Design Recommendations for a Collaborative Game of Bird Call Recognition Based on Internet of Sound Practices*

Emmanuel Rovithis, *AES Member* (emrovithis@ionio.gr)
Nikolaos Moustakas *AES Member* (a11mous@ionio.gr)
Konstantinos Vogklis (voglis@ionio.gr)
Konstantinos Drossos (konstantinos.drossos@tuni.fi)
Andreas Floros *AES Fellow* (floros@ionio.gr)

Ionian University, Corfu, Greece *Tampere University, Tampere, Finland*

Citizen Science aims to engage people in research activities on important issues related to their well-being. Smart Cities aim to provide them with services that improve the quality of their life. Both concepts have seen significant growth in the last years, and can be further enhanced by combining their purposes with Internet of Things technologies that allow for dynamic and large-scale communication and interaction. However, exciting and retaining the interest of participants is a key factor for such initiatives. In this paper we suggest that engagement in Citizen Science projects applied on Smart Cities infrastructure can be enhanced through contextual and structural game elements realized through augmented audio interactive mechanisms. Our inter-disciplinary framework is described through the paradigm of a collaborative bird call recognition game, in which users collect and submit audio data, which are then classified and used for augmenting physical space. We discuss the Playful Learning, Internet of Audio Things, and Bird Monitoring principles that shaped the design of our paradigm, and analyze the design issues of its potential technical implementation.

0 INTRODUCTION

The concept of Smart Cities (SC) describes urban environments enriched with interaction modalities towards the improvement of city functioning and of its inhabitants' life [1]. Aiming to align the technological attainments of the digital era with the urban fabric of the physical world, SC utilize Information and Communication Technologies (ICT), such as mobile devices, embedded sensors, and data collection and analysis tools, and seamlessly integrate them in traditional infrastructure. Thus, the physical environment is transformed into a dynamic source of information, an "intelligent living space", which, based on the adoption of networking advances, provides citizens with the tools, resources, and services to exploit the benefits of this data flow [2]. The realization of intelligent systems to be embedded into SC environments can be broken down into three axes: i) the Internet of Things (IoT) facilitating the inter-connectivity of physical and virtual devices through communication protocols, ii) the Internet of Services (IoS) comprised of the amalgamation of different applications into explicable services, and iii) the Internet of People (IoP) encompassing the interactions between the citizens, who are ultimately the intended users of the system [3]. In order for SC to become truly beneficial innovation ecosystems capable of finding solutions to real-world problems, the citizens need to be excited in terms of creativity and collaboration [4]. To that scope, SC require applications that facilitate large-scale participatory projects, in which emphasis will be placed on the coordination of end-users towards dealing with the targeted issues [4].

The concept of Citizen Science (CS) describes projects, in which people volunteer to contribute to a scientific enquiry by gathering and managing information. The advent of the 21st century has signaled an unprecedented growth of CS initiatives bringing scientists and the public together with the aim of raising awareness and finding solutions about social and environmental issues. Volunteers can now quickly locate a project for CS on a subject of their interest and easily join its active community, whereas ad-

*Correspondence should be addressed to Emmanuel Rovithis
K. Drossos is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 957337, project MARVEL.

vances in human-computer interaction have extended the access to such projects to groups that could not previously be reached [5]. Thus, citizens can take on a vital role in research activities that may vary from simply providing experts with the necessary information to consulting the responsible authorities or even participating in making and implementing decisions [6]. In that process, basic scientific principles must be followed, such as well designed data collection and validation methods, explicit instructions and research questions, and feedback as a reward for participation [7].

We believe that SC can host large-scale participatory CS projects in a mutually beneficial relationship, in which SC provide the technological infrastructure and CS the activities targeting the citizens' well-being. However, the design of such an endeavour must address important challenges related to participants' engagement and cognition. Regarding the former, volunteer dropout has been identified as one of the salient factors that impact the organizational consistency of CS [6, 8]. Regarding the latter, undertaking and accomplishing tasks in a CS context does not necessarily extend beyond the acquisition of knowledge to a deeper understanding of the scientific process [9]. So far, suggestions for exciting and retaining the interest of CS volunteers include providing positive reinforcement, and matching the tasks to their personal skills [6]. Cognitive impact has been approached mostly through student assessments in curriculum-based projects presenting limited evidence that participation enhances scientific knowledge and public awareness. Therefore, more evaluations are deemed necessary for extracting solid conclusions [9]. In this paper we suggest that motivation and understanding can be enhanced through an inter-disciplinary approach that combines structural and contextual game elements with Internet of Audio Thing (IoAuT) technologies [10] to realize CS projects in SC environments. We describe our recommendations for developing an appropriate design framework through the paradigm of a bird call recognition augmented reality audio-based game.

The rest of the paper is organized as follows. In Section 1 are discussed the principles underlying the paradigm's design in terms of its playful learning, audio interaction, and bird recognition aspects, Section 2 describes the conceptual and technical structure of the paradigm and Section 3 outlines the technical design specifications. Finally, Section 4 concludes the paper.

1 DESIGN PRINCIPLES

1.1 Playful Learning

Playful Learning refers to the incorporation of game elements into non-game learning environments [11].The ability to motivate players is the most frequently cited characteristic of games related to knowledge construction [12]. By utilizing a variety of interaction mechanisms games create the conditions for competition, cooperation, exploration, and reflection, and engage participants in immersive experiences [13]. Aiming to investigate the connection of motivation and engagement to the learning outcomes researchers have intensified their efforts in the last decade, whereas educators have been drawing upon the results to systematically use game-based learning practices in their classroom [12]. Non-schooling environments have been also following this trend: museums, libraries, corporations, and government agencies have been integrating game elements in personal or collective activities as the means to enrich users' experience and enhance their construction of knowledge. However, simply adding a leaderboard system based on points of progress may have negative effects, since players with low scores could become frustrated and lose their interest in the competition [14]. Similarly, stereotypical approaches will not necessarily result in increasing and sustaining participation when addressing the broader community [15]. Therefore, careful planning of game elements integrated into non-game systems is needed to ensure motivation at all times of the process.

