Towards an Internet of Multisensory, Multimedia and Musical Things (Io3MT) Environment

Rômulo Vieira
MídiaCom Lab
Fluminense Federal University
Niterói, Brazil
romulo_vieira@midiacom.uff.br

Débora C. Muchaluat-Saade
MídiaCom Lab
Fluminense Federal University
Niterói, Brazil
debora@midiacom.uff.br

Pablo César
Distributed and Interactive Systems (DIS) Group
Centrum Wiskunde & Informatica - CWI
TU Delft
Amsterdam, The Netherlands
garcia@cwi.nl

***Abstract*—The Internet of Multisensory, Multimedia and Musical Things (Io3MT) is a new concept that arises from the confluence of several areas of computer science, arts, and humanities, with the objective of grouping in a single place devices and data that explore the five human senses, besides multimedia aspects and music content. In this paper, we present the first practical environment that adheres to the principles outlined in this area, providing a comprehensive specification of the architecture, devices, protocols, and tools employed in its building. To validate the technical feasibility of our proposed system, we conducted an evaluation of the system's quality of service (QoS). Furthermore, we performed a comparative analysis with related work in the field, providing a comprehensive assessment of our approach contributions and novelties.**

***Index Terms*—Internet of Sounds, Io3MT, networked musical performance, interactive art**

I. Introduction

The Internet of Things (IoT) can be defined as the interconnection of physical devices enriched with embedded electronics and software, facilitating the acquisition and transmission of data among themselves, while collaborating to accomplish a designated task [1], [2]. The proliferation of this concept has an impact on several areas of modern life, from home automation and city management to vehicle networks and e-health.

The rapid advancement within this domain has engendered a need for applications that harness the multimedia nature, thus giving rise to the Internet of Media Things (IoMT) [3]. This paradigm can be characterized as a network comprising devices endowed with distinct identifications and addresses, enabling mutual collaboration to facilitate the seamless exchange of multimedia content, encompassing text, audio, images, video, and more, among users [4], [5].

The studies and research endeavors delving deeper into the realm of sound interaction in the cyberspace culminate in a specialized sub-field known as the Internet of Sounds (IoS) [6]. This area arises from the fusion of engineering and humanities concepts, covering domains such as digital audio processing, acoustic monitoring, music and arts. The primary objective of IoS is to establish a harmonious ecosystem of interconnected devices that facilitate the seamless exchange of sound-related data. This ecosystem serves diverse purposes, ranging from enhancing art pieces, such as facilitating interaction between artists and their audiences, to enabling real-time retrieval and integration of audio files for soundscape composition. Additionally, IoS facilitates the interconnection of smart musical instruments (SMI) [7] and supports the creation of innovative wearable musical equipment.

In addressing both musical and non-musical domains, IoS incorporates two distinct paradigms, namely the Internet of Audio Things (IoAuT) [8] and the Internet of Musical Things (IoMusT) [9]. The former area pertains to a network of computational devices integrated into physical objects, with the overarching objective of facilitating sound information production, analysis, processing, transmission, and comprehension within distributed environments.

On the other hand, IoMusT is oriented towards the domain of music, encompassing systems, networks, and devices engineered to cater to a diverse array of stakeholders (performers, musicians, students, and industry professionals). IoMusT also engenders the development of pioneering models and frameworks that facilitate various applications, including immersive concert experiences, enhanced audience engagement, remote rehearsal capabilities, e-learning opportunities, and the augmentation of functionalities within smart studio environments.

Despite their common origin in the IoT domain and shared features such as the utilization of networked objects with data or media processing capabilities, these research topics have traditionally been treated as distinct subjects, often overlooking the potential benefits that interdisciplinary integration could offer [10]. An illustrative instance pertains to the Internet of Musical Things (IoMusT), where its theoretical framework envisions the integration of multisensory components, specifically incorporating tactile stimuli, gesture tracking, and physiological parameter monitoring to enhance communication among system participants. However, full implementation remains unrealized. These sensory additions mainly enhance musical performances, not being central research focuses. Moreover, IoMusT doesn't prioritize multimedia traffic in its network.

Furthermore, the predominant focus of proposed solutions in these fields has been on the visual and auditory senses, disregarding the substantial role played by non-verbal communication, which encompasses the integration of all five senses: sight, sound, touch, taste, and smell.

To address these challenges, we propose a novel research field known as the Internet of Multisensory, Multimedia and Musical Things (Io3MT), a subcategory of IoT. This pioneering approach draws upon the foundational concepts of the aforementioned areas and also of Networked Music Performance (NMP) [11], Wireless Multimedia Sensor Networks (WMSNs) [12], Internet of Senses [13], Multi-sensorial Media (MulSeMedia) [14], and Interactive Art [15]. Io3MT aims to deliver an immersive user experience, enhancing levels of engagement and perceived quality.

To establish this concept as a focal point within the Internet of Sounds community and demonstrate its practical feasibility, this paper presents the pioneering implementation of an artistic environment aligned with the principles of Io3MT. The work elucidates and apply the architectural framework, outlining its layered structure, device features, protocols, and tools employed to accommodate a diverse range of media and sensory effects. To validate the concepts presented, 10 practical tests are conducted to assess results on latency, jitter and flow. Usability and user experience analyzes are not discussed.

The remainder of this paper is organized as follows. Section II reviews networked musical performances and mulsemidia communication research. Section III discusses the proof of concept (POC) environment's setup. Section IV evaluates quality of service (QoS) and compares features with related works. Section V critically discusses results and insights. Finally, Section VI summarizes the research findings.

II. Related Work

In this section, we review some notable artistical scenarios based on the network and in multimedia concepts to extract valuable insights that can inform the design and development of our Io3MT environment [16].

A. Internet of Musical Things (IoMusT) Environments

The Internet of Musical Things (IoMusT) comprises an assemblage of systems, networks, musical things, protocols, and services that are intricately associated with music in physical and/or digital environments [10]. Specifically, IoMusT represents a distinct branch within the broader IoT framework, tailored to cater to the unique requirements of the music industry and its diverse stakeholders. Consequently, IoMusT stands as one of the foundational pillars of the Internet of Sounds. This domain offers compelling instances of immersive concerts, public engagement in artistic performances, remote rehearsals, e-learning, and smart studios [17].

An archetype created along these lines is the proposal of Turchet, Viola, Fazekas, and Barthet [18]. That exemplary aims to establish an efficient and asynchronous semantic architecture through the use of a message-oriented middleware system based on the publish-subscribe strategy. This approach ensures timely and loosely coupled communication, promoting seamless information exchange within the system.

The network implementation follows the client-server model and utilizes an oriented and labeled graph structure to encode information. The SPARQL language is employed for efficient retrieval of relevant data through querying this graph structure. This approach enables communication between devices with shared attributes, reducing data transmission volume and computational processing requirements, thereby optimizing system performance.

