




Article

Semantic Crowdsourcing of Soundscapes Heritage: A Mojo Model for Data-Driven Storytelling

Marina Eirini Stamatiadou *, Iordanis Thoidis , Nikolaos Vryzas, Lazaros Vrysis 
and Charalampos Dimoulas * 

Multidisciplinary Media & Mediated Communication Research Group (M3C), Aristotle University,
54636 Thessaloniki, Greece; ithoidis@auth.gr (I.T.); nvryzas@auth.gr (N.V.); lvrysis@auth.gr (L.V.)

* Correspondence: mstamat@auth.gr (M.E.S.); babis@eng.auth.gr (C.D.); Tel.: +30-2310-994245 (C.D.)

Abstract: The current paper focuses on the development of an enhanced Mobile Journalism (MoJo) model for soundscape heritage crowdsourcing, data-driven storytelling, and management in the era of big data and the semantic web. Soundscapes and environmental sound semantics have a great impact on cultural heritage, also affecting the quality of human life, from multiple perspectives. In this view, context- and location-aware mobile services can be combined with state-of-the-art machine and deep learning approaches to offer multilevel semantic analysis monitoring of sound-related heritage. The targeted utilities can offer new insights toward sustainable growth of both urban and rural areas. Much emphasis is also put on the multimodal preservation and auralization of special soundscape areas and open ancient theaters with remarkable acoustic behavior, representing important cultural artifacts. For this purpose, a pervasive computing architecture is deployed and investigated, utilizing both client- and cloud-wise semantic analysis services, to implement and evaluate the envisioned MoJo methodology. Elaborating on previous/baseline MoJo tools, research hypotheses and questions are stated and put to test as part of the human-centered application design and development process. In this setting, primary algorithmic backend services on sound semantics are implemented and thoroughly validated, providing a convincing proof of concept of the proposed model.

Keywords: soundscapes; audiovisual heritage; semantic audio; data-driven storytelling; cultural heritage; content crowdsourcing; heritage management



Citation: Stamatiadou, M.E.; Thoidis, I.; Vryzas, N.; Vrysis, L.; Dimoulas, C. Semantic Crowdsourcing of Soundscapes Heritage: A Mojo Model for Data-Driven Storytelling. *Sustainability* **2021**, *13*, 2714. <https://doi.org/10.3390/su13052714>

Academic Editors: Asterios Bakolas
and Marc A Rosen

Received: 12 December 2020

Accepted: 25 February 2021

Published: 3 March 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Cultural Heritage (CH) is considered very important from multiple perspectives of everyday modern human life, including but not limited to education, history, cultivation of cultural awareness, social engagement, entertainment, and well-being. The proliferation of Information and Communication Technologies (ICTs) and especially digital mobile devices has significantly propelled CH projects and associated featured services (websites, multimedia/mobile apps, etc.). In this context, ordinary users can navigate and virtually visit places and artifacts displaying cultural and heritage interests, literately, without time or geographical restrictions. These services can be deployed at the change of attending a physical environment with cultural value for augmenting the whole experience (before, during, and after the visit) or general infotainment activities. Apart from the cases of digital museums and exhibitions concerning artworks, historical buildings, monuments, and other cultural items, intangible CH has flourished through the processes of information capturing, documentation, and digital synthesis of CH storytelling experiences [1–7].

Among others, the audiovisual heritage associated with places, performances, and events can benefit from this progress in recording, managing, and authoring data-driven narratives [5–9]. In this context, average users can become active participants in the processes of contributing and exploiting multimedia content by experiencing, evaluating,

and reinforcing the associated services. For instance, previous works have proved that applicable media assets can be quickly and massively crowdsourced, making use of the inherent audiovisual capturing and networking capabilities that modern mobile devices offer [10–12]. Apart from the data themselves, useful context-, time-, and location-aware meta-data can be extracted to facilitate semantic information management and retrieval [13–16]. Through social tagging, it is possible to gather information about emotionally pleasant or unpleasant sounds in different urban areas [17]. However, as discussed in [10], not many ICT tools and/or services have been developed to support people in contributing audiovisual data, assisting toward the design of a CH framework.

Environments, either physical or artificial, bring together their own acoustic profiles. Distinct sound languages can shape a recognizable identity offering an individual experience to the human's sound perception [18]. The concept of the soundscape was introduced as early as 1977 by R. Murray Schafer, making the first attempts to describe what exactly a human ear hears or listens to, when in a particular and self-explained environment [19]. It was in 2008 when the International Organization for Standardization (ISO) established the working group ISO/TC 43/SC1/WG 54 "Perceptual assessment of soundscape quality." The objective of this group was to assist and promote consistency and compatibility between both theoretical and methodological approaches of soundscape studies and practice, developing the following definition, as given in ISO 12913-1, Section 2.3 [20]:

"Soundscape is an acoustic environment as perceived or experienced and/or understood by a person or people, in context."

Therefore, when discussing soundscapes heritage, the key issue is to focus not only on the meaning of sounds, but on their implicit impact on the everyday quality of life and the opportunity to promote genuine acoustic sustainability. Besides, the interdisciplinary field of soundscape studying and research also lies in the conservation of acoustic heritages [21,22].

Data-driven storytelling is related to the way of making stories through data, i.e., the captured audiovisual content and its associated semantic metadata. In this perspective, possible multisite monitoring (offered by multiple mobile users) can be deployed, offering the option of selecting and/or augmenting the preferred viewpoint/reproduction configuration [14]. This feature makes a good match to the empirical and strongly personalized aspects of perceiving soundscapes, opposed to the somewhat neutral/impersonal acoustic environment capturing and reproduction [18–21]. Hence, the idea is to engage the audience for sound-related CH capturing and semantic description, thus forming a mediated way of experiencing soundscapes. Apparently, there are multiple aspects that can be assembled in this direction, encompassing all spatiotemporal, acoustic, visual, and semantic levels at the reproduction site. Nonetheless, the main goal here is to attract mobile users for collecting and contributing semantically enhanced media assets (i.e., audiovisual records with their pattern-related metadata), equipping them with the necessary Machine Learning (ML) capabilities for on-site sound detection and classification [15,16]. Such mobile applications would allow the description of the associated scenes and sound-fields (both aurally and visually), and to share the soundscape experience as intangible CH storytelling. This notion of soundscapes, which is perceived by the captured content, the offered retrieval/reproduction, and the associated sound (and video) semantics, will be considered throughout the rest of the paper.

The current work focuses on the collaborative collection and documentation of soundscapes and environmental sound semantics, which apart from CH, also significantly impact human life quality in multiple perspectives (as explained in the next sections). The whole approach has many similarities with sophisticated Mobile Journalism (MoJo) services, helping professional and citizen journalists collect news-items and shape them into featured data-driven storytelling [23,24]. Relying on the so-called *MoJo-mate* platform (Mobile Journalism Machine-Assisted Reporting) [23,24], an analysis is held regarding model elaboration and adaptation for the needs of soundscape heritage purposes. In this perspective, state-of-the-art machine and deep learning services are implemented both client- (mobile)

and cloud-wise. This approach allows for multilevel semantic monitoring of sound-related heritage, while offering new insights toward sustainable urban and rural growth. Much emphasis is placed on capturing, preserving, and recreating soundscapes and open ancient theater acoustics, representing important cultural artifacts.

1.1. Related Work

Based on the preceding introduction, there are multiple perspectives concerning the related work around the discussed research domain. Data-driven storytelling, as a form of digital, sensemaking narrative, has recently received significant attention. Recognizing the increasing need to support novel means for integrating data visualization into narrative stories, featured cultural and audiovisual heritage projects deploy state-of-the-art technologies to capture, manage, and publish CH data through rich-media storytelling experiences [1–9]. Among others, related services or cultural activities (tomorrow’s heritage) include tourist promotion and environmental preservation/awareness for landscapes and intangible artifacts [1–3], sites modeling/reconstruction and content restoration/documentation [4,5], and multi-disciplinary collaborations in research and education innovations [6–9]. Audiovisual and soundscape-related heritage initiatives also emerge, focusing on historical sound records and landscapes preserved, re-created, and reproduced as means of intangible CH expressions [25–29]. Furthermore, the impact of environmental sounds, noise, and soundscape components is analyzed on various aspects of modern human life, i.e., examining their associations with the residents’ physical/mental health, perception, and behavior, aiming to unveil factors of sustainable growth and development and overall quality of life as well [30–37]. Social media soundscape information can serve for the prediction of health effects of noise pollution in different areas [38]. In this context, cooperative smart-sensing and crowdsourcing practices have been proposed and launched to raise public awareness toward soundscape conservation, safeguarding, and overall ecological consciousness through multimodal mapping capabilities [39–46].