Large-scale participatory CS projects require attention, coordination, cooperation, and commitment. The few cases, in which CS was organized in the form of a game, have delivered positive results: providing users with a playful interface and allowing them to collaborate with or compete against each other towards a common goal resulted in users coming up with novel ideas [8]. Another approach refers to CS projects, which are embedded in the form of mini-games within larger sand-box game environments, i.e. environments, in which players have freedom of action that is not restricted by a linear narrative. In the case of [16] players completing various stages of the mini-games are rewarded with in-game prizes.

Besides the motivational function, there are other game elements that can be useful. In order for CS to produce an output of equal-to-expert quality, the participants need guidance through protocols, training, and oversight [5]. A game's rules, tutorial, and feedback can address these issues respectively, whereas the addition of a compelling narrative can enhance immersion in the experience. A final issue that we considered is the link of high motivation and engagement to the intended learning outcomes. Drawing upon modern learning theories including Problem-based Learning [17], i.e. learning from the process of striving toward the resolution of a problem, Constructivist Learning [18], i.e. learning from the process of interacting with the environment, and Experiential Learning [19], i.e. learning from the process of reflecting on one's experience, provides the theoretical basis for realizing meaningful learning environments [20]. Yet, researchers stress the need for stronger evidence, before game mechanisms aiming at motivation and immersion are systematically used as the means to achieve the learning objectives [21, 22], a need that large-scale participatory projects can address.

1.2 Internet of Audio Things

IoAuT is an emerging field that refers to embedding computing devices in physical objects towards the reception, processing, and transmission of audio information [23]. It comprises different types of audio collectors, pro-

cessors and transmitters, and facilitates their integration, local and remote accessibility, and multi-directional communication [10]. Despite the plethora of SC initiatives and the need for utilizing state of the art Human-Computer Interaction (HCI) technologies to realize new forms of participation, most approaches have focused on data visualisation techniques and mostly neglected the acoustic aspect of the urban environment [1, 24]. Existing SC applications of distributing information through the auditory channel include the generation of sound content based on urban related data in order to increase users' awareness about their city environment [24, 25], the generation of visual maps based on the perceptual attributes of submitted recordings for monitoring and managing the urban acoustic environment [26], and the augmentation of public spaces with audio information for engaging the audience in social experiences [27]. Furthermore, Wireless Acoustic Sensor Networks can be used for the surveillance and analysis of acoustic scenes, urban noise pollution, environmental anomalies, and wildlife [10].

In our paradigm design we focused on three specific aspects of IoAuT: i) collecting and submitting audio data for analysis, ii) generating a soundscape map, and iii) augmenting physical space with virtual audio components for navigation and interaction within the environment. Focusing on the latter, Augmented Reality Audio (ARA) systems have been applied for well-being purposes by acoustically enriching the working environment of employees [28], indicating the location of security threats [29], realizing non-visual spatial mappings for navigation [30], aurally signalling touristic points of interest [31], assigning audio recordings to locations of cultural importance [32], and aurally signifying city facilities for urban exploration [33]. Interaction in these implementations can be characterised as passive, i.e. users of the system essentially trigger sound events through their position and movement in the augmented space. However, more active modes of interaction can enhance users' communication in competitive or collaborative contexts. In [34] a positive connection was shown between challenging mechanics requiring the performance of gestures with the satisfaction gained from the experience. In [35] the behavior of the virtual sound sources that players need to locate is controlled by the movement of other antagonizing players, whereas in [36] players take up different roles and need to coordinate their actions in the augmented space to achieve the game goal. Sound recognition and audio based analytics [37, 38] can further expand the possibilities for interaction by advancing the responsiveness between the natural and the virtual acoustic environment [39], whereas user experience improvement techniques from the wider frame of AR can be utilized to enhance the system's context-awareness [40].

1.3 Bird Monitoring

The third field that we drew upon for designing our paradigm relates to Bird Monitoring. Bird related ecological projects usually fall into three categories: i)inventory, ii) monitoring, and iii) research [41]. Inventory projects aim to generate a list of species by identifying birds by visual observation and/or their song. Monitoring projects involve recording birds in a region or study site for a period of time. Such projects use geolocation information to pinpoint found birds on Geographic Information System (GIS) overlays. Research projects require experts to formalize and investigate a hypothesis about bird behaviour.

One of the leading active projects in collaborative Bird Monitoring is eBird, a project of the Cornell Lab of Ornithology [42, 43]. eBird evolved from a basic CS project into a collective enterprise through the novel approach of developing cooperative partnerships among experts in a wide range of fields including computer scientists, biologists, and data administrators. eBird data are overlaid on global GIS maps. They are openly available and constitute a major source of biodiversity data, increasing expert knowledge on the dynamics of bird species distributions and aiding the conservation of birds and their habitats. The project involves at the moment more that 100,000 registered users that deliver up-to-date results about bird populations. We suggest that future projects can motivate and retain participation through embedded game mechanisms as described in this paper.