The environment includes five prototypes of musical artifacts that emphasize wireless connectivity using IEEE 802.11ac Wi-Fi. These devices are equipped with embedded processing capabilities powered by Bela protoboard and utilize the Pure Data language as the audio engine. Data transmission within the environment is facilitated through SPARQL requests and Python scripts, while the OSC protocol is employed for exchanging musical data among interconnected devices.

The subsequent scrutinized framework, conceived by Turchet and Barthet [19], endeavors to facilitate interactions between artists and their instruments, as well as interactions linking artists and their audiences, by generating musical accompaniments that delve into sounds of collective provenance. In this way, that representation consists of two stages. The first one focuses on enhancing the interaction between the performer and the smart mandolin [20]. This aims to explore and enrich the musician's interaction capabilities, enabling dynamic expressions [19].

The smart mandolin is an acoustic instrument with integrated technological components, such as Bela protoboard, sensors, and wireless connectivity [20]. It utilizes a Wi-Fi network and OSC messages over UDP for seamless musical data transmission. The audio engine is implemented using Pure Data, enabling real-time sound processing and a wide range of sound effects and responsive behaviors to instrument movements. A mobile application developed with TouchOSC allows the execution of pre-defined tracks and the transmission of relevant information to the instrument. Python programming language is utilized to develop a software component for efficient file retrieval and transmission to the Bela processing board, ensuring smooth communication and data transfer within the system architecture.

The second level explores the interaction between the musician and the audience, utilizing crowd-sourcing techniques. This aspect delves into engaging the audience in the musical performance, leveraging their participation and feedback to shape the overall experience.

Another proposal for IoMusT scenarios also serves as a proof of concept for a protocol known as Sunflower [21], which provides a set of predefined messages for communication among the devices involved in artistic practice. The implementation of this scheme follows the Pipes-and-Filters architecture, wherein data processing occurs within discrete units known as filters. Communication between these filters takes place through pipes. Notably, information exchange solely transpires through the input and output ports of each entity, eliminating the need for devices to have prior knowledge of their neighboring filters. Consequently, this architecture accommodates elements with varying characteristics within the same environment, promoting flexible coexistence.

The architecture is organized into distinct layers based on the musical things, data types, and protocols associated with each layer. These stages include: a digital audio layer responsible for sound generation using the Pure Data; a graphic layer ensuring visual representation capabilities, particularly those created using the Processing language; a control layer facilitating remote modification of properties such as volume, frequency, and BPM; and a management layer enabling system administrators to monitor connected devices and their key characteristics. The management layer ensures connectivity between gadgets capable of data exchange and is implemented using the Python language.

Regarding the network communication, a hybrid approach is employed using Wi-Fi based on the IEEE 802.11n standard, alongside an Ethernet cable for data transmission. The UDP protocol is utilized to transmit general data over the network, while the OSC protocol is employed specifically for music-related information exchange.

There is also the proposal designed by Centenaro, Casari and Turchet [22], which presents an architecture that provides reliability and low latency for mobile devices, while establishing guidelines, services and requirements to facilitate communication in this context. The developers achieve this by combining a next-generation radio access network (NG-RAN) and a 5G core, which efficiently transfers digital audio traffic between musical things while mediating the services.

The proposed musical things within this model encompass dedicated hardware for audio input, output, and processing, integrated with the Elk Audio operating system and a 5G communication module. Additionally, cloud computing plays a crucial role by hosting applications and services, which must either exhibit latency tolerance or be located at the network edge, particularly when they play critical roles within the environment.

Although still in its initial stages, this model offers notable strengths, including support for heterogeneous traffic and the ability to accommodate devices with diverse quality of service requirements. While audio file formats, musical information, and communication protocols specific to mobile transmission have not been addressed at this stage, the model demonstrates considerable potential for future development and advancement.

B. Multiple Sensorial Media (Mulsemedia) Environments

When delivering multimedia content over a network, it becomes imperative to employ mechanisms that guarantee a satisfactory quality of experience (QoE) for the user [23]. This metric focuses on the user's perspective and involves quantifying their perception of specific services, as well as providing a comprehensive evaluation of the level of satisfaction or annoyance experienced with an application [24]. Several factors influence this indicator, including:

- **Content immersiveness:** denotes the capability of an experience to deeply engage the viewer or user's senses and captivate their attention in a profound manner. It encompasses the sensation of complete immersion within an environment or narrative, fostering a profound sense of presence and a feeling of being an integral part of the simulated or created reality;
- **Delay in media presentation:** technique employed to decrease delay and interruptions when displaying different media;
- **Synchronization between sensory effects and media objects:** deals with the precise and intentional coordination of sensory stimuli with specific elements of a media or content;
- **Description of devices:** describes the functional capabilities of devices to ensure compatibility, create engaging experiences, enable personalized interactions, optimize resources, and explore creativity and innovation in multimedia content creation;;
- **Interactivity:** seeks to support user interaction, allowing the user to relate both to the audiovisual content and to a sensory effect that is being presented.

In response to these requirements, various studies in the field have proposed strategies to address each of these factors, such as the proposals by [14], [25]–[27], that suggest adding sensory effects in multimedia applications in order to make the content more immersive for the user, and consequently improve the QoE. Furthermore, Meixner et al. [28] employ a prefetching mechanism to reduce delays and interruptions during the presentation of different media. Conversely, Davison [29] and Su et al. [30] propose automated mechanisms to address this issue, eliminating the need for the multimedia application author to manually control which content should be prefetched [23].

To maintain synchronization between sensory effects and media objects, Yuan et al. [31] use "synchronization regions" that represent temporal intervals in which aroma dispersion and touch sensation effects must be triggered by the actuating devices. When the effect is triggered in an instant within this region, the user considers that it is synchronized with the visual content. In the framework proposed by [32], the metadata of effects encapsulated in an MPEG-2 TS stream (Transport Stream) is used to control media synchronization. The work of [33] highlights that the calculation of the time for sending effect metadata must consider the time required for the preparation of sensory devices, in addition to the transmission time of metadata in the network and the time of presentation of the effect in the application. The work by Su et al. [30] indicates the use of mechanisms to be used in the transmission phase of the application, to avoid synchronization failures between sensory effects and audiovisual content.

With regard to the description of the devices, the study by Choi et al. [34] recommends the use of Part 2 of the MPEG-V standard to obtain information that can be used to provide the streaming service in an appropriate manner, while the interactivity attributes must allow the user to interact both with the audiovisual content and with a sensorial effect that is being presented, which makes the response time to an interaction an important factor for the quality of user experience. In this sense, Santos et al. [35] also presents a proposal to improve response time in event-based mulsemedia applications.