In recent years, mobile devices offered significant advantages in the direction of massive harvesting of large-scale diversified audio and image data, enabling users to exploit their mobile terminals for capturing, recreating, and sharing various events [10–16,44]. Smartphone capabilities can serve for citizen science projects, following a user-centered design and providing motivation factors [45]. The cultural sector is also benefitted from this evolution, adopting these practices to collaboratively collect, share, and annotate heritage sites and artifacts [7,39–42,47]. These processes feature many resemblances with aspects of the MoJo paradigm and other Digital Journalism genres (i.e., Data, Multimedia, Immersive Journalism, etc.) [16–24,48]. Context- and location-aware services can be combined with (multichannel) semantic processing to offer spatiotemporal sound mapping and pattern-related visualizations. Such featured summarization techniques are encountered on generic audio detection and classification tasks, including environmental sound recognition [14,49–59]. In this view, crowdsourced audio data can offer soundscape enhancement with multiple augmentation layers in favor of documentation, data-driven storytelling, and management. The massive research progress on the domain has established multiple pattern recognition schemes and hierarchical semantic audio taxonomies to describe the sound-fields associated with the different social events [13–24,52–59]. Apart from the geographical- and time-related information that a mobile terminal can easily hold, environmental sounds and soundscapes can be classified, filtered, and highlighted based on the associated pattern classification taxonomies, various low-level audio descriptors, other semantic labels concerning the transmitted or perceived emotions, etc. [49–56,60,61]. Furthermore, recent audio and audiovisual captioning trends can offer additional semantic conceptualization meta-data [62–65]. These meta-information augmentation perspectives can accompany the above-discussed sustainable growth and well-being indicators, suggesting added-value innovative services for soundscape preservation and their engaging promotion at environmental, ecological, and heritage views.

A linked popular research topic that significantly propelled multidisciplinary scientific projects and associated knowledge gain is the way of learning by example, through the Machine and Deep Learning (ML/DL) paradigms. Both the audio semantics and the CH domains have also benefited the made breakthroughs and progress [4–7,13–16,52–59]. Hence, sound and acoustic scene recordings can be processed to provide event detection and recognition outcomes, offering pattern-related metadata, content-based description, and management automation (i.e., retrieval, summarization/highlighting, etc.). Coarse classification schemes (i.e., Speech, Music, Other) can be deployed for detecting human activity and other main events, which could be hierarchically extended to additional classes [13–16,54–59,66]. More complex audio patterns have been formed/adapted to the needs of environmental sound monitoring, incorporating additional classes, therefore increasing the pattern recognition difficulty (e.g., the UrbanSound classification task containing 10 environmental sound categories) [15,54–59,66]. These two taxonomies represent the primary/baseline recognition demands that the proposed system should be able to handle (i.e., to extract such class-related metadata). Hand-crafted feature extraction has been extensively used for abstracting audio information to feed ML systems, taking advantage of the perceptual human experience. Early and late integration methods were also deployed, either by temporally fusing base features or by combining multiple classifiers (both in parallel and in cascade order), also increasing the computational load demands [54–59,66]. A recent trend in the field is the use of convolutional networks and DL architectures, shifting from the feature-based representation to automatically forming audio embeddings, as part of the training process [54–59,66,67]. These latest approaches are computationally heavier (especially at the learning phase), while they also require much more labelled/ground-truth samples as inputs, so dedicated datasets are continuously formed to serve the various training and testing needs. Again, the proposed framework should be able to cope with such solutions, as well as to expedite the creation of soundscape-adapted datasets through the process of mobile crowdsourcing.

Summing up, the conducted literature review revealed important aspects of soundscapes, i.e., environmental monitoring, sound and intangible cultural heritage, data-driven documentation, decentralized/smart sensing, etc., with diverse extensions on human health and sustainable growth indicators. Many related publications have attempted to enlighten most of the above viewpoints by utilizing mobile terminals and collaborative mapping [17–19,38–45]. However, to the best of our knowledge, such a multi-faceted approach (like the current one) has not been reported, incorporating sophisticated on-site semantic analysis and crowdsourcing dynamics, as they are advanced in today's ubiquitous society (i.e., in the era of big data and the semantic web). The impact of the anticipated services is also strongly connected to featured projects, which have been deployed to discover and recreate sounds of the past, emanating from the perspectives of acoustic heritage, archaeo-acoustics, and historical acoustics. Such works, supported by limited historical/acoustic data, rely mainly on computational models and simulation outcomes to offer an intangible CH experience, projecting relationships between people and sound over time [46,68–72]. In this direction, we can forestall the dense impact of the proposed Mojo-adapted system, which can document today's soundscapes to be experienced as tomorrow's heritage, taking advantage of semantically enhanced data-driven storytelling. Recalling the importance of ground-truth datasets and crowdsourcing audio semantics in the age of deep learning, the launched model can easily lead to massive soundscape data and metadata. The in-depth analysis of those repositories would reveal finer pattern correlations and taxonomies, with sharper conceptualization capabilities.

1.2. Project Motivation and Research Objectives

The related work presented in the previous section indicates that the field of crowdsourcing soundscape assets is very fruitful and mature, providing significant benefits for cultural heritage preservation and urban development. Audience engagement can be feasible, given a proper framework design. The motivation of the current project em-

anates from the idea of incorporating proper ML/DL analysis for soundscape semantics through a cloud-based architecture. For this reason, early backend implementations for General Audio Classification and Detection are presented and evaluated. The successful implementation of *MoJo-mate*, a mobile application offering machine-assisted reporting with semantically enhanced capture and documentation MoJo facilities [23,24], justifies this approach. The encompassed audio processing and recognition layers exhibit state-of-the-art time-, context- and location-aware ubiquitous computing services, combined with generic/hierarchical pattern classification schemes [13–16]. These content analysis perspectives are considered ideal for meta-information augmentation of environmental sounds and soundscapes, which can be massively crowdsourced as User-Generated Content (UGC) to represent essential sites or places of intangible CH. The multilevel semantic interpretation of audio (and audiovisual) streams, contributed by both experienced and average users, will allow monitoring how the formed soundscapes have evolved and/or are still evolving over time and within special areas of interest. Typical examples include sensitive ecological zones, landscapes with environmental and cultural interest, and places hosting cultural activities (in ancient or modern theaters and music halls), UNESCO world heritage sites, etc.).

The utmost target is to collect the necessary volumes of data in an easy and entertaining way, provide in-situ/real-time and batch semantic analysis modes, augment the physical visiting experience, and enable data-driven storytelling through multiple auralization and visualization layers. Such techniques will allow the monitoring of the way acoustic comfort of historic urban and rural areas is affected by sound space components (e.g., cars, motorbikes, tourists) and, overall, the necessities of improving the environmental qualities. Another important aspect refers to assessing the mediated navigation experience of both physical and virtual visitors, with respect to the offered digital storytelling, derived by soundscapes and environmental acoustics recreation. No doubt, these perspectives are equally important for the processes of intangible CH collection, management, and preservation. In the long-term, sustainable growth and well-being indicators could be systematically monitored, correlated, and predicted in relation to the associated sound-field attributes (e.g., in heritage sites and areas featuring substantial environmental, cultural, or historical interest).

The work presented here is part of a broader project, aiming to collect and document multimedia semantics of soundscape heritage, to be later used for data-driven storytelling. The Logical User-Centered Design (LUCID) [6,7,11,23] was adopted through the whole process, emphasizing the audience engagement and reinforcement part. This was also one of the principal elements that had to be answered in the early beginnings of this undertaking, i.e., the degree to which targeted users would be interested to actively participate and contribute in this effort, which is aligned with the Analysis/Communication phase of standard application development procedures. Hence, a related survey was carefully set-up and executed to serve the needs of audience analysis. The second key factor would be to investigate whether mobile terminals and the associated algorithmic backend can be adapted to the task of crowdsourcing soundscape semantics. In this perspective, ML and DL systems were implemented as the initial/piloting algorithmic solutions and were thoroughly evaluated at various levels to provide a convincing proof-of-concept of the tested scenario.

Based on the above analysis, Research Hypotheses (RH) are stated and put to test, providing a convincing proof of concept of the proposed model, its feasibility, and effectiveness, emphasizing the semantic processing part:

Research Hypothesis 1 (RH1): *It is both feasible and innovative to launch a Mobile Journalism application for soundscape heritage crowdsourcing and data-driven storytelling, and there is an audience willing to use the application and contribute.*

Research Hypothesis 2 (RH2): *General Audio Detection and Classification techniques can be implemented by means of Machine and Deep Learning to serve the required soundscape semantics.*

In this context, risen Research Questions (RQ) accommodated to the listed hypotheses are as follows:

Research Question 1 (RQ1): *How can the MoJo framework be configured for soundscape heritage capturing and documentation? How can the crowdsourced media assets serve the needs for data-driven storytelling?*

Research Question 2 (RQ2): *What are the main classification taxonomies that can be incorporated in the initial backend implementations of soundscape recognition? What is the estimated accuracy and computational load of these algorithmic systems?*

The rest of the paper is organized as follows. The system architecture and concept, as well as the experimental procedures, are presented and justified in the Materials and Methods section. Results and discussion illustrate the corresponding outcomes (and their thorough evaluation), providing multi-perspective analysis with regard to the stated hypotheses and questions. Conclusions are finally drawn, stressing the novel aspects and the contribution of the whole project, followed by the respective Summary section.

2. Materials and Methods

2.1. Integration of State-of-the-Art Audio and Soundscape Semantics on the Cloud

The main target of the current paper is to enhance the semantic aspects of capturing, managing, and recreating soundscapes, engaging the audience in the direction of mobile crowdsourcing and sharing related audio events. In this context, crowdsourced audio data can be comprehended in various ways, one of them being monitoring encountered soundscapes. Theoretically, this can be achieved by manually matching and managing different input streams from end-users, exploiting the aspects of semantic tagging, and annotation at different levels of hierarchy. However, in real-world conditions, difficulties regarding user- and context-related heterogeneities arise, which require the employment of intelligent audio processing and interaction methods, to utilize and benefit from the underlying semantic information of audio data.