2 PROPOSED FRAMEWORK

2.1 Scenario Design

In terms of structure our paradigm consists of four stages:

- In the first stage, users undertake the task of collecting bird songs for classification using the recording tool of the application. The recordings are checked internally in the mobile device regarding their authenticity and clarity, and, if they meet the criteria, they are matched to the corresponding bird species. Users are then provided with the respective information including the bird's name and photos.
- In the second stage, users submit their successful entry to the system's remote server along with the related meta-data including the recording's date and location. This information is used by the system for the creation of a virtual 2d map representing bird presence in the urban environment. Users receive a message informing that the map was updated to include their latest entry. The virtual map is dynamically shaped according to users' position in physical space and the meta-data gathered by all submissions.
- In the third stage, users activate the augmented version of the map. The system merges virtual components with physical space into an augmented environment, in which users can immerse. This augmentation relies purely on audio information. Essentially, the meta-data submitted by the users is periodically refreshed and translated into aural stimuli by shaping the parameters for the occurrence of sound events stored in the device. Each bird has been assigned its own sound, which acts as a symbol for the bird's presence, while at the same time demonstrates

its characteristic song for further study. Data sonification and sound spatialization techniques are used to express the targeted aspects of the aerial fauna: panning, playback volume, and playback rate are dynamically modified to hint at the birds' direction and proximity to the user, and at the amount of the submitted recordings respectively. Users interpret the audio information to navigate through the augmented urban landscape. Our suggested paradigm has a 24-hour storage cycle, through which the meta-data can be recalled on a day-by-day basis.

- In the fourth stage, users experience an artistic aspect of the project, in which all submitted recordings are streamed by the system to the users, aimed at exciting their enjoyment, engagement, and collaboration. This stage can be activated once users are within the range of an audio source. In case many recordings have been assigned to a specific location, the system performs processing and mixing of the material through reverberation and temporal allocation algorithms to distinguish the recordings from one another and create a clear and appealing virtual soundscape before superimposing it onto the real acoustic environment. Thus, users feel like they are participating in a collaborative artwork that enriches their scientific duties.

Our proposed paradigm suggests further enhancing the aforementioned stages through game elements applied on game scenarios. Thus, citizen scientists become players of a large-scale participatory game that aspires to boost their engagement through fun, challenge, attainment, competition, and collaboration.

Regarding game elements, the following are suggested:

- Level advance: players unlock different game scenarios according to their successful submissions and commitment to the project.
- Badges and rewards: players' progress is also made public through titles to be gained as rewards for their performance.
- Collaborative mode: players can work together towards achieving more complex goals that require cooperation and coordination with each other.

Regarding game scenarios, the following are suggested:

- Quiz: players are asked to recognize the bird songs to gain points for their correct answers. They can familiarize themselves with the topic by studying the stored patterns in the device, a process which enhances the project's educational aspect.
- Treasure Hunt: players focus on a specific bird and report different locations of its presence within a limited time frame. They can consult their maps (2d or augmented), as they are dynamically shaped by the submissions of other players.
- Time Travel: players follow the route of a specific bird within a certain time period by visiting past corresponding locations that are saved in the system, and thus study its migratory mobility.
- Adopt a bird: players monitor the activity of a specific bird for a certain amount of consecutive days. In the process more information about the species is disclosed, such as feeding and mating habits. This game scenario aims to engage players more deeply with caring for the subject of their study.

In terms of audio interaction, all aforementioned stages, elements, and scenarios rely on three different modes:

- Constant Listening: players are exposed to the complete augmented audio environment at any time, and can also select to isolate specific information.
- Focused Listening: players turn their device like opening a window to a specific direction and only listen to the audio information that the device is facing.
- Interactive Listening: players perform specific actions with the device, such as pressing a virtual button to record or tilting their device to activate the virtual soundscape and/or bring front specialized information.

2.2 Architecture Design

The design of our paradigm's architecture (fig. 1) is based on the four stages defined in the scenario, and involves a three-layer IoAuT setup. The Sensing Layer includes the sensors, and the recording and playback module in the user's mobile device, which allow for producing audio content and analyzing phenomena associated with auditory events. The Network Layer is responsible for data transfer from the Sensing Layer, and the Application Layer includes the web services and the virtual soundscape construction module.

Focusing on exploring the ARA environment, the architecture design segments the concept into two primary modes. The first mode is designed to facilitate passive interaction, as users walk through the real environment. After the desired filters are set through the menu of the mobile app, the sound monitoring mechanism reproduces the real acoustic environment in real-time, and the playback mechanism delivers the captured sound, when its source is in the user's proximity. All audio components are mixed together and delivered through the audio headset. An amplification coefficient, which is adjustable by the user, is applied to audio capture for improving the recognition of bird sound. Once users hear something of interest, they can enter the second mode and actively search the dynamically generated virtual 2d map or augmented audio map to locate points of interest and interact with them.

Once the preview mechanism is derived using the above procedure, audio recording is implemented into the existing capture procedure model. The user then responds to this procedure by annotating the part of the waveform, which contains the bird sound. A Convolutional Neural Network (CNN) model embedded in the mobile device checks to recognize specific bird classes. If the model classifies the annotating sound to a specific class, then the local save pro-

cedure is enabled. More specifically, the following data are stored: i) the annotation of the audio file, ii) the tag of the bird class, iii) the tag of the time of the event, including date, year, and hour, and iv) the GPS coordinates that are captured using the GPS features of the mobile device. As soon as the local saved data are ready, the final upload procedure to the web server can be made.

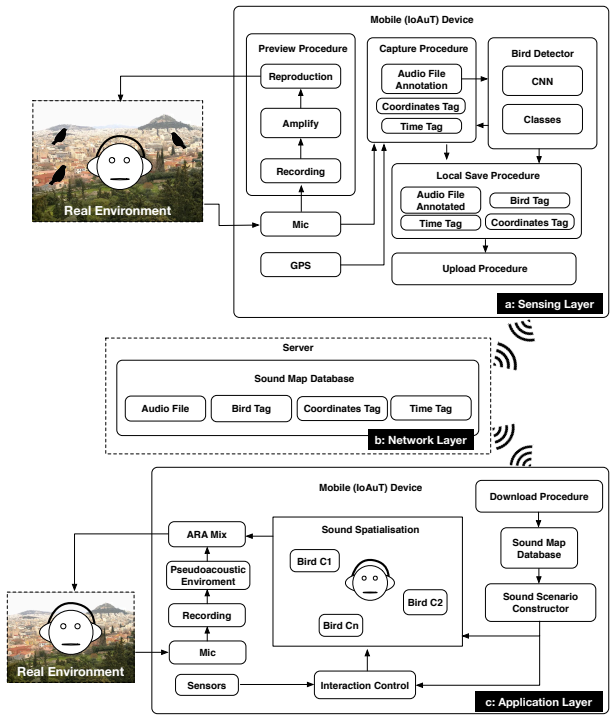


Fig. 1: The Concept Architecture: a) Sensing Layer, b) Network Layer, c) Application Layer