III. Io3MT Environment Design

Our system design encompasses the exposition of the architecture, interfaces, protocols, data structures, and decision-making processes that collectively contribute to the successful execution of a specific task. Consequently, this section is devoted to providing a comprehensive overview of such information pertaining to an Io3MT environment.

A. Architecture

Given that this model is the inaugural proposition of an Io3MT-based environment, a qualitative research approach and a selective bibliographic review were undertaken, drawing insights from the works presented in Section II and the relevant literature on the Internet of Sounds. The investigation revealed that an effective network structure for artistic-musical presentations should possess certain overarching characteristics, including **low latency**, **interoperability**, and **scalability** [36].

Each system also has its specific requirements. In the case of Io3MT, it needs to be able to deal with different types of signals, including audio, video, images and sensory stimuli, among others. Additionally, it should facilitate the inclusion and active participation of non-musicians, thereby fostering a collaborative and inclusive creative process.

The proposed Io3MT architecture consists of five distinct layers, as summarized in Table I. Each layer serves specific functions and is designed to meet the requirements of the preceding layers. Moreover, the stakeholders associated with each level are also identified, highlighting their roles and responsibilities within the Io3MT environment.

TABLE I: Io3MT Layers.

Layer Name	Goal	Stakeholder
Things Layer	Virtualize data coming from physical devices	Musicians, Performers, Artists, Audience, Designers, Industry
Link Layer	Connect objects to the network	Technical Groups, Manufacturers, Suppliers and Researchers
Network Layer	Responsible for packet forwarding infrastructure	Technical Groups, Manufacturers, Suppliers and Researchers
Middleware Layer	Data Management	Entertainment/Tech Industry, Labels, Publishers, Distributors, Manufacturers
Application Layer	Provides services requested by customers	Musicians, Performers, Artists, Audience, Students, Teachers

B. Devices

An additional noteworthy aspect is the behavior exhibited by the devices within the Io3MT environment, encompassing both physical and virtual entities. These devices are classified as “Multisensory, Multimedia, and Musical Things” (3MT). From a functional perspective, they are characterized by their ability to perform at least one multisensory, multimedia, or musical action, either by generating or responding to such content. Moreover, these devices are equipped with on-board electronics, wireless communication capabilities, and support functionalities for remote control, customization, or upgrade. Furthermore, the 3MT devices possess several essential attributes. They are uniquely identifiable and addressable, ensuring seamless communication and interaction. The devices are designed to be scalable, accommodating the potential expansion of the environment. Additionally, they exhibit persistence and reliability, allowing for continuous and dependable operation within the Io3MT framework.

In terms of functionalities, these devices perform various tasks, including but not limited to motion detection, media rendering, object tracking, and media compression/streaming. Examples of such gadgets are diverse, encompassing a range of technologies. These include cameras, media storage devices, wearables (such as smart glasses and virtual reality headsets), smart lights, wind fans, scent dispersers, heat lights, smoke machines, smartphones, tablets, loudspeakers, microphones, and smart musical instruments. Each device contributes to the immersive and interactive nature of the Io3MT environment, enabling a rich and dynamic user experience.

From a technical point of view, they must contain the ability to send beacons, route frames, contain their own power supply and be context sensitive, minimizing redundant data acquisition. As these are devices with artistic concerns, aesthetic, expressive and ergonomic factors are also relevant [37].

C. Protocol Stack

In the protocol stack of the environment, the network layer employs the Internet Protocol (IP), while the transport layer utilizes the UDP and TCP. The selection of IP at the network layer is motivated by its widespread adoption and ability to facilitate seamless communication among devices. UDP is chosen in the transport layer to transfer sound data, since it does not incorporate mechanisms for packet re-transmission, in-order delivery, or congestion control, being suitable for applications that require low latency and real-time data transfer, in line with the requirements of the audio specifications in an Io3MT environment. TCP, on the other hand, is responsible for adaptive video streaming, since it has reliability, congestion control, and delivery guarantee, which makes the delivery of this content more efficient, fluid, and adapted to the bandwidth available in the user’s connection.

When considering the symbolic representation of multimedia information, the MIDI protocol emerges as a well-established framework for exchanging musical data. However, we have chosen to employ the OSC protocol. This model offers enhanced flexibility for organizing and transmitting sound control information across networks. Noteworthy features of OSC include URL-like symbolic nomenclature, support for high data resolution, the ability to specify multiple recipients for a single message, and concurrent packet delivery. These qualities make OSC an advantageous choice for facilitating efficient and versatile communication of multimedia information within the Io3MT framework [18].

D. Tools

A valuable tool for facilitating this process is the Nested Context Language (NCL) - ITU-T H.761 [38]. NCL is a language specifically designed for defining hypermedia documents composed of various media elements such as text, images, graphics, audio, video, and animations. It enables authors to describe the behavior and interactivity of these objects within the document. Notably, NCL utilizes a declarative approach, ensuring that the meaning of each statement is independent of its specific usage and execution details. In a nutshell, NCL acts as a "glue" that binds media objects together in a presentation, defining how they are structured and how they are related, not restricting or pre-screening the types of contents that are accepted by media objects [39]. While NCL has historically been employed in Digital TV services, its utilization has expanded to encompass a wide array of multisensory applications, leveraging the increasing prevalence of sensors and interconnected devices within the IoT paradigm.

Even though NCL is easy to learn for people with no programming experience, it lacks flexibility and lack of imperative assistance, aspects that directly impacts mathematical processing, text manipulation, use of the interactivity channel, and animations and collisions between objects. Furthermore, its use is advantageous only when the application depends on resources foreseen in the scope of the language [40].

To enhance the expressive capabilities of NCL and enable its seamless adaptation to the requirements of Io3MT, integration with a scripting language becomes imperative. Lua [41], a lightweight, portable, and efficient scripting language designed for simplicity and ease of embedding within other languages, presents a suitable solution to address this gap. The primary distinguishing feature of Lua lies in its utilization of a single type of data structure known as a table, which represents an associative array or dictionary [42]. By incorporating Lua into the NCL framework, the combined solution becomes more flexible and versatile, empowering content authors to effectively manipulate and control the behavior of media objects.

To ensure the platform-independent execution of NCL and Lua files, the integration of a middleware layer becomes necessary. This software layer serves the crucial role of facilitating compatibility and support across different receiving platforms. Given the multimedia nature of the application, the middleware must possess adaptability to cater to the diverse requirements of the network. Moreover, it should enable the provisioning of services for pervasive end users, offer distinct services, and accommodate multiple users [3].