While many related processing strategies can be deployed on mobile computing environments, resources for processing and analyzing vast amounts of audio data in a mobile device are typically limited [10–24]. Thus, a strong motivation for embracing cloud-based services emerges in this scenario. In this direction, accessible and highly capable cloud-based computing environments can facilitate the binding of semantically relevant content, by incorporating previous knowledge on individual soundscape characteristics (i.e., the rules that a listener would associate to a specific soundscape) [73].

Prevailing research on intelligent audio analysis and sound recognition is highly focused on the sub-fields of General Audio Detection and Classification (GADC) and Environmental Sound Recognition (ESR). The analysis aims at the semantic description of complex acoustic scenes, relying on a system that inputs an audio signal and outputs the semantic description of that signal. Hence, in this case, the meaningful aspects of a soundscape are to be detected and identified.

State-of-the-art approaches in computer audio intelligence motivate data-driven modeling, through machine learning. A wide variety of pre-processing and classification algorithms can deliver a solid generalization performance, given large amounts of training data. Moreover, the performance of these models is strongly dependent on the quality of the utilized data. For this reason, mobile devices can offer significant advantages in the direction of large-scale diversified labeled audio data gathering and the construction of generic ground-truth semantic audio databases [15].

Efficient pre-processing and semantic monitoring techniques can also be deployed as a front-end client-based system, given the ability to adapt to the variance in the acoustic environments and the respective sound recording conditions. This process can locally interact with the input signal and map it into a latent space, allowing users to on-site-monitor soundscape semantics, with the option to define patterns of interest and associate them with specific audio features, geolocations, and/or visual content [56,57]. The proposed

modular architecture allows the attachment of multi-channelled ambisonics sensors to the client terminal (i.e., soundfield microphones), to apply more sophisticated spatiotemporal localization and mapping that could facilitate the audiovisual content description and management [49–51,74,75]. On the other side, more demanding semantic analysis can be performed on a batch processing mode, as a cloud service, making use of recent advantages on Convolutional Neural Networks (CNN), Deep Learning (DL), and multimodal decision-making systems [58–65]. The focus here lies in the discrimination of time-concurrent audio events in a hierarchical classification taxonomy. This processing type is more adapted to the audio domain and may have considerable advantages over end-to-end solutions. Moreover, a soundscape crowdsourcing approach is favored in the proposed methodology for constructing big datasets, as users are encouraged to contribute with new labeled data while making use of the services. This real-world soundscape intervention approach to audio management systems can offer further conceptual analysis perspectives of crowdsourced audio data, layered on top of existing semantic analysis assets.

2.2. The Implemented Sound Heritage and Storytelling Model

Soundscapes can tell the story of spaces through time. While the acoustic scenes can characterize certain places and ecosystems, they are also in constant movement and evolution, as they change as a whole, and as temporary events occur, breaking the perceived continuity of sound. Treating environmental recordings in this scope allows the design of an interactive storytelling mode, where varying soundscapes can be in the spotlight of the narration.

When a crowdsourcing approach is adopted, definitive and linear storytelling is replaced by a collective narration, formed with the combination of the provided audio recordings and audiovisual assets provided by the users. The criteria that individual listeners follow to access the available files define different perspectives and can form a vast amount of stories that emerge from the provided soundscape recordings. An intuitive design can support interactive storytelling, facilitating the exploration of the dataset in creative ways. Two of the main aspects of treating soundscapes have already been mentioned and they refer to their spatial and temporal evolution. An interactive map, with a supplementary timeline option, can provide the functionality for filtering the data, using both the geographical and temporal information of the recordings. The user can access environmental recordings using an interactive world map, while the option of selecting the time interval within which the recording was created is available. Context-aware content-creating applications can provide such information without manual annotation at the time of the recording [11–16,74,75].

Besides the straightforward spatiotemporal filtering of results, content-based retrieval can form different storytelling paths. Soundscapes that are far away in terms of distance or time may capture similar acoustic scenes, e.g., open theaters, cities, forests. Manual annotation from the content creators can provide a tagging scheme to retrieve relevant assets. By providing a data-driven analysis system on the cloud, several soundscape descriptors can be extracted automatically from the audio characteristics of the recordings. Users can form queries to browse through the dataset, based on the manual and automated tagging of data. In this approach, the integration of featured personalization and recommendation modules can push relevant content to the users, based on their queries and, overall, the monitoring of their behavior and interests.

So far, several scenarios of searching for audio content through textual input, as well as extracting textual descriptors from audio content, have been presented. However, modern trends in Human–Computer Interaction demand more intuitive query processes. In the context of soundscape storytelling, it is possible to retrieve audiovisual content through an audiovisual input query. By recording or providing a soundscape, users should be able to search the database through similarity checks (e.g., pattern matching). This will result in accessing content with audio characteristics that match the input. In the same way, by accepting not only audio recordings but also videos, or accompanying assets (e.g., users

can upload photographs along with the environmental recordings), a mapping between different modalities can be created. By providing certain soundscapes, relevant content can be generated, and vice-versa. This interaction can provide great possibilities in the paths a user can follow to access different stories.

Another meaningful parameter that can boost interactive storytelling functionality is the acoustic modeling of distinctive soundscapes, especially those related to cultural heritage (i.e., the cases of notorious ancient open theaters). This process of defining a transfer function can be used to estimate and imitate the acoustic behavior of a scene. In the case such a functionality is offered, users can provide studio-quality or close-miking recordings with no reverberation and simulate the reproduction of their recordings as if they had been held within various soundscapes [5,56]. It is essential to mention that related functionalities have been recently deployed on the *MoJo-mate* application, facilitating time-, context-, and location-aware audiovisual recordings with significant semantic enhancements concerning the encountered audio patterns and the surrounding acoustic behavior [11,12,16–24]. While these modalities have been successfully integrated and evaluated for the needs of *MoJo* capturing and publishing services, the proposed re-orientation can be even more valuable in the direction of preserving and demonstrating soundscape heritage. Furthermore, the collection of big data in a more organized manner and the gradual construction of semantically enhanced audio (and audiovisual) repositories can force added-value services toward implementing diverse ground-truth sets and their utilization on more sophisticated semantic conceptualization automations. As already stated, such analysis perspectives can be correlated with human well-being, cultural heritage, and sustainable development indicators, which is very important in today's rapidly changing ubiquitous society.

2.3. The Proposed Model Architecture

The work presented emanates from the particularities residing in the vast increase in UGC. Apparently, mobile devices offer significant advantages in the direction of massive harvesting of large-scale diversified labeled audio data. Users' smartphones make the procedures of recording, recreating, and sharing audio and audiovisual material as simple as possible. Professional and nonprofessional users capture audiovisual content using mobile devices (smartphones and tablets) and upload it to the platform. However, multimedia data that are collected through crowdsourcing are often of low quality, due to nonprofessional hardware limitations and the lack of proper training. In this direction, mobile automations add a level of intelligence to assist the process. Difficulties regarding user- and context-related heterogeneities are overcome through the adoption of dedicated audio processing and interaction techniques for the semantic tagging and annotation of audio events.

To this end, the implementation of a 4-layer, cloud-based architecture is shown in Figure 1, offering audio-driven multimedia analysis and classification. Mobile terminals offer sensory and recording software to capture sound and audiovisual data, which can be enhanced with time, geolocation, and other context-aware metadata. The user can upload the created files on the cloud for analysis. The data handling layer is responsible for orchestrating and distributing the incoming data depending on the resource allocation, while also extracting audio tracks from audiovisual material and selecting the channels/segments to be further processed. Next, the audio processing and classification layer takes over, resulting in an assembly of salient (human-crafted) audio features, as presented on the left side of Figure 1 (terminal-wise analysis). A set of dedicated temporal feature integration processes is involved [54,57,59], attempting to classify the sounds identified in the given soundscape through typical Multi-Layer Perceptron (MLP) architectures. Apart from this on-site analysis, heavier processing is deployed on the cloud, utilizing state-of-the-art CNN architectures for machine-driven convolutional feature engines and finer pattern recognition (right side of Figure 1). Overall, these two independent flows employ different-complexity (and computational load) machine learning models, associated with the client-wise and server-wise (cloud) perspectives, as previously stated. The resulting entities are stored

in a repository along with their semantic representation. Based on this information, an interactive map is created, augmented with a timeline bar and multiple semantic filtering options, taking into consideration time, location, and pattern-related tags. The captured audio streams are pinned in this multilevel information mapping so that spatiotemporal monitoring and auralization processes are offered as part of the storytelling. Hence, both the UGC contributors (displayed at the bottom of Figure 1) and the end-users/consumers (depicted at the top of Figure 1) can reproduce the evolution of sound and soundscapes over time, and in relation to the available semantic layers. The main goals of the proposed architecture concern the efficient and purposeful employment of cloud services and mobile artificial intelligence for the support of interactive soundscape exploration. More specifically, the current paper evaluates the individual and ensemble potentials of the two different semantic analysis processes (terminal- and cloud-wise), thus making a convincing proof of concept for their usefulness in the attempted CH data-driven storytelling.