The Application Layer is organized in a client-server architecture. The mobile app, as the client, requests to download a sound scene by sending the phone position. The server part manages the data regarding the audio files of the annotated birds with corresponding tags of bird classification, GPS coordinates and time. The sound scene data are imported in the sound scenario constructor, which manages the processes of gestural interaction control, and the sound spatialization engine, which is responsible for placing the recordings in virtual space. The variations of sound scenario constructor and interaction controller shape each game mode.

2.3 Risk Assessment

Our paradigm stands for a preliminary approach that requires implementation and testing for evaluation and optimization. As a first step we have performed an assessment of potential risks and we suggest ways that they could be dealt with.

- Bad data: users could submit sounds that have been downloaded from the web, or recorded from other prerecorded playback. This risk can be countered by a) allowing users to make a recording only via the in-app recording tool, and b) performing a frequency range check to establish that the audio captured is a result of natural wild-life recorded in situ and not elsewhere. Special techniques used in voice anti-spoofing (see results from `https://www.asvspoof.org/`) can be also applied.
- Wildlife disturbance: naive users might disturb the birds' natural habitat in their attempt to perform the game actions. The fact that the proposed scenarios rely on capturing audio information, a process which does not require visual contact with the subject under examination, reduces that risk.
- Acoustic interference: in connection to the previous risk, there is the possibility that the playback of the application's audio content interferes with the natural acoustic environment and its inhabitants. This risk can be eliminated, if the application works only in headphones mode and does not emit sound from the speakers.
- Data overload: the streaming of too much information might cause system lag. Towards reducing that risk, the sounds used for user's navigation will be stored in the device, which will receive from the server only playback specifications. Furthermore, the streaming soundscape made from the all users' submissions, will be created periodically on the server. Each user entering the same physical area would download the same soundscape audio.

We understand that wildlife disturbance is a complex and sensitive issue and poses a major challenge in the gamification of the bird recognition process. Careful design must be applied to ensure that a set of appropriate instructions regarding user behavior is clearly communicated without thwarting the application's game aspect. Furthermore, an expert evaluation [44] is intended to take place and provide valuable insight towards the prevention of possible negative consequences.

3 TECHNICAL DESIGN SPECIFICATIONS

3.1 Audio Capture

The capture of the real acoustic environment is done by means of the sound recording features of the device. The user defines the appropriate recording option according to their equipment. The available options include mono and binaural recording technology. The capture procedure also involves outputting the monitor audio to the default playback device. Mono recording uses modules that are typically available in almost all modern smartphones: the built-in microphone for detecting sounds of interest, and a set of headphones as playback acoustic equipment. The binaural recording option is performed using in-ear microphones embedded on a stereo headset like Sennheiser Ambeo Smart Headset[1]. In both options, there is an optional

[1] Sennheiser AMBEO Smart Headset-Mobile binaural recording headset, URL `https://en-de.sennheiser.com/in-ear-headphones-3d-audio-ambeo-smart-headset`, (Accessed on 03/19/2021)

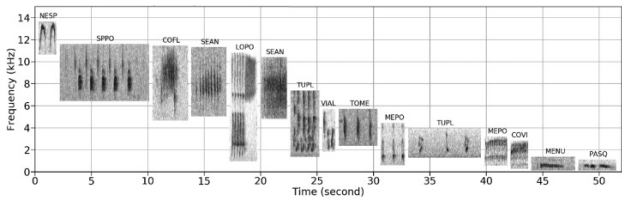


Fig. 2: Spectrograms of 14 distinct tropical birds from Puerto Rico. Taken from [45]

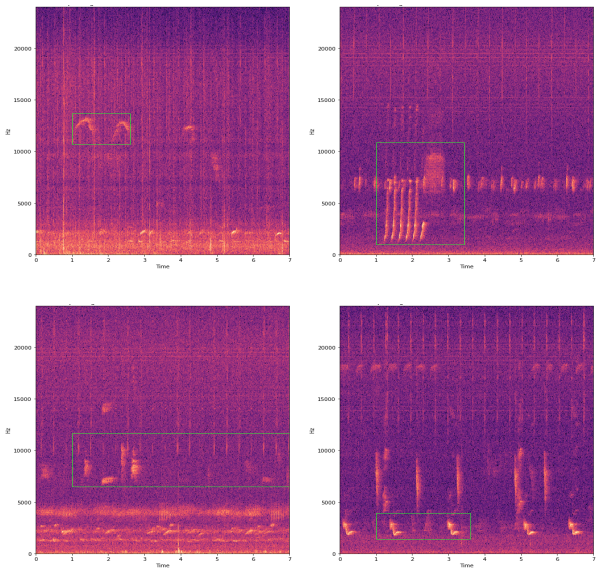


Fig. 3: Time-frequency representation of 4 distinct bird sounds from [45], with an annotation bounding box created by an expert

gain control of the environmental monitoring system for personalized sound detection procedure (fig.1).

3.2 Bird Call classification

Automatic bird sound classification plays an important role in monitoring and protecting biodiversity. Recent advances in machine listening and deep learning models for bird audio detection provide a novel way for improving bird call recognition to expert level. The fact that interspecies bird sounds exhibit such a distinctive spectral structure, motivated researches to employ typical hand-crafted time-frequency representations as an input to various deep learning models. The most prominent one is the usage of mel-scaled band energies as the time-frequency representation, which is given as an input to 2D CNNs [45]. Figure 2 shows the distinctive spectral structure of 14 tropical birds.