An exemplary middleware that fulfills these requirements is Ginga (ITU-T H.761) [43], [44]. The architecture of Ginga comprises four primary elements. Firstly, there are the Exhibitors (or Media Players), which possess specific functionalities associated with the respective media types they handle. Additionally, there is the Formatter (or Orchestrator), responsible for controlling the temporal and spatial synchronization of the NCL presentation. The Parser undertakes the crucial task of conducting syntactic and semantic analysis of the content, while the Document represents the object tree of the application [45]. The relationships among these components are depicted in Figure 1 [46].

In a broad sense, Ginga operates by receiving an NCL application within its formatter component. Subsequently, the parser translates the specifications of the application into data structures that Ginga-NCL can comprehend, enabling control over the application's presentation. The scheduler component is then initiated to orchestrate the presentation, employing techniques such as content pre-fetching, assessment of link conditions, and scheduling of actions. Further, the scheduler is responsible for instantiating the relevant players capable of handling specific media types. Through this process, Ginga ensures seamless execution and synchronization of multimedia content within the NCL application [45].

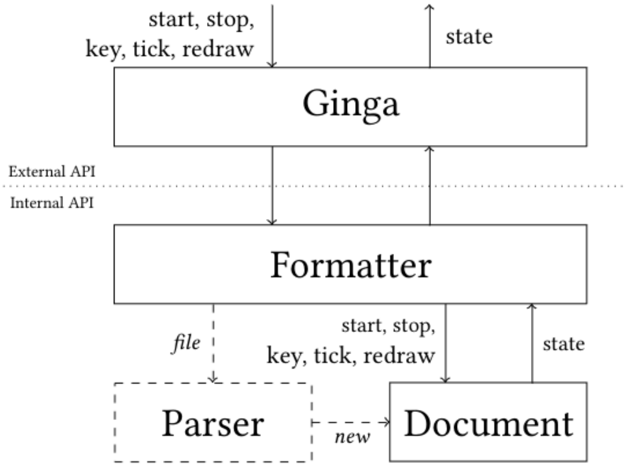


Fig. 1: Ginga middleware architecture [46].

During the analysis of the related work to Io3MT in Section II, it is evident that Pure Data is extensively used as an audio engine, responsible for generating and capturing audio data for network transmission. Language's popularity stems from its cross-platform compatibility, extensibility, and versatility, making it suitable for various devices and integration with Io3MT. Its reprogramming capabilities, code reuse, and real-time language updates facilitate easy patch development and scalability are also points that corroborate your choice. This tool supports diverse audio and video files, as well as MIDI and OSC protocols for exchanging musical information. This versatility enables flexible data flow and integration of devices, enhancing system heterogeneity. In the creative process, artists and users of different proficiencies can actively participate [47], [48]. Pd also serves as a conduit for electroacoustic instruments to access the network, transcending its role as an audio engine. In view of these advantages, we also adopted Pure Data in our system.

IV. Technical Validation

To validate our proposed concepts, we implemented a practical scenario that simulates a live performance within the Io3MT paradigm, as shown in Figure 2. We used physical equipment like HP 256 Laptop (Computer A), Acer Predator Helios 300 G3 Laptop (Computer B) both running Linux Mint xfce 20.04, a Condenser Microphone, 6-String Electric Guitar, Behringer UMC202HD soundcard, Moodo AIR fragrance dispenser, Yeelight E27 smart bulb, and a projection device. In addition, we created three virtual 3MT instances using Pure Data. The first two instances captured audio data from the microphone and guitar, transmitting it over the network, while the third generated controllable visual effects. NCL 4.0 served as the central platform for orchestrating these actions, integrating content, managing presentation timing, synchronizing sensory effects and media objects, describing devices, and facilitating Io3MT ecosystem interactivity.

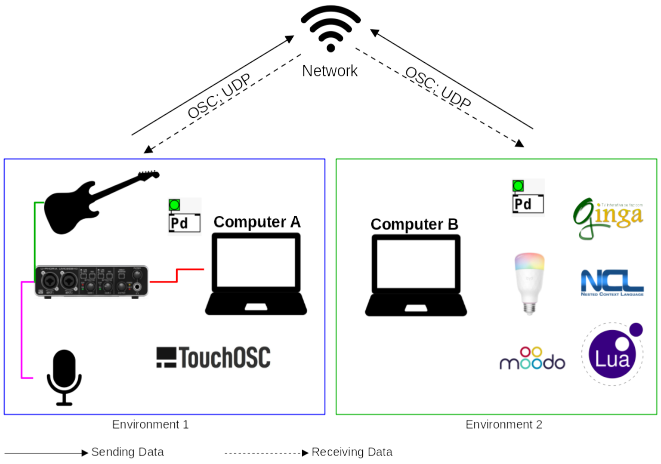


Fig. 2: Environment Setup.

Due to the lack of native support for Pure Data patches in the NCL language, Lua scripts were employed to act as players for the guitar and microphone media. These scripts are responsible for starting, stopping, and modifying these events' properties. Considering the limitations of Pure Data in terms of data structures, iteration, and recursion capabilities, an external module called GingaPD was developed to facilitate communication between the patches and the hypermedia document. This module was created based on the pd-lua library. As a result, the control object implemented in GingaPD enables the transmission of values from the guitar and microphone volume and reverb sliders, which control the intensity of light and aroma effects. Additionally, it can receive information from these multisensory objects to modify their properties accordingly.

The properties associated with the different elements in the environment, including microphone volume, reverb settings (room size, damping level, wet signal), guitar volume, intensity of light and aroma effect, and geometric shapes of visual art, were simulated using TouchOSC. This allowed for easier control of these parameters by the artist and provided a visual representation of the environment's behavior. Figure 3 illustrates the TouchOSC interface used for this purpose.

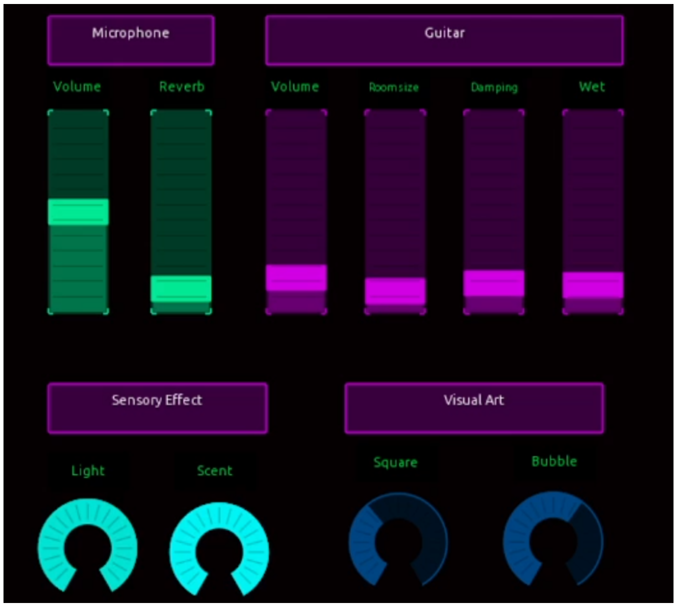


Fig. 3: The layout of the controller app.