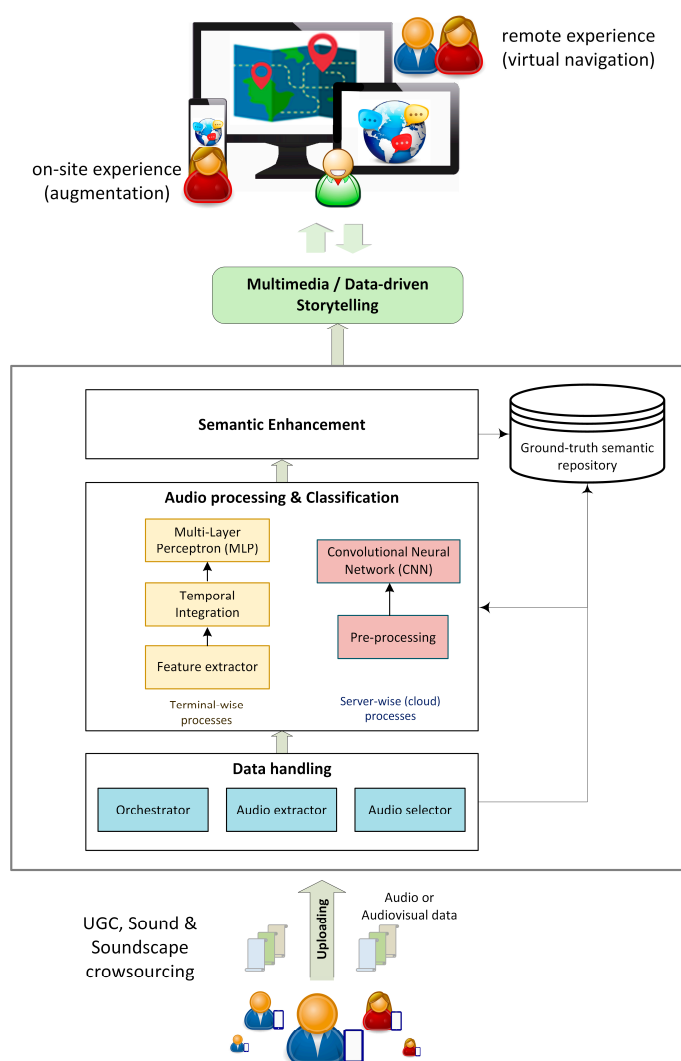


Figure 1. The adopted semantic crowdsourcing model architecture. Terminal-site audio semantics is deployed through feature extraction, temporal integration (enhanced temporal integration (ETi)), and multi-layer perceptron (MLP)-driven pattern recognition. Server-wise semantics are applied in heavier processing modes using convolutional neural networks (CNN) architectures for end-to-end content-based recognition. Captured audio (and audiovisual) data are enhanced with diverse semantic tags and pattern-related metadata, which are documented in the formed ground-truth repository. These media assets also augment the proposed data-driven cultural heritage (CH) storytelling model.

2.4. Experimental Setup

2.4.1. Concept Validation: Preparation of a Questionnaire Survey

The initial hypothesis (RH1) can be examined by answering typical questions for soundscape capturing, sharing, exploration, and specific aspects regarding users' cultural interests and habits, thus retrieving vital feedback. In order to grasp and monitor users' preferences the research utilized a quantitative survey method for data collection, with the formation of a corresponding online questionnaire.

Detailed information regarding this survey is provided in the associated results section, along with the assessment outcomes. An overview of the chosen inquiries is presented here, aiming to justify the adoption and configuration of the formed questionnaire. Hence, background-related questions (soundscape knowledge, relevance, previous use, etc.) were structured in a categorical form of potential answers, with 5-point Likert scales (1–5, from “Totally Disagree” to “Totally Agree” or from “Not at all” to “Very Often”). Binary values (i.e., gender) and higher-dimensional lists were also involved. The items were divided into three subsets, with the former involving basic characteristics/demographics of the users (questions 1–4), the second implicating questions on the participants' background/knowledge on soundscapes (questions 5–10), and the latter containing suggested modalities and usability characteristics of the proposed mobile application (questions 11–17, in Table 1). The test formation was validated after discussions and focus groups with representative users and authorities of various kinds. Specifically, there were involved journalists, cultural and soundscape heritage enthusiasts, multimedia producers/programmers, technologists and researchers in machine/deep learning, environmental sound recognition, audio semantics, etc. The survey was updated based on the received feedback, investigating the audience interest in soundscapes and soundscape heritage, while also estimating the anticipated dynamics of the proposed approach.

Table 1. The analysis questionnaire.

#	Question (Indicative Answers—Range)
1	Age (intervals: <18, 18–25, 26–35, 36–45, 46–55, >55)
2	Gender (Male, Female, N/A)
3	Education (Secondary, University, Master, Student, PhD)
4	Profession (Employee, Freelancer, Student, Unemployed, Retired)
5	Knowledge of what a soundscape is (Yes, Probably Yes, Probably No, No, Don't know)
6	Frequency of soundscape search last year (<5, 5–10, 10–20, >20)
7	Interest in soundscape heritage (1–5)
8	Yearly participation frequency in cultural events containing soundscapes (<3, 3–5, 5–10)
9	Soundscape capturing frequency (Never, Rarely, Sometimes, Frequently, Very often)
10	Most preferred soundscapes to capture (Culture, Environment, People, Urban, Other)
11	Privacy and/or copyright (Both, Privacy only, Copyright only, None)
12	Probability of using the app for soundscape capturing and sharing (1–5)
13	Probability of using the app for own soundscape re-production (1–5)
14	Probability of using the app for others' soundscape re-production (1–5)
15	Technological maturity (Yes, No, Don't know)
16	App capability for soundscape sustainability (1–5)
17	App extra usability features and modalities (6 suggestions: Soundscape search from soundscape, image search from soundscape, soundscape recommendations, given soundscape's related subjects, augmented reality, user sharing and combination)

Table 1 synthesizes the final set of questions selected for the needs of this survey. During the survey preparation, all ethical approval procedures and rules suggested by the “Committee on Research Ethics and Conduct” of the Aristotle University of Thessaloniki were followed. The respective guidelines and information is available online at <https://www.rc.auth.gr/ed/> (accessed on 2 March 2021). Moreover, the declaration of Helsinki and the MDPI directions for the case of pure observatory studies were also taken into account. Specifically, the formed questionnaire was fully anonymized, and the potential participants were informed that they agree to the stated terms upon sending their final answers, while they have the option of quitting anytime, without submitting any data.

2.4.2. Configuration and Validation of the Audio-Semantic Modalities

Aiming to conduct an objective evaluation for both terminal- and server-side classification algorithms, a comparative evaluation between a lightweight feature-based method and a deep learning approach was decided. As already explained, these two approaches represent the earliest algorithmic implementations that the project should launch, so they are investigated in this first research. Specifically, an Enhanced Temporal Integration (ETi) model [57] with a fully connected neural network (i.e., MLP) and typical 2-dimensional CNN topologies [58], proposed as the terminal and server-side classification approaches, respectively, were tested on typical audio classification scenarios, utilizing common datasets. Again, the specific pattern analysis taxonomies are thought of as the minimum, though entirely adequate, pilot developments to provide a convincing proof of concept, while initiating the semantic crowdsourcing process and the gradual construction of the anticipated ground-truth repository, as well.

The classification scenarios involve two datasets, according to a 3-class generic classification and an environmental 10-class scheme. The first one is simulated using the LVLlib-v3 dataset [59], which follows the Speech/Music/Other (SMO) taxonomy, while the 10-class task is based on the UrbanSound8K dataset [55]. This decision is justified by the fact that the Other class of the LVLlib-v3 can be hierarchically split into more classes, which for instance, can follow the scheme of the UrbanSound8k [15]. On the one hand, LVLlib-v3 includes 1.5 h of recordings, and it is available online at m3c.web.auth.gr/research/datasets (accessed on 2 March 2021) and specifies a 3-fold cross-validation strategy to make the results comparable across the algorithms of different creators. On the other hand, UrbanSound8K is a standard benchmark for environmental sound recognition and contains 8.75 h of field recordings, divided into 10 environmental sound categories.

Regarding the classification units, as aforementioned, the ETi with an MLP and a 2-dimensional CNN and were deployed. It is a fact that the latest deep learning approaches can process raw waveform data [58], but the 2-dimensional topologies deliver the best balance between performance and computational cost and were selected in this case. In addition to this, the ETi method proved to be a lightweight solution for conventional feature-based classification, offering decent performance [59]. The CNN processes mel-spectrogram patches, with a shape of 84 time-steps \times 56 bands. Spectral analysis is executed on a 512/256 sample size/step basis with a sampling rate of 22,050 Hz. The convolutional network consists of four consecutive CPD blocks (each one containing successive Convolutional, Pooling, and Dropout layers), a Global Average Pooling (GAP), and two Fully Connected (FC) layers with an additional Dropout layer in between. The number of filters is 16, 32, 64, and 128 for the convolutional layers with a kernel size of 3×3 , while the pooling size is set to 2×2 . The number of neurons of the FC layers was set to 64 and according to the number of classes, respectively. A schematic of the deployed CNN architecture is given in Figure 2. The MLP configuration takes as input 200 features, extracted in a 512/256 sample size/step basis and integrated according to the ETi method. The extracted baseline features are 12 MFCCs, Perceptual Sharpness, Perceptual Spread Spectral Centroid, Spectral Decrease, Spectral Flatness, Spectral Flux, Spectral Kurtosis, Spectral Rolloff, Spectral Skewness, Spectral Slope, Spectral Spread, Spectral Variation, and Zero Crossing Rate. These features are temporally integrated

using the Mean Value, Standard Deviation, Skewness, Kurtosis, Mean Absolute Sequential Difference, Mean Crossing Rate, Flatness, and Crest Factor metrics [54]. A typical network setup was deployed with two hidden layers, featuring 64 and 32 neurons. Concerning the rest of the parameters, both networks follow the same configuration: The ReLU function was used as activation for all intermediate (Convolutional and Fully Connected) layers and SoftMax for the output layer, Categorical Cross-Entropy as the loss function, and Adam as the optimizer. Dropout was set to 25%.