The most recent example of a Bird audio classification model is BirdNet, introduced by Cornell University [46, 47]. It involves a ResNet-like CNN model, containing 127 layers and 27 million parameters, and capable of identifying 984 North American and European bird species by sound. BirdNet achieves a Mean Average Precision (mAP) of 0.791 for single-species recordings. However, typical CNNs cannot model long temporal dependencies that are usually needed in machine listening tasks, such as bird-audio detection [48, 37]. To address this issue, different published papers have adopted the Convolutional Recurrent Neural Network (CRNN) model [48, 49]. The CRNN model consists of a series of 2D CNNs, followed by Recurrent Neural Networks (RNNs) and a linear layer. A time-frequency representation of audio is given as an input to the a CRNN model, and the 2D CNNs learn time-frequency patterns according to the targeted task (e.g. bird audio detection). Then, the RNNs take as input the output of the CNNs, and focus on learning temporal patterns. Finally, the linear layer is fed the output of the RNNs and performs the classification. The CRNN model achieved high performance in DCASE Challenge tasks on bird-audio detection, ranking among the top 5 systems [48, 50].

3.2.1 Deep learning workflow

In order to implement the bird call classification module the typical workflow for training image deep learning models will be followed. The first important step is to record the primary data sources, while keeping extensive metadata about time and location and recording quality. Sound data must be carefully curated by experts to reflect also the background sounds of the field under surveillance. The next steps comprise standard Deep Learning pipelines. CNNs are applied to classify time-frequency visualization of bird sounds (see Figure 3), using a hold out validation set to control the generalization to unknown cases. There are two major modeling options: i) clip-wise annotation/inference, where the model classifies a bunch of seconds at once, and ii) frame-wise annotation/inference, where the model outputs prediction for each discrete time step (e.g. milliseconds).

3.2.2 Delivering the model

A trained CNN model is delivered to the proposed application in two ways: i) recognition as a service where a back-end server (powered by GPU) is used to accept audio chunks, perform the pre-processing and the inference, and report the classification result and ii) recognition on the edge-device, where all process is performed on the user side. Each serving paradigm has pros and cons. Recognition as a service needs network access and a powerful back-end server, whereas recognition on the edge can accommodate relatively small sized models and has no network requirements.

3.3 GIS based repository

Since bird call detection will be accompanied by GPS based geographical coordinate and time stamp, all detection data can be presented in the form of cartographic GIS data model. This will allow users to easily locate their findings and the findings of others. GIS data can be aggregated by zoom levels and give finer grain detection the further the user zooms in. The basic user interface screen could be a map with overlaid information about the user's current geo-location and already existing bird call detection. A side effect of this online collaborative GIS repository

would be the extraction of migratory journeys of species of bird as well as the abundance of specific species.

3.4 Augmented Reality Audio

The ARA environment consists of a real and a virtual acoustic component with the real sound recording being mixed with the spatial reproduction of the classified captured birds' sounds into a pseudoacoustic environment, a mix that needs to take place as seemlessly as possible. The importance of a dynamic ARA mix of the gain difference between the real and the virtual acoustic environment compared to the static mix gain in legacy ARA mix models [51] has been pointed out in [52]. The comparison of the legacy and the adaptive ARA mix model has shown that the latter demonstrated significantly better performance in terms of auditory perception [53]. Thus, in the proposed framework, we employ this dynamic and adaptive ARA mixing strategy that focuses on the impact of dynamic fluctuations of the real and the virtual environment to acoustic perception, taking in consideration acoustic phenomena, such as auditory masking.

Furthermore, the location awareness in ARA systems refers to the capability of a device to determine its location in terms of coordinates through active or passive human-computer interaction. Several ARA works have shown the necessity to utilize spatialization techniques, in order to combine data extracted from location awareness systems with virtual sound sources [28, 54]. Our proposed system includes a spatialization module for positioning the 3d virtual sound sources, a set of sensors including gyroscope, accelerometer, and GPS, that facilitates gestural interaction, and a headset for reproducing the augmented acoustic environment. With this setup users are free to move their head in both horizontal and median plane, and listen to the entire 3d acoustic space, while transmitting their location, movement, and gestural activity to the system's engine.

4 CONCLUSION

We have presented a framework for enriching a Citizen Science project in a Smart City environment with game elements using Internet of Sound technologies. The aim of our inter-disciplinary approach is to seek ways to enhance the public's motivation for participation, engagement in the experience, and deeper understanding of the subjects and processes at hand. We focused on Bird Call Recognition and Monitoring collaborative activities, in which participants can report and study the state and flux of the urban aerial ecosystem in the context of playful scenarios. In accordance with modern learning theories we propose the use of game elements including targeted quests, structured levels, and progress points and badges. User interaction relies on Internet of Audio Things mechanisms including recording and submitting audio data, and exploring the augmented environment through GIS-related navigation and gestural performance with the mobile device. Sound classification takes place through a CNN network, and the augmented soundscape is constructed by an adaptive ARA mixing system.

A core aspect of CS and SC is to focus on citizens' problems and needs. The users of our proposed system are seen, on the one hand, as active units that exploit the benefits of the enhanced world around them, and, on the other, as interconnected members of the community that collaborate with each other to improve that world. SC facilitate a safe environment to observe, reflect, and experiment, whereas CS provides with specific problems to solve and thus contribute to the scientific community. We hope that our proposed framework will serve as future reference towards enhancing the appeal of CS to the involved stakeholders, and providing novel ways to realize personal and collective interactive experiences based on a network of audio devices able to collect, evaluate, process and distribute acoustic data in urban environments.

5 REFERENCES

[1] M. Batty, K. W. Axhausen, F. Giannotti, A. Pozdnoukhov, A. Bazzani, M. Wachowicz, G. Ouzounis, Y. Portugali, "Smart cities of the future," *The European Physical Journal Special Topics*, vol. 214, no. 1, pp. 481–518 (2012), doi:10.1140/EPJST/E2012-01703-3.

