.

[297] E. A. Drott, "Music as a technology of surveillance," *J. Soc. Amer. Music*, vol. 12, no. 3, pp. 233–267, 2018.

[298] F. Morreale and M. Eriksson, "My library has just been obliterated: Producing new norms of use via software update," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2020, pp. 1–13.

[299] A. Holzapfel, B. Sturm, and M. Coeckelbergh, "Ethical dimensions of music information retrieval technology," *Trans. Int. Soc. Music Inf. Retrieval*, vol. 1, no. 1, pp. 44–55, 2018.

[300] F. Morreale, A. Bin, A. McPherson, P. Stapleton, and M. Wanderley, "A NIME of the times: Developing an outward-looking political agenda for this community," in *Proc. New Interfaces Musical Exp.*, 2020, pp. 1–8.

[301] G. Baldini, M. Botterman, R. Neisse, and M. Tallacchini, "Ethical design in the Internet of Things," *Sci. Eng. Ethics*, vol. 24, no. 3, pp. 905–925, 2018.

[302] R. Weber, "Internet of Things: Privacy issues revisited," *Comput. Law Security Rev.*, vol. 31, no. 5, pp. 618–627, 2015.

[303] A. Nautsch, C. Jasserand, E. Kindt, M. Todisco, I. Trancoso, and N. Evans, "The GDPR & speech data: Reflections of legal and technology communities, first steps towards a common understanding," in *Proc. Interspeech*, 2019, pp. 3695–3699.

[304] A. Nautsch et al., "The privacy ZEBRA: Zero evidence biometric recognition assessment," in *Proc. Interspeech*, 2020, pp. 1698–1702.

[305] N. Tomashenko et al., "Introducing the VoicePrivacy initiative," in *Proc. Interspeech*, 2020, pp. 1693–1697.

[306] B. Brüggemeier and P. Lalone, "Perceptions and reactions to conversational privacy initiated by a conversational user interface," *Comput. Speech Lang.*, vol. 71, Jan. 2022, Art. no. 101269.

[307] T. Bäckström et al., "Intuitive privacy from acoustic reach: A case for networked voice user-interfaces," in *Proc. 1st ISCA Symp. Security Privacy Speech Commun.*, 2021, pp. 1–12.

[308] European Data Protection Board. "Guidelines 02/2021 on Virtual Voice Assistants." 2021. [Online]. Available: https://edpb.europa.eu/system/files/2021-07/edpb_guidelines_202102_on_vva_v2.0_adopted_en.pdf

[309] Open Voice Network. "Privacy Guidelines and Capabilities Unique to Voice." 2021. [Online]. Available: https://openvoicenetwork.org/post/voice-specific-privacy/

[310] T. Bäckström, B. Brüggemeier, and J. Fischer, *Privacy in Speech Interfaces*, VDE News, Frankfurt, Germany, 2020.

[311] R. Roman, P. Najera, and J. Lopez, "Securing the Internet of Things," *Computer*, vol. 44, no. 9, pp. 51–58, 2011.

[312] P. P. Zarazaga, T. Bäckström, and S. Sigg, "Acoustic fingerprints for access management in ad-hoc sensor networks," *IEEE Access*, vol. 8, pp. 166083–166094, 2020.

[313] M. A. Pathak and B. Raj, "Privacy-preserving speaker verification and identification using Gaussian mixture models," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 2, pp. 397–406, Feb. 2013.

[314] M. A. Yakubu, P. K. Atrey, and N. C. Maddage, "Secure audio reverberation over cloud," in *Proc. 10th Annu. Symp. Inf. Assurance (ASIA)*, 2015, p. 39.

[315] E. Crockett. "A Low-Depth Homomorphic Circuit for Logistic Regression Model Training." 2020. [Online]. Available: https://eprint.iacr.org/2020/1483

[316] X. Meng and J. Feigenbaum. "Privacy-Preserving XGBoost Inference." 2020. [Online]. Available: https://arxiv.org/pdf/2011.04789.pdf

[317] A. A. Badawi et al., "Towards the AlexNet moment for homomorphic encryption: HCNN, the first homomorphic CNN on encrypted data with GPUs," *IEEE Trans. Emerg. Topics Comput.*, vol. 9, no. 3, pp. 1330–1343, Jul./Sep. 2021.