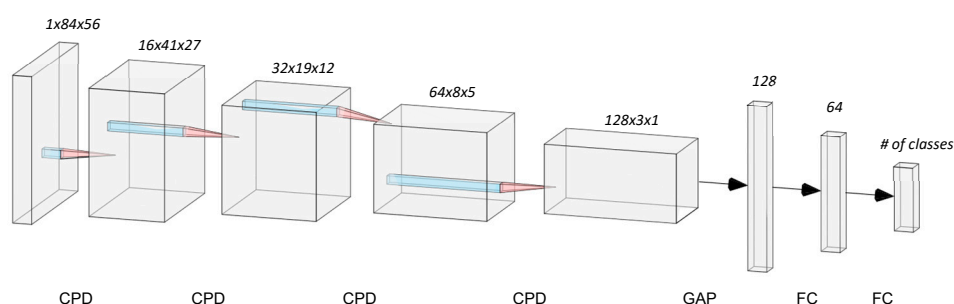


Figure 2. Schematic of the deployed CNN architecture, where the succession of the used convolutional, pooling, and dropout blocks (CPD), global average pooling (GAP), and fully connected (FC) blocks and layers is presented. The evolution of the data format along the network is also depicted.

3. Experimental Results

3.1. Concept Validation: Audience Analysis Results

To examine the proposed research question regarding the usefulness of an application similar to the one proposed, we undertook an online survey (N = 171). Data collection via an online survey appeared to be the most realistic and feasible method to reach a broad audience that would lead to a representative sample. From the collected sample, 61.4% of the responders were females, 36.4% were males, while 2.3% preferred not to state their gender. Regarding sample's distribution in the given age groups 18–25, 26–35, 36–45, 46–55, and above 55, the results are 30.4%, 48%, 15.8%, 5.3%, and 0.6% respectively. In general, the results showed that many people are not familiar with what a soundscape is. In more detail, given six (6) common acoustic scenarios, the participants were asked to identify which of them could be considered soundscapes. The study shows that over 70% of the participants were able to identify the cases in which actual soundscapes were given (e.g., sound of a bell in a village), while on the other hand, about 40% of them had difficulty distinguishing what was rather a false-positive soundscape (e.g., a teleconference). The majority of the participants expressed their interest in the mediated soundscape experience that is aimed within the current project, as thoroughly analyzed below.

In order to balance the diversity of the sample, we selected 104 out of the 171 participants, the ones positively posed against soundscape heritage, considering it an important factor for sustainability, especially in cultural places. This division was also dictated by the fact that some of the questions require a basic background and understanding of soundscapes. Thus, it would be unreliable or biased to equally balance the replies on soundscape heritage and semantics of those without a basic comprehension of the associated terms. The results from the selected sample (N = 104) show that only 30% of the participants explore soundscapes once a month. In addition, 30% of the participants record sounds and soundscapes frequently, while 66% of them record mostly cultural-related content. Moreover, 40% stated that they want soundscapes to be available for future reference and/or exploration. Moreover, the selected sample featured a clear interest in soundscape preservation over time, while the majority of them (69%) stated that they use their mobile devices for soundscape capturing and sharing. On the other hand, from the smaller percentage of participants not showing interest in sound heritage (13%) or being moderate

about it (26%), almost half of them capture soundscapes quite often, thus constituting a group of potential application users.

It is noteworthy that although soundscape capturing, sharing, and reproduction is not that widespread, the selected participants showed a high interest in the proposed application. More specifically, 89% of the participants would use an application like the one proposed for soundscape capturing and sharing. Further, 77% would use the application for the reproduction of what was once recorded, either by themselves or other users. Finally, 87.5% of the participants believe that an application similar to the one proposed here would assist in the sustainability of soundscapes' heritage.

Figure 3 provides graph statistics for both the whole ($N = 171$) and the subset group ($N = 104$), concerning some of the important questions (namely, #12, #13, #14, and #16). It can be noticed that most users are willing to capture and contribute soundscape recordings, especially the ones belonging to the selected subset (a mean value of 4.03 is observed with a st.dev of ± 1.11 , compared to the 3.47 ± 1.11 respective values of the entire population). Likewise, almost all participants consider it very likely to reproduce their own or other soundscapes, appraising the impact of the application to sound and soundscape heritage (again, the mean values are higher and with slightly smaller dispersion in the case of the selected sub-group). In summary, the results of the conducted survey validate the first hypothesis (RH1) and the associated research question (RQ1) that there is an audience willing to use the suggested Mojo application, contributing to soundscape heritage crowdsourcing and the subsequent data-driven storytelling (even subjects that do not fully comprehend the underlying principles of the soundscape semantic).

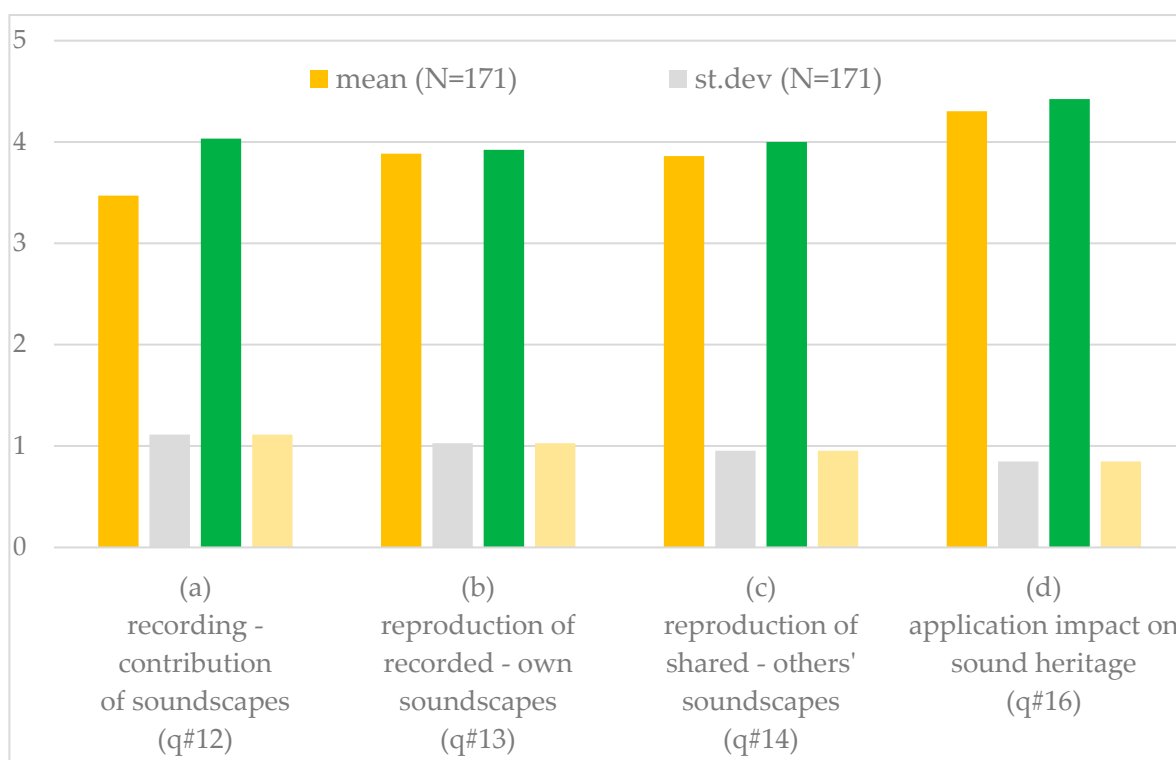


Figure 3. Results on the probability (a) to record—contribute soundscapes (q#12); (b) to reproduce recorded (own) soundscapes (q#13); (c) to reproduce recorded (others') soundscapes (q#14); and (d) on the estimated impact of the application in sound heritage. Statistical moments of mean and standard deviation (st.dev) are presented both for the entire population ($N = 171$) and the selected subset ($N = 104$).

3.2. Audio Classification Results

Classification results are presented (Table 2) in terms of accuracy statistics (mean value/standard deviation) as they have been extracted by the associated evaluation in

unknown samples (testing dataset). In this manner, it is anticipated the expected generalization performance of the trained modalities, i.e., their ability to provide accurate classification estimates to entirely new/unknown data. On the LVLlib-v3 dataset, both classification modules perform almost equally, achieving high scores, similar to relevant tests conducted in previous works [57–59,66]. As expected, the CNN classifier performs slightly better. As already explained, the Urbansound8k dataset involves a 10-class scheme, making the classification problem more demanding, compared to the 3-class LVLlib-v3 taxonomy. While a reasonable accuracy drop is noticed for this reason (i.e., in Urbansound8k), the performance ratings of these models are in line with the current state-of-the-art on the same datasets. Concerning further the UrbanSound8k dataset, the high learning capacity of the CNN is more evident, making the performance gap wider, where the deep network clearly outperforms the conventional model. However, the utilized temporal integration technique ensures decent classification accuracy, making the feature-based approach capable of successfully accomplishing the more demanding task.

Table 2. Classification accuracy (mean \pm st.dev%) on the LVLlib-v3 and UrbanSound8k Datasets.