[2] A. Urbieta, A. González-Beltrán, S. B. Mokhtar, M. A. Hossain, L. Capra, "Adaptive and context-aware service composition for IoT-based smart cities," *Future Generation Computer Systems*, vol. 76, pp. 262–274 (2017), doi:10.1016/j.future.2016.12.038.

[3] J. M. Hernández-Muñoz, J. B. Vercher, L. Muñoz, J. A. Galache, M. Presser, L. A. H. Gómez, J. Pettersson, "Smart cities at the forefront of the future internet." presented at the *Future internet assembly*, pp. 447–462 (2011), doi:10.1007/978-3-642-20898-0_32.

[4] H. Schaffers, N. Komninos, M. Pallot, B. Trousse, M. Nilsson, A. Oliveira, "Smart cities and the future internet: Towards cooperation frameworks for open innovation," presented at the *The future internet assembly*, pp. 431–446 (2011), doi:10.1007/978-3-642-20898-0_31.

[5] R. Bonney, J. L. Shirk, T. B. Phillips, A. Wiggins, H. L. Ballard, A. J. Miller-Rushing, J. K. Parrish, "Next steps for citizen science," *Science*, vol. 343, no. 6178, pp. 1436–1437 (2014), doi:10.1126/science.1251554.

[6] C. C. Conrad, K. G. Hilchey, "A review of citizen science and community-based environmental monitoring: issues and opportunities," *Environmental monitoring and assessment*, vol. 176, no. 1, pp. 273–291 (2011), doi:10.1007/s10661-010-1582-5.

[7] J. Silvertown, "A new dawn for citizen science," *Trends in ecology & evolution*, vol. 24, no. 9, pp. 467–471 (2009), doi:10.1016/j.tree.2009.03.017.

[8] E. Hand, "Citizen science: People power," *Nature News*, vol. 466, no. 7307, pp. 685–687 (2010), doi:10.1038/466685a.

[9] R. Bonney, T. B. Phillips, H. L. Ballard, J. W. Enck, "Can citizen science enhance public understanding of science?" *Public Understanding of Science*, vol. 25, no. 1, pp. 2–16 (2016), doi:10.1177/0963662515607406.

[10] L. Turchet, G. Fazekas, M. Lagrange, H. S. Ghadikolaei, C. Fischione, "The Internet of Audio Things: State of the Art, Vision, and Challenges," *IEEE internet of things journal*, vol. 7, no. 10, pp. 10233–10249 (2020), doi:10.1109/JIOT.2020.2997047.

[11] J. L. Plass, B. D. Homer, C. K. Kinzer, "Foundations of game-based learning," *Educational Psychologist*, vol. 50, no. 4, pp. 258–283 (2015), doi:10.1080/00461520.2015.1122533.

[12] C. Costa, K. Tyner, S. Henriques, C. P. G. Sousa, "A Review of Research Questions, Theories and Methodologies for Game-Based Learning," *Journal of Content, Community and Communication*, vol. 4 (2016).

[13] D. Dicheva, C. Dichev, G. Agre, G. Angelova, "Gamification in education: A systematic mapping study," *Journal of Educational Technology & Society*, vol. 18, no. 3, pp. 75–88 (2015).

[14] T. Reiners, L. C. Wood, J. Dron, "From chaos towards sense: A learner-centric narrative virtual learning space," in *Gamification for human factors integration: Social, education, and psychological issues*, pp. 242–258 (IGI Global) (2014), doi:10.4018/978-1-4666-5071-8.CH015.

[15] S.-K. Thiel, M. Reisinger, K. Röderer, P. Fröhlich, "Playing (with) democracy: A review of gamified participation approaches," *JeDEM-eJournal of eDemocracy and Open Government*, vol. 8, no. 3, pp. 32–60 (2016), doi:10.29379/JEDEM.V8I3.440.

[16] CCP, "Project Discovery: citizen science begins with you," (2020), URL `https://www.eveonline.com/discovery`.

[17] D. Boud, G. Feletti, *The challenge of problem-based learning* (Psychology Press) (1997).

[18] T. M. Duffy, D. H. Jonassen, "Constructivism: New implications for instructional technology?" *Educational Technology*, vol. 31, no. 5, pp. 7–12 (1991 May).

[19] D. A. Kolb, *Experiential learning: Experience as the source of learning and development* (FT press) (2014).

[20] K. Kiili, *On Educational Game Design: Building Blocks of Flow Experience*, Tampere University of Technology. Publication (Tampere University of Technology) (2005 Dec.), awarding institution:Tampere University of Technology¡br/¿Submitter:Made available in DSpace on 2008-09-24T10:58:32Z (GMT). No. of bitstreams: 1 kiili.pdf: 2269572 bytes, checksum: a198e0fd0dec407e8b1f9aecfdeb1fb9 (MD5) Previous issue date: 2005-12-14.

[21] C. Perrotta, G. Featherstone, H. Aston, E. Houghton, "Game-based learning: Latest evidence and future directions," *Slough: NFER* (2013).

[22] J. Hamari, D. J. Shernoff, E. Rowe, B. Coller, J. Asbell-Clarke, T. Edwards, "Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning," *Computers in human behavior*, vol. 54, pp. 170–179 (2016), doi:0.1016/J.CHB.2015.07.045.

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

[24] P. Sarmento, O. Holmqvist, M. Barthet, "Musical Smart City: Perspectives on Ubiquitous Sonification," *arXiv preprint arXiv:2006.12305* (2020).

[25] K. Drossos, A. Floros, N.-G. Kanellopoulos, "Affective Acoustic Ecology: Towards Emotionally Enhanced Sound Events," presented at the *Proceedings of the 7th Audio Mostly Conference: A Conference on Interaction with Sound*, AM '12, pp. 109–116 (2012), doi:10.1145/2371456.2371474, URL `http://doi.acm.org/10.1145/2371456.2371474`.