The link layer used Wi-Fi in the IEEE 802.11n standard, with a theoretical transmission limit of up to 100 Mbps, using a TP-Link AC 1200 Archer C5 router. In order to ensure optimal performance and minimize latency and packet loss, a dedicated network configuration was established, exclusively connecting the devices involved in the communication. For communication with the smart lamp and aroma diffuser, REST API requests were utilized. These requests offer a range of functions, including establishing a connection with the physical devices, activation and deactivation of the devices, and commands to adjust the intensity of the effect rendering. Additionally, specific functionalities can be implemented, such as commands to modify the color of a light effect. As communication with the Moodo device occurs through REST API requests, it involves the use of a JSON file for input and output of information and the TCP protocol for sending data over the network, while messages for smart bulb also follow a format similar to that transmitted by the TCP protocol. The exchange of audio data is facilitated through the UDP protocol, while the control of music and multimedia information is carried out using the OSC protocol.

The network layer assumed the responsibility of data addressing through the utilization of the IP protocol, as previously indicated. In this context, the audio data from the microphone was transmitted from Computer A (192.168.0.20) to Computer B (192.168.0.22) via port 3000, whereas the guitar data was relayed through port 3001. Concurrently, control data was conveyed via port 10000. As the audio played on Computer B was subject to modifications due to changes in media and sensory effects commanded by the NCL program, it passed such information on to Computer A using port

20000. This clear differentiation brought transparency to the system, that is, it was noticeable which data altered a certain parameter, and also allowed a subsequent analysis of such data, identifying latency points and the like.

Next, the middleware layer relies on Ginga to process and orchestrate the media specified in the NCL document, in addition to synchronizing all the elements present in this scenario, performing concurrency control and transitions, ensuring persistence and stream processing.

Finally, the fifth layer displays the artistic application of this scenario. In this sense, computer A sends real-time audio to computer B. Concurrently, computer B displays a series of videos that contribute to the storytelling. From actions in audio and videos (as they are started, paused, or finished), events are triggered in sensory equipment, causing changes in colors and light intensity, and dispersion of different fragrance effects in Moodo. In this way, an interoperable and hierarchical environment is created, where musical properties are controlled by video actions, images, or other types of media, the intensity of an aroma is impacted by the lamp color, and so on. A demonstration of the environment's operation can be observed through a video available on YouTube[1], and the source codes are accessible in a GitHub[2] repository.

V. Evaluation and Discussion

Before testing, it's essential to explain Pure Data's operation and the latency it introduces. In Pure Data, audio processing happens in blocks of samples, with varying sizes like 64, 128, 256, 512, 1024, or 2048 samples per block. These blocks move from kernel space to user space, and their processing time depends on the block size, remaining constant within the same soundcard but varying across multiple machines. This processing time corresponds to Pure Data's latency. In our environment, with a sampling rate of 44100 Hz, this latency is 1.45 ms [21]. Note that this latency is inherent to audio systems, and our analysis focuses on block-based information processing, excluding network processing time.

With the aforementioned information, we can now delve into the testing phase. A total of ten sessions, each lasting approximately five minutes, were conducted to collect data for assessing the QoS in the proposed environment. This metric focuses on analyzing three key aspects: latency, jitter, and throughput. Latency can be determined by calculating the average difference between the relative sound time and the expected time, along with the minimum latency time introduced by Pure Data, which has been determined to be 1.45 ms. In artistic network presentations, it is crucial to ensure that the latency value does not exceed 40 ms [49].

Jitter was assessed by measuring the standard deviation of latency values. The third parameter, throughput, was calculated by dividing the total number of packets transmitted by the duration of each test. The time measurement encompassed the interval from the initial sample to the final one, taking into account any observed gaps between samples.

It is worth noting that the expressive qualities of the musician, such as intensity, accentuation, and articulation, contribute to variations in the interpretation of the same composition during the 10 tests. As a result, the number of packets transmitted differs across these sessions. Table II provides a summary of the total packets transmitted in each session, further categorizing them into audio data and control data. These values were captured using the Wireshark software.

TABLE II: Number of packets sent in each of the tests.

Test	Audio Packets	Control Packets
Test 1	10416	273
Test 2	11582	185
Test 3	10698	148
Test 4	10715	250
Test 5	11511	248
Test 6	11885	418
Test 7	11259	268
Test 8	12471	392
Test 9	12468	149
Test 10	12343	375

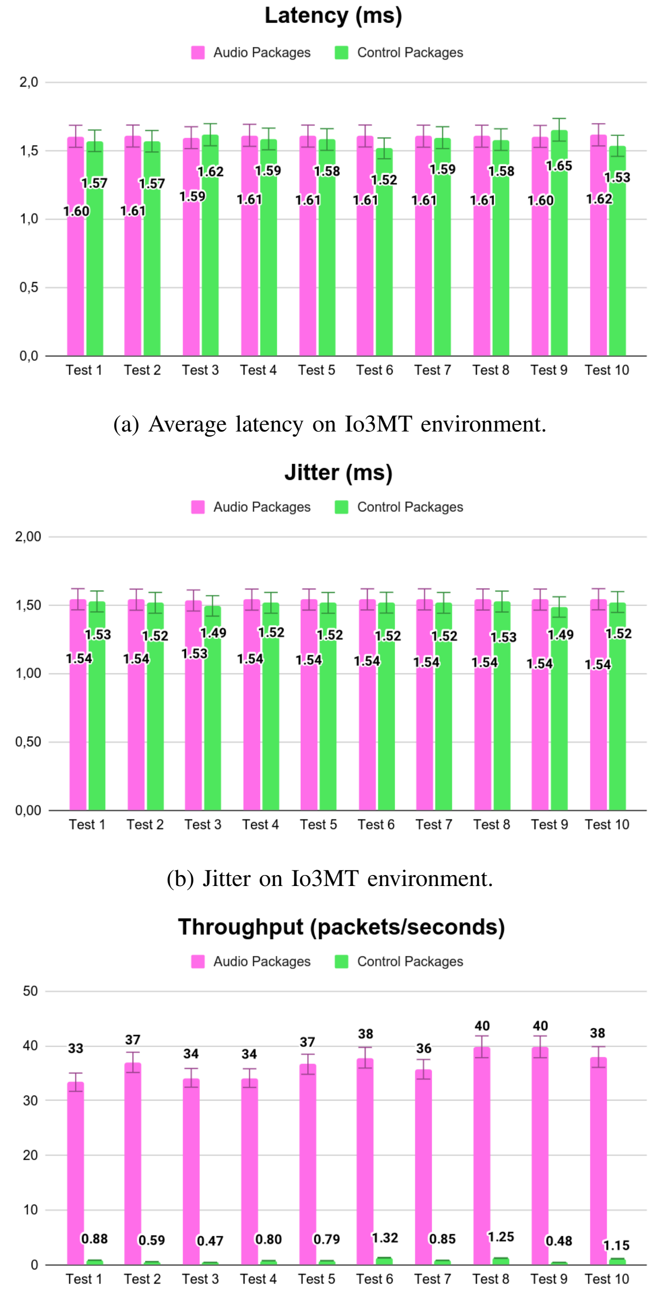
Figure 4 presents the measured values for each individual test. The analysis reveals that the latency for audio data ranged from 1.59 ms to 1.62 ms, with test 3 yielding the best results. As for control data, the latency varied between 1.53 ms and 1.65 ms. Since changes in control data occur linearly (e.g., moving the volume from 0 to 10 requires sending 10 packets), expressiveness has a significant impact on this metric. The varying number of packets transmitted in different tests directly affects latency. Consequently, test 6 demonstrated the best results in terms of control data latency.