	LVLlib-v3	UrbanSound8k
ETi	84.4 \pm 4.1	68.4 \pm 3.9
CNN	86.4 \pm 3.9	72.2 \pm 5.9

The results show that the ETi lives up to the standards of deep learning approaches, especially when computational resources are limited [13,16,56]. This was further investigated, and a computational complexity evaluation was also executed. The additional evaluation involves the measurement of prediction times for both models, and a relative presentation of the results was decided because absolute measurements can significantly vary on different processing units. Table 3 depicts the computational cost in terms of network size and prediction times.

Table 3. Network size and relative computational complexity for the ETi and CNN models.

	Number of Parameters	Complexity
ETi	15k	1 \times
CNN	100k	2 \times

It can be noticed that in the case of the ETi approach, network size is significantly smaller, facilitating the deployment on devices with low processing power. Nevertheless, the size of the CNN is not that large to make the deployment of the model impossible in the modern mobile computing devices. Summing up, the CNN can equip both client- and cloud-wise semantic analysis services, while the ETi provides adequate performance at the lowest processing cost. These findings directed our decision for selecting the ETi and the CNN as client- and cloud-wise classification solutions, respectively.

Overall, based on evaluation results of the trained models, and the justification concerning the selection of these two demanding datasets, the remaining research hypothesis (RH2) and question (RQ2) are validated/positively answered. Hence, the adopted audio classification schemes, suited for pattern-related soundscape semantics, can be served through relatively light-weight (concerning the required memory and computation load) ML and DL modules. Two related systems have been successfully trained and evaluated as the initial algorithmic backend solutions. The accuracy of those models is already more than satisfactory. However, it can be further enhanced through the users' feedback (and the implicated semi-supervised learning features) deployed within the proposed MoJo framework. Furthermore, the hierarchical and/or hybrid combination of the two taxonomies, along with the initiation of the crowdsourcing process, would lead to the gradual construction of a dedicated dataset. This problem-adapted ground-truth repository

would facilitate the training of more sophisticated ML and DL networks with superior performance and additional semantic conceptualization perspectives.

4. Discussion

The current paper introduces MoJo services updated and adapted to the need of semantic soundscape, crowdsourcing, management, and data-driven storytelling. Based on the conducted experiments, the stated hypotheses have been fully verified, i.e., the audience is interested in such a mobile application (RH1). Furthermore, current technology is adequately mature to reliably deliver the wanted functionalities through General Audio Detection and Classification techniques deployed through Machine and Deep Learning networks to serve the required soundscape semantics (RH2). Furthermore, specific audio processing and semantic analysis features were tested in an effort to quantify the implementation parameters set in RQ1. The configured modalities, both client- and server-wise, exhibit remarkable accuracy with acceptable computational load. Based on the previous experience with the *MoJo-mate* platform [11–24], especially for the data shaping, presentation, and publishing part, the proposed model can efficiently deploy the desired data-driven storytelling and management services, which have a heavy impact on the CH domain. Concerning the technological adequacy and reliability that RQ2 inquires, the proposed integration seems to overcome the expected difficulties and to suitably serve the desired semantic enhancement, documentation, and auralization/reproduction perspectives. Specifically, along with the above-mentioned low-level measurement modes, the software also provides long-term audio analysis capabilities, based on semantic audio processing concepts [56]. This higher-level mode brings real-time audio-pattern recognition, visually resulting in an event detection markup timeline. A dynamic audio-samples database is used as a pattern-storing matrix, which is configurable by users. Samples can be added, by making a simple recording, and deleted as well. Relying on the *MoJo-mate* application experience, a user-friendly measurement session manager is feasible, allowing each measurement to be easily stored on the mobile terminal memory and recalled on demand. Additional session measurement data can be stored, including title, location, user's comments, etc., while the position is automatically determined utilizing the device GPS. Likewise, timestamps are easily overlaid by the device, while a handy interface allows photo and video capturing of the measurement location, i.e., the recorded soundscape. A cloud-based session manager handles all the users' data, aiming at building a user-generated, spatiotemporal digital map used for storing measurements. Users can store, update, and retrieve raw audio data and their corresponding analysis output. All measurements uploaded to the cloud are accessible by anyone who uses the application. By exploiting the GPS sensor and cellular data capabilities, the application can easily classify and group measurements by geographical location and kind. Thus, a user can instantly check and confirm the correctness of a specific measurement by comparing it to similar ones, provided by other users. They can even obtain the desired data without making a measurement.

Audio recognition usually refers to different recognition tasks, like acoustic scene detection, speech recognition, and speaker recognition. Systems that implement such models are oriented to specific scenarios of recognition. Applying audio recognition to soundscape management is a much more complicated task. The information that can be extracted from the recordings is not pre-defined. Environmental noise can contain multiple layers of audio information and includes a great variety of possible temporal audio events. In the proposed approach, an ensemble of algorithms is proposed to compose a hierarchical classification scheme. For example, an algorithm for acoustic scene classification can classify an acoustic scene as "river," while an audio event detection can recognize a "speech" audio event at a certain time, triggering algorithms that extract information concerning speaker diarization and spoken language, thus triggering algorithms that transform speech-to-text, etc. This approach results in several layers or perspectives of audio monitoring, giving the user the possibility to browse through the data with different levels of information abstraction. In the context of environmental recordings, several information layers concerning

acoustic characteristics, noise levels, etc. can also be included in the defined hierarchical scheme. Another interesting approach for analyzing complex scenes is automated audio (and audiovisual) captioning. This defines an end-to-end model that maps acoustic scenes to descriptive texts but can also correlate them with associated visual entities.

5. Summary

The current work focuses on the collaborative collection and documentation of soundscapes and environmental sound semantics. The whole approach has many similarities with sophisticated Mobile Journalism services, assisting professional and citizen journalists in collecting news-items and shaping them into featured data-driven storytelling. Crowdsourcing media assets for cultural heritage is a fruitful field that can engage an audience through successful design and motivation decisions. Along with audio/multimedia content and metadata, semantic annotation can be incorporated through typical sound classification scenarios. A comparative evaluation between a lightweight feature-based machine learning network and a convolutional deep learning architecture was decided for the terminal and server-side algorithmic approaches, employing two different classification taxonomies with applicable audio datasets. Adopting the LUCID design and development methodology, audience engagement and reinforcement was triggered through an online survey, confirming that users are willing to contribute and appraise the impact of the application to crowdsource sound semantic and soundscape heritage.

The innovation of the paper lies in the incorporation of sophisticated on-site semantic analysis and crowdsourcing dynamics, as they are advanced in today's ubiquitous society (i.e., in the era of big data and the semantic web). Specifically, one of the advantages of this approach, which also highlights one of the main novelties of our work, is that besides collecting and storing resources (recordings of soundscapes and corresponding metadata) from users, it is possible to provide semantically enhanced services on the cloud. Environmental sound recognition is addressed in the paper as one of the featured functionalities using machine learning techniques. Relying on the so-called *MoJo-mate* platform, an analysis is held regarding model elaboration and adaptation for the needs of soundscape heritage. A four-layer, cloud-based architecture was deployed, incorporating two independent flows that employ different-complexity (and computational load) ML/DL models, associated with the client-wise and server-wise (cloud) perspectives for soundscape semantics. The achieved model performance supports the feasibility of the proposed system. The impact of the proposed MoJo-adapted system lies in the ability to document today's soundscapes to be experienced as tomorrow's heritage, taking advantage of semantically enhanced data-driven storytelling.

Author Contributions: Conceptualization, C.D. and M.E.S.; methodology, N.V., L.V. and I.T.; software, L.V. and N.V.; validation, M.E.S., L.V. and C.D.; formal analysis, N.V. and I.T.; investigation, M.E.S.; resources, C.D. and L.V.; data curation, C.D., L.V. and N.V.; writing—original draft preparation, all authors; writing—review and editing, all authors; visualization, M.E.S.; supervision, C.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available in a publicly accessible repository that does not issue DOIs. Publicly available datasets were analyzed in this study. This data can be found here: [m3c.web.auth.gr/research/datasets] (accessed on 2 March 2021)].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yao, D.; Zhang, K.; Wang, L.; Law, R.; Zhang, M. From Religious Belief to Intangible Cultural Heritage Tourism: A Case Study of Mazu Belief. *Sustainability* **2020**, *12*, 4229. [[CrossRef](#)]