[26] J. Kang, F. Aletta, E. Margaritis, M. Yang, "A model for implementing soundscape maps in smart cities," *Noise Mapping*, vol. 5, no. 1, pp. 46–59 (2018), doi:10.1515/noise-2018-0004.

[27] P. K. Nikolic, H. Yang, "Designing playful cities: audio-visual metaphors for new urban environment experience," *Mobile Networks and Applications*, pp. 1–7 (2020), doi:10.1007/s11036-020-01514-6.

[28] E. D. Mynatt, M. Back, R. Want, R. Frederick, "Audio Aura: Light-weight audio augmented reality," presented at the *Proceedings of the 10th annual ACM symposium on User interface software and technology*, pp. 211–212 (1997), doi:10.1145/263407.264218.

[29] V. Sundareswaran, K. Wang, S. Chen, R. Behringer, J. McGee, C. Tam, P. Zahorik, "3D audio augmented reality: implementation and experiments," presented at the *The Second IEEE and ACM International Symposium on Mixed and Augmented Reality, 2003. Proceedings.*, pp. 296–297 (2003), doi:10.1109/ISMAR.2003.1240728.

[30] S. Holland, D. R. Morse, H. Gedenryd, "AudioGPS: Spatial audio navigation with a minimal attention interface," *Personal and Ubiquitous computing*, vol. 6, no. 4, pp. 253–259 (2002), doi:10.1007/s007790200025.

[31] D. McGookin, S. Brewster, P. Priego, "Audio bubbles: Employing non-speech audio to support tourist wayfinding," presented at the *International Conference on Haptic and Audio Interaction Design*, pp. 41–50 (2009), doi:10.1007/978-3-642-04076-4_5.

[32] J. Reid, E. Geelhoed, R. Hull, K. Cater, B. Clayton, "Parallel worlds: immersion in location-based experiences," presented at the *CHI'05 extended abstracts on Human factors in computing systems*, pp. 1733–1736 (2005), doi:10.1145/1056808.1057009.

[33] J. R. Blum, M. Bouchard, J. R. Cooperstock, "What's around me? Spatialized audio augmented reality for blind users with a smartphone," presented at the *International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services*, pp. 49–62 (2011), doi:10.1007/978-3-642-30973-1_5.

[34] E. Rovithis, N. Moustakas, A. Floros, K. Vogklis, "Audio Legends: Investigating Sonic Interaction in an Augmented Reality Audio Game," *Multimodal Technologies and Interaction*, vol. 3, no. 4, p. 73 (2019), doi:10.3390/mti3040073.

[35] N. Moustakas, A. Floros, N. Grigoriou, "Interactive audio realities: An augmented/mixed reality audio game

prototype," presented at the *Audio Engineering Society Convention 130* (2011 may).

[36] R. Pellerin, N. Bouillot, T. Pietkiewicz, M. Wozniewski, Z. Settel, E. Gressier-Soudan, J. R. Cooperstock, "Soundpark: Exploring ubiquitous computing through a mixed reality multi-player game experiment," *Studia Informatica Universalis*, vol. 8, no. 3, p. 21 (2009).

[37] K. Drossos, S. I. Mimilakis, S. Gharib, Y. Li, T. Virtanen, "Sound Event Detection with Depthwise Separable and Dilated Convolutions," presented at the *2020 International Joint Conference on Neural Networks (IJCNN)* (2020 Jul.), doi:10.1109/IJCNN48605.2020.9207532.

[38] K. Drossos, S. Lipping, T. Virtanen, "Clotho: An Audio Captioning Dataset," presented at the *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2020 May), doi: 10.1109/ICASSP40776.2020.9052990.

[39] V. Pulkki, S. Delikaris-Manias, A. Politis, *Parametric time-frequency domain spatial audio* (Wiley Online Library) (2018 October), doi:10.1002/9781119252634.

[40] J. Aliprantis, G. Caridakis, "A survey of augmented reality applications in cultural heritage," *International Journal of Computational Methods in Heritage Science (IJCMHS)*, vol. 3, no. 2, pp. 118–147 (2019), doi: 10.4018/IJCMHS.2019070107.

[41] M. E. Hostetler, "Bird Monitoring Projects for Youth: Leader's Guide1," (2002 October).

[42] B. L. Sullivan, C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, S. Kelling, "eBird: A citizen-based bird observation network in the biological sciences," *Biological conservation*, vol. 142, no. 10, pp. 2282–2292 (2009), doi: 10.1016/J.BIOCON.2009.05.006.

[43] B. L. Sullivan, J. L. Aycrigg, J. H. Barry, R. E. Bonney, N. Bruns, C. B. Cooper, T. Damoulas, A. A. Dhondt, T. Dietterich, A. Farnsworth, *et al.*, "The eBird enterprise: an integrated approach to development and application of citizen science," *Biological Conservation*, vol. 169, pp. 31–40 (2014), doi:10.1016/J.BIOCON.2013.11.003.

[44] H. Petrie, N. Bevan, "The Evaluation of Accessibility, Usability, and User Experience." *The universal access handbook*, vol. 1, pp. 1–16 (2009), doi:10.1201/9781420064995-c20.

[45] J. LeBien, M. Zhong, M. Campos-Cerqueira, J. P. Velev, R. Dodhia, J. L. Ferres, T. M. Aide, "A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network," *Ecological Informatics*, vol. 59, p. 101113 (2020), doi: https://doi.org/10.1016/j.ecoinf.2020.101113.

[46] M. Arif, R. Hedley, E. Bayne, "Testing the Accuracy of a birdNET, Automatic bird song Classifier," *ERA - University of Alberta's open access digital archive* (2020 July), doi:10.7939/R3-6KHB-KZ18.