[318] L. Gabrielli and L. Turchet, "Towards a sustainable Internet of sounds," in *Proc. 17th Int. Audio Mostly Conf.*, 2022, pp. 231–238.

[319] T. M. Fernández-Caramés and P. Fraga-Lamas, "A review on the use of blockchain for the Internet of Things," *IEEE Access*, vol. 6, pp. 32979–33001, 2018.

[320] A. Reyna, C. Martín, J. Chen, E. Soler, and M. Díaz, "On blockchain and its integration with IoT. Challenges and opportunities," *Future Gener. Comput. Syst.*, vol. 88, pp. 173–190, Nov. 2018.
[321] H.-N. Dai, Z. Zheng, and Y. Zhang, "Blockchain for Internet of Things: A survey," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 8076–8094, Oct. 2019.
[322] M. A. Uddin, A. Stranieri, I. Gondal, and V. Balasubramanian, "A survey on the adoption of blockchain in IoT: Challenges and solutions," *Blockchain Res. Appl.*, vol. 2, no. 2, 2021, Art. no. 100006.
[323] L. Turchet and C. Ngo, "Blockchain-based Internet of musical things," *Blockchain Res. Appl.*, vol. 3, no. 3, 2022, Art. no. 100083.
[324] B. D. Mayton, G. Dublon, N. Joliat, and J. A. Paradiso, "Patchwork: Multi-user network control of a massive modular synthesizer," in *Proc. Int. Conf. New Interfaces Musical Expression*, 2012.

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

He is an Associate 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 European Space Agency, the Italian Minister of Foreign Affairs, and the Danish Research Council. He is a Co-Founder of the music-tech company Elk, Stockholm, Sweden.

Dr. Turchet serves as an Associate Editor for IEEE ACCESS and the JOURNAL OF THE AUDIO ENGINEERING SOCIETY, and is Co-Editor of the book *Ubiquitous Music Ecologies*. He is the Chair of the IEEE Emerging Technology Initiative on the Internet of Sounds and the President of the Internet of Sounds Research Network. His main research interests are in music technology, Internet of Things, human–computer interaction, and multisensory perception.

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

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 is a CNRS Research Scientist with LS2N, a French laboratory dedicated to cybernetics. 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.

Cristina Rottondi (Senior Member, IEEE) received the bachelor's and master's degrees "(*cum laude*)" in telecommunications engineering and the Ph.D. degree in information engineering from the Politecnico di Milano, Milan, Italy, in 2008, 2010, and 2014, respectively.

She is an Associate Professor with the Department of Electronics and Telecommunications of Politecnico di Torino, Torino, Italy. From 2015 to 2018, she had a research appointment with the Dalle Molle Institute for Artificial Intelligence, Lugano, Switzerland. She has coauthored more than 80 scientific publications in international journals and conferences. Her research interests include optical networks planning and networked music performance.

Dr. Rottondi is a co-recipient of the 2020 Charles Kao Award, three Best Paper Awards (FRUCT-IWIS 2020, DRCN 2017, and GreenCom 2014), and the one Excellent Paper Award (ICUFN2017). She served as an Associate Editor for IEEE ACCESS from 2016 to 2020 and is currently an Associate Editor of the IEEE/OSA JOURNAL OF OPTICAL COMMUNICATIONS AND NETWORKING.

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

He is a Senior Lecturer with the Center for Digital Music, Queen Mary University of London, London, U.K. He is an Investigator of UKRI's £6.5M Centre for Doctoral Training in Artificial Intelligence and Music and he was the QMUL's Principal Investigator on the H2020 funded Audio Commons Project. He published over 150 papers in the fields of MIR, 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 the Papers Co-Chair of the AES 53rd International Conference on Semantic Audio.

Nils Peters (Senior Member, IEEE) received the Dipl-Ing. degree in electrical and audio engineering from the University of Technology in Graz, Graz, Austria, in 2005, and the Ph.D. degree in music technology from McGill University Montreal, Montreal, QC, Canada, in 2010.