2. Pollice, F.; Rinella, A.; Epifani, F.; Miggiano, P. Placetelling[®] as a Strategic Tool for Promoting Niche Tourism to Islands: The Case of Cape Verde. *Sustainability* **2020**, *12*, 4333. [\[CrossRef\]](#)
3. Carta, M.; Ronsivalle, D. Neoanthropocene Raising and Protection of Natural and Cultural Heritage: A Case Study in Southern Italy. *Sustainability* **2020**, *12*, 4186. [\[CrossRef\]](#)
4. Doulamis, A.; Voulodimos, A.; Protopapadakis, E.; Doulamis, N.; Makantasis, K. Automatic 3D Modeling and Reconstruction of Cultural Heritage Sites from Twitter Images. *Sustainability* **2020**, *12*, 4223. [\[CrossRef\]](#)
5. Dimoulas, C.; Kalliris, G.; Chatzara, E.; Tsipas, N.; Papanikolaou, G. Audiovisual production, restoration-archiving and content management methods to preserve local tradition and folkloric heritage. *J. Cult. Herit.* **2014**, *15*, 234–241. [\[CrossRef\]](#)
6. Chatzara, E.; Kotsakis, R.; Tsipas, N.; Vrysis, L.; Dimoulas, C. Machine-Assisted Learning in Highly-Interdisciplinary Media Fields: A Multimedia Guide on Modern Art. *Educ. Sci.* **2019**, *9*, 198. [\[CrossRef\]](#)
7. Psomadaki, O.; Dimoulas, C.; Kalliris, G.; Paschalidis, G. Digital storytelling and audience engagement in cultural heritage management: A collaborative model based on the Digital City of Thessaloniki. *J. Cult. Herit.* **2019**, *36*, 12–22. [\[CrossRef\]](#)
8. Oomen, J.; Ordelman, R. Accessing Audiovisual Heritage: A Roadmap for Collaborative Innovation. *IEEE Multimed.* **2011**, *18*, 4–10. [\[CrossRef\]](#)
9. Liew, C.L. Digital audiovisual heritage: An exploration of challenges and a community-based approach to preservation. In Proceedings of the 9th International Conference on Communities & Technologies-Transforming Communities, Vienna, Austria, 3–7 June 2019; pp. 76–80.
10. Dimoulas, C.; Veglis, A.; Kalliris, G. Application of mobile cloud based technologies in news reporting: Current trends and future perspectives. In *Mobile Networks and Cloud Computing Convergence for Progressive Services and Applications*; Rodrigues, J., Lin, K., Lloret, J., Eds.; IGI Global: Hershey, PA, USA, 2014; Chapter 17; pp. 320–343.
11. Sidiropoulos, E.A.; Vryzas, N.; Vrysis, L.; Avraam, E.; Dimoulas, C.A. Collecting and Delivering Multimedia Content during Crisis. In Proceedings of the EJTA Teacher's Conference 2018: Crisis Reporting, Thessaloniki, Greece, 18–19 October 2018.
12. Vryzas, N.; Sidiropoulos, E.; Vrysis, L.; Avraam, E.; Dimoulas, C.A. A mobile cloud computing collaborative model for the support of on-site content capturing and publishing. *J. Media Crit. [JMC]* **2018**, *4*, 349–364.
13. Vrysis, L.; Tsipas, N.; Dimoulas, C.; Papanikolaou, G. Mobile Audio Intelligence: From Real Time Segmentation to Crowd Sourced Semantics. In Proceedings of the 10th Audio Mostly (A Conference on Interaction with Sound), Thessaloniki, Greece, 7–9 October 2015.
14. Dimoulas, C.A.; Symeonidis, A.L. Syncing Shared Multimedia through Audiovisual Bimodal Segmentation. *IEEE Multimed.* **2015**, *22*, 26–42. [\[CrossRef\]](#)
15. Vrysis, L.; Tsipas, N.; Dimoulas, C.; Papanikolaou, G. Crowdsourcing Audio Semantics by Means of Hybrid Bimodal Segmentation with Hierarchical Classification. *J. Audio Eng. Soc.* **2016**, *64*, 1042–1054. [\[CrossRef\]](#)
16. Vryzas, N.; Sidiropoulos, E.; Vrysis, L.; Avraam, E.; Dimoulas, C. jReporter: A Smart Voice-Recording Mobile Application. In Proceedings of the 146th Audio Engineering Society Convention, Dublin, Ireland, 20–23 March 2019.
17. Aiello, L.M.; Schifanella, R.; Quercia, D.; Aletta, F. Chatty maps: Constructing sound maps of urban areas from social media data. *R. Soc. Open Sci.* **2016**, *3*, 150690. [\[CrossRef\]](#)
18. Bollini, L.; Della Fazio, I. Situated Emotions: The Role of the Soundscape in a Geo-Based Multimodal Application in the Field of Cultural Heritage. In Proceedings of the International Conference on Computational Science and Its Applications, Cagliari, Italy, 1–4 July 2020; Springer: Cagliari, Italy, 2020; pp. 805–819.
19. Krause, B. Anatomy of the soundscape: Evolving perspectives. *J. Audio Eng. Soc.* **2008**, *56*, 73–80.
20. International Organization for Standardization (ISO). *ISO/DIS 12913-1 Acoustics—Soundscape—Part 1: Definition and Conceptual Framework*; ISO: Geneva, Switzerland, 2013.
21. Comunità, M.; Gerino, A.; Lim, V.; Picinali, L. Web-based binaural audio and sonic narratives for cultural heritage. In *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*, York, UK, 27–29 March 2019; Audio Engineering Society: York, UK, 2019.
22. Chourmouziadou, K.; Sakantamis, K. Soundscape: Investigation and application of an innovative urban design parameter. In *Proceedings of International Conference on “Changing Cities”: Spatial, Morphological, Formal & Socio-Economic Dimensions*, Thessaloniki, Greece, 18–21 March 2013; Aristotle University of Thessaloniki: Thessaloniki, Greece, 2013.
23. Sidiropoulos, E.; Vryzas, N.; Vrysis, L.; Avraam, E.; Dimoulas, C. Growing Media Skills and Know-How in Situ: Technology-Enhanced Practices and Collaborative Support in Mobile News-Reporting. *Educ. Sci.* **2019**, *9*, 173. [\[CrossRef\]](#)
24. Vryzas, N.; Sidiropoulos, E.; Vrysis, L.; Avraam, E.; Dimoulas, C. Machine-assisted reporting in the era of Mobile Journalism: The MOJO-mate platform. *Strategy Dev. Rev.* **2019**, *9*, 22–43. [\[CrossRef\]](#)
25. Suárez, R.; Alonso, A.; Sendra, J.J. Intangible cultural heritage: The sound of the Romanesque cathedral of Santiago de Compostela. *J. Cult. Herit.* **2015**, *16*, 239–243. [\[CrossRef\]](#)
26. Bijsterveld, K. (Ed.) *Soundscapes of the Urban Past: Staged Sound as Mediated Cultural Heritage*; Sound Studies; Transcript Verlag: Bielefeld, Germany, 2013.
27. Bartalucci, C.; Luzzi, S. The soundscape in cultural heritage. In *IOP Conference Series: Materials Science and Engineering*; IOP Publishing: Bristol, UK, 2020; Volume 949, p. 012050.
28. Maffei, L.; Brambilla, G.; Di Gabriele, M. Soundscape as part of the cultural heritage. In *Soundscape and the Built Environment*; Kang, J., Schulte-Fortkamp, B., Eds.; CRC Press: Cleveland, OH, USA, 2015.