[47] S. Kahl, C. M. Wood, M. Eibl, H. Klinck, "BirdNET: A deep learning solution for avian diversity monitoring," *Ecological Informatics*, vol. 61, p. 101236 (2021), doi:10.1016/j.ecoinf.2021.101236.

[48] E. Cakir, S. Adavanne, G. Parascandolo, K. Drossos, T. Virtanen, "Convolutional recurrent neural networks for bird audio detection," presented at the *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 1744–1748 (2017), doi:10.23919/EUSIPCO.2017.8081508.

[49] E. Çakir, G. Parascandolo, T. Heittola, H. Huttunen, T. Virtanen, "Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1291–1303 (2017), doi:10.1109/TASLP.2017.2690575.

[50] S. Adavanne, K. Drossos, E. Çakir, T. Virtanen, "Stacked convolutional and recurrent neural networks for bird audio detection," presented at the *2017 25th European signal processing conference (EUSIPCO)*, pp. 1729–1733 (2017), doi:10.23919/EUSIPCO.2017.8081505.

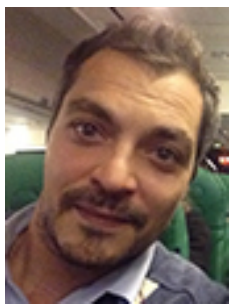
[51] M. Karjalainen, V. Riikonen, M. Tikander, "An augmented reality audio mixer and equalizer," presented at the *Audio Engineering Society Convention 124* (2008 May).

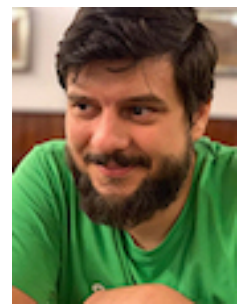
[52] N. Moustakas, A. Floros, E. Rovithis, K. Vogklis, "Augmented audio-only games: A new generation of immersive acoustic environments through advanced mixing," presented at the *Audio Engineering Society Convention 146* (2019 March).

[53] N. Moustakas, E. Rovithis, K. Vogklis, A. Floros, "Adaptive Audio Mixing for Enhancing Immersion in Augmented Reality Audio Games," presented at the *Companion Publication of the 2020 International Conference on Multimodal Interaction*, pp. 220–227 (2020), doi:10.1145/3395035.3425325.

[54] Y. Vazquez-Alvarez, I. Oakley, S. A. Brewster, "Auditory display design for exploration in mobile audio-augmented reality," *Personal and Ubiquitous computing*, vol. 16, no. 8, pp. 987–999 (2012), doi: 10.1007/s00779-011-0459-0.

THE AUTHORS




Dr. Emmanouel Rovithis was born in Athens, Greece in 1978. He holds an MA in Music Composition from the Anglia Polytechnic University in Cambridge, UK., and a PhD (first-class honours) in Electronic Music Composition from the Department of Music Studies of the Ionian University in Corfu, Greece. He is currently employed as Laboratory Research and Teaching Staff at the Department of Audio & Visual Arts of the Ionian University, teaching subjects related to Art and Technology in Education, Digital Signal Processing, and Music Programming. His research focuses on the design of Audio Games, and Augmented Reality Audio applications for educational, entertaining, and creative purposes. He has designed numerous interactive installations, educational games and software, seminars and workshops on behalf of prominent cultural institutions, and has a strong background in music composition and sound design for theatre and cinema.

●

Nikolaos Moustakas was born in Piraeus, Greece in 1987. He received his Master Degree in Audiovisual Arts from the department of Audio & Visual Arts, Ionian University in 2011, and in 2021 his PhD from the same department. His research focus on digital audio signal processing, adapting techniques, auditory perception and mixed reality environments. He was involved in the development of augmented reality audio technologies and audio-only games.

●

Konstantinos Vogklis was born in Ioannina, Greece, in 1978. He received the diploma degree in Computer Science from the University of Ioannina, Greece, in 1999, and M.Sc. and Ph.D degrees in computer science, in 2002 and 2010 respectively, from the same department. His research interests include the development of high-performance parallel algorithms and machine learning techniques with emphasis on big data, image and sound processing. In the last couple of years he has been involved in the design and implementation of augmented reality applications on mobile devices aiming on both entertainment and the emergence of cultural heritage. His published work includes 30 papers in peer-reviewed scientific journals and conferences.

●

Dr. Konstantinos Drossos was born in Thessaloniki, Greece. He holds a BEng in Sound Technology (first-class honours), a BSc in Informatics, an MSc in Sound & Vibration Research, and a PhD (first-class honours) in the field of machine listening. Currently, he is a senior researcher at the Audio Research Group (ARG), Finland. He has been a postdoc researcher at ARG and a postdoc fellow at Montreal Institute for Learning Algorithms, Canada, and at Music Technology Group, Spain. He has authored over 45 research papers, has pioneered the field of audio captioning, has organized scientific challenges, special sessions, and workshops in international conferences, serves as a reviewer for top journals and conferences, and has served as the Chairman of the Finnish IEEE Joint Chapter of SP&CAS. His research interests include audio captioning, domain adaptation, multimodal translation, source separation, detection and classification of acoustic scenes and events, and machine listening.

●

Andreas Floros was born in Drama, Greece in 1973. In 1996 he received his engineering degree from the department of electrical and computer engineering, University of Patras, and in 2001 his Ph.D. degree from the same department. His research was mainly focused on digital audio signal processing and conversion techniques for all-digital power amplification methods. He was also involved in research in the area of acoustics. H serves as Professor of Audio Technology and Electroacoustics at the department of Audiovisual Arts, Ionian University. His current research interests focus on Analysis, processing and conversion of digital audio signals, intelligent digital audio effects and sound synthesis, creative intelligence, audio-only games, auditory interfaces and displays, augmented reality audio foundations and applications as well as the investigation of the impact of sound events to human emotions.