He is a Professor of Audio Signal Processing with the International Audio Laboratories, University of Erlangen-Nuremberg, Erlangen, Germany. His work focuses on the possibilities and challenges of future audio applications using the Internet of Things. Before joining the International Audio Laboratories, he was a Senior Staff Research Engineer and the Team Lead of Spatial Audio with Qualcomm's Multimedia Research and Development Department, San Diego, CA, USA, a Postdoctoral Fellow with the University of California at Berkeley, Berkeley, CA, USA, and a Visiting Researcher with IRCAM, Paris, France.

Dr. Peters is a Co-Chair of the Technical Committee for Spatial Audio at the Audio Engineering Society.

Jan Østergaard (Senior Member, IEEE) received the M.Sc. degree in electrical engineering from Aalborg University, Aalborg, Denmark, in 1999, and the Ph.D. degree (*cum laude*) from Delft University of Technology, Delft, The Netherlands, in 2007.

From 1999 to 2002, he worked as a Research and Development Engineer with ETI A/S, Aalborg, and from 2002 to 2003, he was a Research and Development Engineer with ETI Inc., Virginia Beach, VA, USA. From September 2007 and June 2008, he was a Postdoctoral Researcher with The University of Newcastle, Callaghan, NSW, Australia. He has been a Visiting Researcher with Tel Aviv University, Tel Aviv, Israel, and also with Universidad Técnica Federico Santa María, Valparaíso, Chile. He is currently a Full Professor of Information Theory and Signal Processing, the Head of the Section on AI and Sound, and the Head of the Centre on Acoustic Signal Processing Research, Aalborg University.

Dr. Østergaard has received the Danish Independent Research Council's Young Researcher's Award, the Best Ph.D. Thesis Award by the European Association for Signal Processing (EURASIP), and fellowships from the Danish Independent Research Council and the Villum Foundation's Young Investigator Programme.

Frederic Font received the M.Sc. and Ph.D. degrees in sound and music computing from Universitat Pompeu Fabra, Barcelona, Spain, in 2010 and 2015, respectively.

He is a Senior Researcher with the Music Technology Group, Department of Information and Communication Technologies, Universitat Pompeu Fabra. He is the Coordinator of the Freesound Website and related research and development projects, and has recently coordinated the EU funded Audio Commons Initiative. His current research is focused on the understanding and analysis of large audio collections, including sound characterization and classification, to improve sound retrieval techniques and to facilitate the reuse of large audio collections in creative and scientific contexts.

Tom Bäckström (Senior Member, IEEE) received the master's and Ph.D. degrees from Aalto University, Espoo, Finland, in 2001 and 2004, respectively, which was then known as the Helsinki University of Technology.

He has been an Associate Professor with the Department of Signal Processing and Acoustics, Aalto University, since 2016. He was a Professor with the International Audio Laboratory Erlangen, Friedrich-Alexander University, Erlangen, Germany, from 2013 to 2016, and a Researcher with Fraunhofer IIS, Erlangen, from 2008 to 2013. His research interests include technologies for spoken interaction, emphasizing efficiency and privacy, and in particular in multidevice and multiuser environments.

Dr. Bäckström has contributed to several international speech and audio coding standards and is the Chair and Co-Founder of the ISCA Special Interest Group "Security and Privacy in Speech Communication."

Carlo Fischione (Senior Member, IEEE) received the Laurea degree (*summa cum laude*) in electronic engineering and the Ph.D. degree in electrical and information engineering from the University of L'Aquila, L'Aquila, Italy, in 2001 and 2005, respectively.

He is a Full Professor with the KTH Royal Institute of Technology, Stockholm, Sweden. He is a Co-Founder and the Scientific Director of the music-tech company Elk, Stockholm, Sweden. He has held research positions with Massachusetts Institute of Technology, Cambridge, MA, USA, (Visiting Professor); Harvard University, Cambridge (Associate); 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 TRANSACTIONS ON COMMUNICATIONS publication of 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 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 IEEE TRANSACTIONS ON COMMUNICATIONS and an Associated Editor of *Automatica* (IFAC).

Open Access funding provided by 'Università degli Studi di Trento' within the CRUI CARE Agreement