29. Masullo, M.; Castanò, F.; Toma, R.A.; Maffei, L. Historical Cloisters and Courtyards as Quiet Areas. *Sustainability* **2020**, *12*, 2887. [\[CrossRef\]](#)
30. Berkouk, D.; Bouzir, T.A.K.; Maffei, L.; Masullo, M. Examining the Associations between Oases Soundscape Components and Walking Speed: Correlation or Causation? *Sustainability* **2020**, *12*, 4619. [\[CrossRef\]](#)
31. Torresin, S.; Albatici, R.; Aletta, F.; Babich, F.; Kang, J. Assessment Methods and Factors Determining Positive Indoor Soundscapes in Residential Buildings: A Systematic Review. *Sustainability* **2019**, *11*, 5290. [\[CrossRef\]](#)
32. Torresin, S.; Aletta, F.; Babich, F.; Bourdeau, E.; Harvie-Clark, J.; Kang, J.; Lavia, L.; Radicchi, A.; Albatici, R. Acoustics for Supportive and Healthy Buildings: Emerging Themes on Indoor Soundscape Research. *Sustainability* **2020**, *12*, 6054. [\[CrossRef\]](#)
33. Zhao, Z.; Wang, Y.; Hou, Y. Residents' Spatial Perceptions of Urban Gardens Based on Soundscape and Landscape Differences. *Sustainability* **2020**, *12*, 6809. [\[CrossRef\]](#)
34. Hong, X.-C.; Zhu, Z.-P.; Liu, J.; Geng, D.-H.; Wang, G.-Y.; Lan, S.-R. Perceived Occurrences of Soundscape Influencing Pleasantness in Urban Forests: A Comparison of Broad-Leaved and Coniferous Forests. *Sustainability* **2019**, *11*, 4789. [\[CrossRef\]](#)
35. Calleri, C.; Astolfi, A.; Pellegrino, A.; Aletta, F.; Shtrepi, L.; Bo, E.; Di Stefano, M.; Orecchia, P. The Effect of Soundscapes and Lightscares on the Perception of Safety and Social Presence Analyzed in a Laboratory Experiment. *Sustainability* **2019**, *11*, 3000. [\[CrossRef\]](#)
36. Zuo, L.; Zhang, J.; Zhang, R.J.; Zhang, Y.; Hu, M.; Zhuang, M.; Liu, W. The Transition of Soundscapes in Tourist Destinations from the Perspective of Residents' Perceptions: A Case Study of the Lugu Lake Scenic Spot, Southwestern China. *Sustainability* **2020**, *12*, 1073. [\[CrossRef\]](#)
37. Sztubecka, M.; Skiba, M.; Mrówczyńska, M.; Mathias, M. Noise as a Factor of Green Areas Soundscape Creation. *Sustainability* **2020**, *12*, 999. [\[CrossRef\]](#)
38. Gasco, L.; Schifanella, R.; Aiello, L.M.; Quercia, D.; Asensio, C.; de Arcas, G. Social media and open data to quantify the effects of noise on health. *Front. Sustain. Cities* **2020**, *2*, 41. [\[CrossRef\]](#)
39. Li, C.; Liu, Y.; Haklay, M. Participatory soundscape sensing. *Landsc. Urban Plan.* **2018**, *173*, 64–69. [\[CrossRef\]](#)
40. Brambilla, G.; Pedrielli, F. Smartphone-Based Participatory Soundscape Mapping for a More Sustainable Acoustic Environment. *Sustainability* **2020**, *12*, 7899. [\[CrossRef\]](#)
41. Yelmi, P.; Kuşcu, H.; Yantaç, A.E. Towards a sustainable crowdsourced sound heritage archive by public participation: The soundsslike project. In Proceedings of the 9th Nordic Conference on Human-Computer Interaction, Gothenburg, Sweden, 23–27 October 2016; pp. 1–9.
42. Yelmi, P.; Kaki, S. Designing an Experiential Exhibition for Raising Public Awareness of Cultural Sounds to Safeguard Sonic Intangible Cultural Heritage Values. *Int. J. Soc. Sci. Humanit.* **2019**, *9*. [\[CrossRef\]](#)
43. Dumyahn, S.L.; Pijanowski, B.C. Soundscape conservation. *Landsc. Ecol.* **2011**, *26*, 1327. [\[CrossRef\]](#)
44. Radicchi, A. Hush City: A new mobile application to crowdsource and assess 'everyday quiet areas' in cities. In Proceedings of the Invisible Places: The International Conference on Sound, Urbanism and the Sense of Place, São Miguel Island, Portugal, 7–9 April 2017; pp. 7–9.
45. Luna, S.; Gold, M.; Albert, A.; Ceccaroni, L.; Claramunt, B.; Danylo, O.; Haklay, M.; Kottmann, R.; Kyba, C.; Piera, J.; et al. Developing mobile applications for environmental and biodiversity citizen science: Considerations and recommendations. In *Multimedia Tools and Applications for Environmental & Biodiversity Informatics*; Springer: Cham, Switzerland, 2018; pp. 9–30.
46. Milone, F.; Camarda, D. Modeling Knowledge in Environmental Analysis: A New Approach to Soundscape Ecology. *Sustainability* **2017**, *9*, 564. [\[CrossRef\]](#)
47. Hristova, D.; Aiello, L.M.; Quercia, D. The new urban success: How culture pays. *Front. Phys.* **2018**, *6*, 27. [\[CrossRef\]](#)
48. Katsaounidou, A.; Dimoulas, C.; Veglis, A. *Cross-Media Authentication and Verification: Emerging Research and Opportunities*; IGI Global: Hershey, PA, USA, 2018.
49. Dimoulas, C.; Avdelidis, K.; Kalliris, G.; Papanikolaou, G. Sound Source Localization and B-Format Enhancement Using Sound Field Microphone Sets. In Proceedings of the 122nd AES Convention, Vienna, Austria, 5–8 May 2007.
50. Dimoulas, C.; Kalliris, G.; Avdelidis, K.; Papanikolaou, G. Improved Localization of Sound Sources Using Multi-Band Processing of Ambisonic Components. In Proceedings of the 126th AES Convention, Munich, Germany, 7–10 May 2009.
51. Dimoulas, C.; Kalliris, G.; Avdelidis, K.; Papanikolaou, G. Spatial Audio Content Management within the MPEG-7 standard of Ambisonic Localization and Visualization Descriptions. In Proceedings of the 126th AES Convention, Munich, Germany, 7–10 May 2009.
52. Salamon, J.; Bello, J.P. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Process. Lett.* **2017**, *24*, 279–283. [\[CrossRef\]](#)
53. Abdoli, S.; Cardinal, P.; Koerich, A.L. End-to-end environmental sound classification using a 1D convolutional neural network. *Expert Syst. Appl.* **2019**, *136*, 252–263. [\[CrossRef\]](#)
54. Bountourakis, V.; Vrysis, L.; Konstantoudakis, K.; Vryzas, N. An Enhanced Temporal Feature Integration Method for Environmental Sound Recognition. *Acoustics* **2019**, *1*, 410–422. [\[CrossRef\]](#)
55. Salamon, J.; Jacoby, C.; Bello, J.P. A dataset and taxonomy for urban sound research. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 18–19 November 2014; pp. 1041–1044.
56. Vrysis, L.; Dimoulas, C.; Kalliris, G.; Papanikolaou, G. Mobile Audio Measurements Platform: Towards Audio Semantic Intelligence into Ubiquitous Computing Environments. *Audio Eng. Soc. Conv.* **2013**, *134*, 8912.

57. Vrysis, L.; Tsipas, N.; Dimoulas, C.; Papanikolaou, G. Extending Temporal Feature Integration for Semantic Audio Analysis. *Audio Eng. Soc. Conv.* **2017**, *142*, 9808.
58. Vrysis, L.; Thoidis, I.; Dimoulas, C.; Papanikolaou, G. Experimenting with 1D CNN Architectures for Generic Audio Classification. *Audio Eng. Soc. Conv.* **2020**, *148*, 10329.
59. Vrysis, L.; Tsipas, N.; Thoidis, I.; Dimoulas, C. 1D/2D Deep CNNs vs. Temporal Feature Integration for General Audio Classification. *J. Audio Eng. Soc.* **2020**, *68*, 66–77. [[CrossRef](#)]
60. Vryzas, N.; Vrysis, L.; Matsiola, M.; Kotsakis, R.; Dimoulas, C.; Kalliris, G. Continuous Speech Emotion Recognition with Convolutional Neural Networks. *J. Audio Eng. Soc.* **2020**, *68*, 14–24. [[CrossRef](#)]
61. Kotsakis, R.; Dimoulas, C.; Kalliris, G.; Veglis, A. Emotional Prediction and Content Profile Estimation in Evaluating Audiovisual Mediated Communication. *Int. J. Monit. Surveill. Technol. Res. (IJMSTR)* **2014**, *2*, 62–80. [[CrossRef](#)]
62. Wenbin, C.X.; Fan, R.; Xiong, D. Zhao. Visual Relationship Embedding Network for Image Paragraph Generation. *IEEE Trans. Multimed.* **2020**, *22*, 2307–2320.
63. Wang, C.; Liaw, P.; Liang, K.; Wang, J.; Chang, P. Video Captioning Based on Joint Image–Audio Deep Learning Techniques. In Proceedings of the 9th International Conference on Consumer Electronics, Berlin, Germany, 8–11 September 2019.
64. Drossos, K.; Adavanne, S.; Virtanen, T. Automated audio captioning with recurrent neural networks. In Proceedings of the 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 15–18 October 2017; pp. 374–378.
65. Drossos, K.; Lipping, S.; Virtanen, T. Clotho: An Audio Captioning Dataset. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 736–740.
66. Vrysis, L.; Tsipas, N.; Thoidis, I.; Dimoulas, C. Enhanced Temporal Feature Integration in Audio Semantics. *J. Audio Eng. Soc.* **2021**. [[CrossRef](#)]
67. Tsipas, N.; Vrysis, L.; Konstantoudakis, K.; Dimoulas, C. Semi-supervised audio-driven TV-news speaker diarization using deep neural embeddings. *J. Acoust. Soc. Am.* **2020**, *148*, 3751–3761. [[CrossRef](#)] [[PubMed](#)]
68. Katz, B.F.; Murphy, D.; Farina, A. The Past Has Ears (PHE): XR Explorations of Acoustic Spaces as Cultural Heritage. In *International Conference on Augmented Reality, Virtual Reality and Computer Graphics*; Springer: Cham, Switzerland, 2020; pp. 91–98.
69. Aletta, F.; Kang, J. Historical acoustics: Relationships between people and sound over time. *Acoustics* **2020**, *2*, 128–130. [[CrossRef](#)]
70. Suárez, R.; Alonso, A.; Sendra, J.J. Archaeoacoustics of intangible cultural heritage: The sound of the Maior Ecclesia of Cluny. *J. Cult. Herit.* **2016**, *19*, 567–572. [[CrossRef](#)]
71. Kytö, M.; Rémy, N.; Uimonen, H.; Acquier, F.; Bérubé, G.; Chelkoff, G.; Said, N.G.; Laroche, S.; Mcoisans, J.; Tixier, N.; et al. European Acoustic Heritage. Ph.D. Thesis, Tampere University of Applied Sciences, Tampere, Finland, 2012; p. 108.
72. Gasco, L.; Clavel, C.; Asensio, C.; de Arcas, G. Beyond sound level monitoring: Exploitation of social media to gather citizens subjective response to noise. *Sci. Total Environ.* **2019**, *658*, 69–79. [[CrossRef](#)] [[PubMed](#)]
73. Davies, W.J.; Bruce, N.S.; Murphy, J.E. Soundscape reproduction and synthesis. *Acta Acust. United Acust.* **2014**, *100*, 285–292. [[CrossRef](#)]
74. Dimoulas, C.; Vegiris, C.; Avdelidis, K.; Kalliris, G.; Papanikolaou, G. Automated audio detection, segmentation, and indexing with application to postproduction editing. In Proceedings of the 122nd Audio Engineering Society Convention, Vienna, Austria, 5–8 May 2007.
75. Vegiris, C.; Dimoulas, C.; Papanikolaou, G. Audio Content Annotation, Description, and Management Using Joint Audio Detection, Segmentation, and Classification Techniques. In Proceedings of the 126th Audio Engineering Society Convention, Munich, Germany, 7–10 May 2009.