The jitter values exhibit a similar pattern to latency, with minimal variation observed in audio data and a wider range observed in control data, ranging from 1.49 ms to 1.52 ms. Test 3 performed better in terms of audio data jitter, while tests 2 and 9 excelled in control data jitter. Given the significantly higher amount of audio data in the environment, the average throughput of packets per second ranged from 33 to 40, with tests 8 and 9 transmitting the highest volume of data. In contrast, the sparse nature of control information resulted in an average of 0.5 to 1.2 packets per second. Test 10 demonstrated the best performance in terms of control data throughput.

After an in-depth exploration of the lower layers, significant insights can be derived concerning the Middleware and Application layers. Specifically focusing on the Middleware layer, its central responsibility revolves around the management of the diverse data types prevalent within the environment. Notably, the adoption of NCL stands out, as it operates by establishing associations between different media objects without explicitly specifying the underlying mechanisms. This particular approach has proven to be both valuable and effective in ensuring the synchronization of video, audio, and sensory effects. Furthermore, the implementation has demonstrated a notable absence of delays during media initiation, as well as the mitigation of execution issues such as quality

[1] https://www.youtube.com/watch?v=FCDt4SbeqXM

[2] https://github.com/blindreview182/io3mt-environment



(a) Average latency on Io3MT environment.

(b) Jitter on Io3MT environment.

(c) Throughput on Io3MT environment.

Fig. 4: Io3MT environment network test results.

degradation, frame loss, and freezing occurrences. Also, the real-time capabilities of Pure Data have further facilitated dynamic manipulations and swift adjustments, thereby enhancing control and adaptability during performance.

Regarding the Application Layer, it facilitated the provision of bespoke services tailored to individual artists/users, incorporating a diverse array of video media and audio tracks, whether pre-recorded or performed live, in order to construct an artistic narrative. Additionally, the Application Layer enabled the automation of light colors and aroma effects in accordance with specific scenes, thereby affording the artist the luxury of concentrating on the sonic aspects of the presentation. Consequently, this automation contributed to a more streamlined execution of the creative art and also led to a reduction in the overall workload involved.

Regarding possible challenges in the evolution of this scenario, technological issues can be mentioned, such as real-time audio quality and device synchronization, as well as the need to create new data formats. Additionally, there are multisensory challenges related to integrating sensory effects, such as sound and odor, into wearable devices. Privacy and security are important concerns given the sensitive information transmitted on Io3MT. Economic issues involve the transformation of the music industry and the balance between automation and artistic quality. Social challenges include the uneven distribution of technology and its impact on socioeconomic disparities. Finally, there are environmental challenges related to the life cycle of connected devices and their environmental impact.

To contextualize the Io3MT environment within the broader field of Internet of Sounds and to enhance the relevance of this study for researchers, musicians, and designers, we conducted a comparative analysis with related works discussed in Section II. An overview of this comparison is provided in Table III.

The comparative analysis presented in the table reveals that the examined ecosystems predominantly utilize Pure Data as an audio engine, with Python being commonly employed as a scripting language. In contrast, our proposed model places emphasis on the synergy between NCL, Lua, and Ginga for handling hypermedia and multisensory content, while leveraging the capabilities of Pure Data for real-time audio processing tasks.

In terms of media utilization, it is evident that each model places an emphasis on audio. However, Sunflower [21] also incorporates videos and animations to increase system heterogeneity and validate the protocol. In contrast, our proposed ecosystem places multimediality as a foundational element within its architecture. Consequently, real-time audio processing (utilizing Pure Data through PCM encoding), pre-recorded tracks (WAV and OGG formats), and videos (MP4 and AVI files) are employed. The technologies integrated within its framework also allow for the inclusion of images (JPEG, PNG, etc.), textual objects (TXT, PDF, etc.), imperative objects (Lua, Pd), declarative objects (HTML, LIME, SVG, etc.), and so forth.

Concerning the use of multisensory effects, the Turchet, Viola, Fazekas & Barthet model [18] and Sunflower [21] touch this point. Despite not using this type of content in their practical presentations, they indicate that the technologies employed by each of them support a considerable variety of sensory elements, from light effects to smoke machines and screens with visual content. However, they treat these contents as support tools in the artistic presentation, while our model practically uses light and aroma effects, and in addition, does not prioritize this type of information about the audio, all of which are extremely important for our proposal. The

TABLE III: Synthesis of Environments Comparison.

Feature	**Turchet, Viola, Fazekas & Barthet Model** [18]	**Turchet & Barthet Model** [19]	**The Sunflower Model** [21]	**Centenaro, Casari & Turchet Model** [22]	**Our proposal: Io3MT Environment**
Technologies	Pure Data, SPARQL, Python	Pure Data, Python	Pure Data, Python, Processing	Elk Audio OS	Pure Data, NCL, Lua, Ginga, JSON
Media	Audio	Audio	Audio, Video, Animations	Audio	Audio, Video, Images, Text, etc.
Multisensory Effects	Light, visual screens, smoke machines	N/A	Audio, visual screens	N/A	Light and scent effects
Protocols	HTTP, UDP, OSC, Wi-Fi	Wi-Fi, UDP, OSC	UDP, OSC, MIDI, Wi-Fi	5G	HTTP, UDP, TCP, OSC, Wi-Fi

Io3MT system also supports haptic information, and sensory information from the skin (for example, airflow, temperature, humidity, etc.). Furthermore, sensory effects can impact and/or be impacted by media and sound content, allowing for an interchangeable and dynamic system, and opening up new possibilities for performers and composers, who will now have new types of elements in their artworks.

Upon examining the protocols utilized, it is evident that this aspect demonstrates significant convergence among the compared examples. With the exception of Centenaro, Casari & Turchet proposal [22], which emphasizes communication via 5G, all other environments utilize Wi-Fi as the underlying network infrastructure, the UDP protocol for audio and control data exchange, and OSC (with partial MIDI support) for the transmission of musical and multimedia information.

VI. Conclusion

This paper presented a proof of concept for an Io3MT environment, encompassing its layered architecture, devices, protocols, and tools employed in its development. The study demonstrates promising results by encompassing artistic presentations that involve the exchange of multisensory, multimedia, and musical information over the network. This not only impacts artistic creation but also holds potential for various sectors such as entertainment, tourism, healthcare, and more. To the best of our knowledge, there are no similar works documented in the existing literature, thus addressing a significant gap in the advancement of this field. Additionally, we introduce a novel utilization of NCL/Ginga and develop an external Pure Data module to support this environment, which can also be applied in other applications.

However, there are some limitations. The implemented use case focused on localized configurations, where all devices were connected through a Wi-Fi-based wireless LAN. In this way, the architecture can be extended to support remote interactions over a wide area network. The network can also be improved to accommodate numerous users and equipment, maintaining low latency and file download capacity, also avoiding congestion. From an artistic point of view, the system has an inherently heterogeneous trait, which makes it difficult to deal with some aspects such as input integration from different multimodal interfaces, delay, responsiveness, sensory effect intensities, wearables and other varied devices to deliver sensory effects. These points still require evolution in our proposal.

As part of future work, our intention is to conduct a formal evaluation of the user experience and the level of creativity support offered by our system. Additionally, we propose to enhance the NCL language by integrating features specifically tailored for real-time audio processing. This enhancement aims to eliminate the necessity of Lua scripts for specifying mulsemedia applications, enabling NCL authors to seamlessly incorporate real-time audio functionalities within their creations.

It is important to emphasize that this work should not be perceived as a monolithic solution, but rather as an open framework that welcomes contributions from other users and developers, as well as the potential for integration into their own projects. We encourage the community to further evolve the concepts presented in this study and explore their application in various domains, such as public perception, interactive arts, education, and aesthetics. By fostering collaborative efforts and embracing diverse perspectives, new avenues for research and innovation can be explored, ultimately advancing the field and facilitating novel discoveries and insights.

Acknowledgment

The authors would like to thank the support of the following funding agencies CAPES (Grant Number 88887.668167/2022-00), CAPES Print, CNPq, and FAPERJ.

References

[1] L. Atzori, A. Iera, and G. Morabito, “The internet of things: A survey,” *Computer Networks*, vol. 54, no. 15, pp. 2787–2805, 2010.

[2] D. M. G. Pereira, F. J. da Silva e Silva, C. de Salles Soares Neto, and Álan Lívio Vasconcelos Guedes, “A middleware perspective for integrating ginga-ncl applications with the internet of things,” in *Anais Estendidos do XXIII Simpósio Brasileiro de Sistemas Multimídia e Web*. Porto Alegre, RS, Brasil: SBC, 2017, pp. 70–75.

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

[4] A. Nauman, Y. A. Qadri, M. Amjad, Y. B. Zikria, M. K. Afzal, and S. W. Kim, “Multimedia internet of things: A comprehensive survey,” *IEEE Access*, vol. 8, pp. 8202–8250, 2020.

[5] A. Floris and L. Atzori, “Quality of experience in the multimedia internet of things: Definition and practical use-cases,” in *2015 IEEE International Conference on Communication Workshop (ICCW)*, 2015, pp. 1747–1752.

[6] L. Turchet, M. Lagrange, C. Rottondi, G. Fazekas, N. Peters, J. Østergaard, F. Font, T. Bäckström, and C. Fischione, “The internet of sounds: Convergent trends, insights, and future directions,” *IEEE Internet of Things Journal*, vol. 10, no. 13, pp. 11 264–11 292, 2023.

[7] L. Turchet, "Smart musical instruments: Vision, design principles, and future directions," *IEEE Access*, vol. 7, pp. 8944–8963, 2019.

[8] 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 of Things Journal*, vol. 7, no. 10, pp. 10 233–10 249, 2020.

[9] L. Turchet, C. Fischione, G. Essl, D. Keller, and M. Barthet, "Internet of musical things: Vision and challenges," *IEEE Access*, vol. 6, pp. 61 994–62 017, 2018.

[10] L. Turchet and C. Rottondi, "On the relation between the fields of networked music performances, ubiquitous music, and internet of musical things," *Personal and Ubiquitous Computing*, 10 2022.

[11] B. Loveridge, "Networked music performance in virtual reality: current perspectives," *Journal of Network Music and Arts*, vol. 2, no. 1, p. 2, 2020.

[12] M. Al Nuaimi, F. Sallabi, and K. Shuaib, "A survey of wireless multimedia sensor networks challenges and solutions," in *2011 International Conference on Innovations in Information Technology*. IEEE, 2011, pp. 191–196.

[13] D. Panagiotakopoulos, G. Marentakis, R. Metzitakos, I. Deliyannis, and F. Dedes, "Digital scent technology: Toward the internet of senses and the metaverse," *IT Professional*, vol. 24, no. 3, pp. 52–59, 2022.

[14] G. Ghinea, C. Timmerer, W. Lin, and S. R. Gulliver, "Mulsemedia: State of the art, perspectives, and challenges," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 11, no. 1s, oct 2014.

[15] T. Cerratto-Pargman, C. Rossitto, and L. Barkhuus, "Understanding audience participation in an interactive theater performance," in *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, ser. NordiCHI '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 608–617.

[16] R. Vieira and D. C. Muchaluat-Saade, "A survey on the internet of musical things: Environment challenges, standards, services, and future visions," in *2022 IEEE 8th World Forum on Internet of Things (WF-IoT)*, 2022, pp. 1–6.

[17] L. Turchet and C. N. Ngo, "Blockchain-based internet of musical things," *Blockchain: Research and Applications*, vol. 3, no. 3, p. 100083, 2022.

[18] L. Turchet, F. Viola, G. Fazekas, and M. Barthet, "Towards a semantic architecture for the internet of musical things," in *IEEE Open Innovations Association*, 11 2018.

[19] L. Turchet and M. Barthet, "Jamming with a smart mandolin and freesound-based accompaniment," in *2018 23rd Conference of Open Innovations Association (FRUCT)*, 11 2018.

[20] L. Turchet, "Smart mandolin: Autobiographical design, implementation, use cases, and lessons learned," in *Proceedings of the Audio Mostly 2018 on Sound in Immersion and Emotion*, ser. AM'18. New York, NY, USA: Association for Computing Machinery, 2018.

[21] R. Vieira and F. Schiavoni, "Sunflower: An environment for standardized communication of iomust," in *Proceedings of the 16th International Audio Mostly Conference*, ser. AM '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 175–181.

[22] M. Centenaro, P. Casari, and L. Turchet, "Towards a 5g communication architecture for the internet of musical things," in *2020 27th Conference of Open Innovations Association (FRUCT)*, 09 2020, pp. 38–45.

[23] M. Josué, M. Moreno, and D. Muchaluat-Saade, "Mulsemedia preparation: A new event type for preparing media object presentation and sensory effect rendering," in *Proceedings of the 10th ACM Multimedia Systems Conference*, ser. MMSys '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 110–120.

[24] P. Juluri, V. Tamarapalli, and D. Medhi, "Measurement of quality of experience of video-on-demand services: A survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 401–418, 2016.

[25] E. Saleme, C. Saibel Santos, and G. Ghinea, "A mulsemedia framework for delivering sensory effects to heterogeneous systems," *Multimedia Systems*, vol. 25, p. 27, 08 2019.

[26] J. Kim, C.-G. Lee, Y. Kim, and J. Ryu, "Construction of a haptic-enabled broadcasting system based on the mpeg-v standard," *Image Commun.*, vol. 28, no. 2, p. 151–161, feb 2013.

[27] M. Waltl, C. Timmerer, and H. Hellwagner, "Improving the quality of multimedia experience through sensory effects," in *2010 Second International Workshop on Quality of Multimedia Experience (QoMEX)*, 2010, pp. 124–129.

[28] B. Meixner and C. Einsiedler, "Download and cache management for html5 hypervideo players," in *Proceedings of the 27th ACM Conference on Hypertext and Social Media*, ser. HT '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 125–136.

[29] B. D. Davison, "Predicting web actions from html content," in *Proceedings of the Thirteenth ACM Conference on Hypertext and Hypermedia*, ser. HYPERTEXT '02. New York, NY, USA: Association for Computing Machinery, 2002, p. 159–168. [Online]. Available: https://doi.org/10.1145/513338.513380

[30] Z. Su, Q. Yang, and H.-J. Zhang, "A prediction system for multimedia pre-fetching in internet," in *Proceedings of the Eighth ACM International Conference on Multimedia*, ser. MULTIMEDIA '00. New York, NY, USA: Association for Computing Machinery, 2000, p. 3–11.

[31] Z. Yuan, T. Bi, G.-M. Muntean, and G. Ghinea, "Perceived synchronization of mulsemedia services," *IEEE Transactions on Multimedia*, vol. 17, pp. 1–1, 07 2015.

[32] K. Yoon, "End-to-end framework for 4-d broadcasting based on mpeg-v standard," *Image Commun.*, vol. 28, no. 2, p. 127–135, feb 2013.

[33] M. Waltl, C. Timmerer, B. Rainer, and H. Hellwagner, "Sensory effects for ambient experiences in the world wide web," *Multimedia Tools and Applications*, vol. 70, 08 2011.

[34] B. Choi, E.-S. Lee, and K. Yoon, "Streaming media with sensory effect," in *2011 International Conference on Information Science and Applications*, 2011, pp. 1–6.

[35] E. B. Saleme, C. A. S. Santos, and G. Ghinea, "A conceptual architecture and a framework for dealing with variability in mulsemedia systems," in *Anais Estendidos do XXVI Simpósio Brasileiro de Sistemas Multimídia e Web*. Porto Alegre, RS, Brasil: SBC, 2020, pp. 5–8.

[36] L. Turchet and F. Antoniazzi, "Semantic web of musical things: Achieving interoperability in the internet of musical things," *Web Semant.*, vol. 75, no. C, jan 2023.

[37] R. Vieira, L. Gonçalves, and F. Schiavoni, "The things of the internet of musical things: defining the difficulties to standardize the behavior of these devices," in *2020 X Brazilian Symposium on Computing Systems Engineering (SBESC)*, 2020, pp. 1–7.

[38] J. Russell and R. Cohn, *Nested Context Language*. Book on Demand Ltd, 2013.

[39] Álan L. V. Guedes, M. Cunha, H. Fuks, S. Colcher, and S. D. Barbosa, "Using ncl to synchronize media objects, sensors and actuators," in *Anais Estendidos do XXII Simpósio Brasileiro de Sistemas Multimídia e Web*. Porto Alegre, RS, Brasil: SBC, 2016, pp. 184–189.

[40] M. Josué, R. Abreu, F. Barreto, D. Mattos, G. Amorim, J. dos Santos, and D. Muchaluat-Saade, "Modeling sensory effects as first-class entities in multimedia applications," in *Proceedings of the 9th ACM Multimedia Systems Conference*, ser. MMSys '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 225–236.

[41] R. Ierusalimschy, L. H. de Figueiredo, and W. Celes, "The evolution of lua," in *Proceedings of the Third ACM SIGPLAN Conference on History of Programming Languages*, ser. HOPL III. New York, NY, USA: Association for Computing Machinery, 2007, p. 2–1–2–26.

[42] R. Ierusalimschy, *Programming in lua*. Lua Org, 2006.

[43] G. Baum and L. F. G. Soares, "Ginga middleware and digital tv in latin america," *IT Professional*, vol. 14, no. 4, pp. 59–61, 2012.

[44] M. F. Moreno, R. S. Marinho, and L. F. Gomes Soares, "Ginga-ncl architecture for plug-ins," in *Proceedings of the 1st Workshop on Developing Tools as Plug-Ins*, ser. TOPI '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 12–15.

[45] R. O. Rodrigues, M. I. P. Josué, R. S. Abreu, G. F. Amorim, D. C. Muchaluat-Saade, and J. A. F. d. Santos, "A proposal for supporting sensory effect rendering in ginga-ncl," in *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web*, ser. WebMedia '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 273–280.

[46] L. F. Gomes Soares, M. F. Moreno, C. D. Salles Soares Neto, and M. F. Moreno, "Ginga-ncl: Declarative middleware for multimedia iptv services," *IEEE Communications Magazine*, vol. 48, no. 6, pp. 74–81, 2010.

[47] J. Kreidler, *Loadbang: Programming Electronic Music in Pure Data*, 1st ed. Wolke Verlagsges, 2013.

[48] R. A. V. Costa and F. L. Schiavoni, "Internet of musical things and pure data: Perfect match?" *Cultural Arts Research and Development*, vol. 2, no. 1, pp. 1–10, 2022. [Online]. Available: http://ojs.bilpub.com/index.php/card/article/view/13

[49] F. Schiavoni, M. Queiroz, and F. Iazzetta, "Medusa - a distributed sound environment," in *Linux Audio Conference*, 08 2011.