electronics

MDPI

Article

Audio Feature Engineering for Occupancy and Activity Estimation in Smart Buildings

Gabriela Santiago [1], Marvin Jiménez [2], Jose Aguilar [1,3,4,*] and Edwin Montoya [4]

[1] Centro de Microcomputación y Sistemas Distribuidos (CEMISID), Universidad de Los Andes, Mérida 5101, Venezuela; gabrielas@ula.ve
[2] Departamento de Ingeniería Industrial, Universidad del Sinú, Montería 230029, Colombia; marvinjimeneznarvaez@gmail.com
[3] Universidad de Alcalá, Departamento de Automática, 28805 Alcalá de Henares, Spain
[4] Grupo de Investigación, Desarrollo e Innovación en Tecnologías de Información y Comunicación (GIDITIC), Universidad EAFIT, Medellín 50022, Colombia; emontoya@eafit.edu.co
* Correspondence: jose.aguilar@uah.es or aguilar@ula.ve or jlaguilarc@eafit.edu.co

Abstract: The occupancy and activity estimation are fields that have been severally researched in the past few years. However, the different techniques used include a mixture of atmospheric features such as humidity and temperature, many devices such as cameras and audio sensors, or they are limited to speech recognition. In this work is proposed that the occupancy and activity can be estimated only from the audio information using an automatic approach of audio feature engineering to extract, analyze and select descriptors/variables. This scheme of extraction of audio descriptors is used to determine the occupation and activity in specific smart environments, such that our approach can differentiate between academic, administrative or commercial environments. Our approach from the audio feature engineering is compared to previous similar works on occupancy estimation and/or activity estimation in smart buildings (most of them including other features, such as atmospherics and visuals). In general, the results obtained are very encouraging compared to previous studies.

Keywords: audio feature engineering; acoustic features; activity estimation; occupancy estimation; smart buildings

Citation: Santiago, G.; Jiménez, M.; Aguilar, J.; Montoya, E. Audio Feature Engineering for Occupancy and Activity Estimation in Smart Buildings. *Electronics* **2021**, *10*, 2599. https://doi.org/10.3390/electronics10212599

Academic Editor: Maria D. R-Moreno

Received: 22 September 2021
Accepted: 22 October 2021
Published: 24 October 2021

1. Introduction

Most research about occupancy is related to minimize the energy consumption in smart environments, so they use mainly atmospheric conditions data to measure the energy produced. Based on the previous work of Jimenez et al. [1], this article proposes the research in occupancy and activity estimation for smart buildings using audio information. Works such as [2,3] use audio information from statistical theory and sound engineer, which include duration, frequency, loudness and sonority, among others, to extract useful information that can be interpreted. In [4,5], Huang and Hu et al. use audio information to determine occupancy in given spaces, but they use additionally speaker recognition or visual data.

To investigate the problem of descriptor extraction for sound content, in [1] Jimenez et al. present the extraction of audio descriptors from the time series theory [6], that is, considering the audios as a set of time series, to use these time series characteristics as audio descriptors. The developed approach allows the analysis and selection of descriptors from a given audio context, with a hybrid scheme of extraction of those audio descriptors based on sound variables, descriptive statistics or time series. It also defines metric-based audio descriptors from the time series domain, allowing the comparison with other audio descriptor extraction techniques, in classification and clustering.

Considering this as a start point, in this work is presented an automatic scheme of feature engineering to determine occupation in smart environments from audio information, which also is used to determine the activities in spaces according to specific acoustic

features: dominant frequency (in Hz), loudness (in dB) and reverberation time (in secs). The system will use the audio descriptors to determine the occupation in a smart environment, and depending on the type of building (academic, administrative or commercial), it is possible to connect the acoustic features with the context to obtain as result the activity that may be taking place.

Thus, the main contribution of this work is to propose an autonomous approach to estimate the occupation and activity in a smart building, considering only acoustic descriptors, extracted automatically using a characteristic engineering process in the environment. This article is organized as follows: related works are presented in Section 2. Section 3 contains the theoretical framework where concepts related to acoustic features and the audio descriptors are explained. In Section Section 4, the approach to occupancy and activity estimation is presented. Section 5 presents the experiments and results for the estimation of occupation and activity, with a description of the experimental protocols and a comparison with previous works. Finally, the conclusions and future works are presented.

2. Related Works

There are in the literature research that makes contributions about occupancy estimation in buildings, as well as in the definition of audio descriptors. For example, in the work of [7], Nasir et al. use data analytics and sensors to define an occupant detection system that monitors the number of occupants in a space, which detected/inferred information helps with better energy management. In addition, Afuosi et al. [8] present a signal processing technique and noise-suppressing algorithm to minimize the acoustic reflection, refraction and diffraction for indoor positioning.

On the other hand, Rana et al. [9] present a sensor fusion method that uses the phone microphone and accelerometer as feed to determine occupancy and activity. They obtain 90% accuracy in estimating occupancy and 95% accuracy in their activity classification algorithm. In [10], Huang implements a combination of hybrid sensors (CO_2 and light sensors) to achieve a more accurate occupancy estimation, in a scheme that is non-intrusive and low cost.

In activity recognition, in [11,12] Zou et al. present a device-free human activity recognition scheme that distinguishes activities using WiFi-enabled IoT devices. Their experimental results demonstrate high recognition accuracy of common human activities.

In the field of audio descriptors and extraction of audio features, in [13] Wülfing et al. use a characteristic learning method to automatically predict music genres, with local patches from the time-frequency-transformed audio signal. Something similar is presented by Costa et al. [14], where the audio signal is converted into spectrograms to compare two sets of different textural characteristics for the automatic classification of the musical genre.

Beyond the scope of the music scenario, in [15] Muaidi et al. propose an architecture called ARANEWS for the news in the Arabic language, using automatic voice recognition to isolate Arabic. This retrieval system is based on modelling signal waves and measuring the similarity between features.

These works establish different perspectives on occupancy estimation, activity estimation and sound descriptors, using different elements and sources. This work aims at using only audio information to determine occupancy and activity with an acceptable level of accuracy. In general, few works simultaneously combine the estimation of occupation and activities, and even less using only sound information.

3. Theoretical Framework

In this section are described the terms related to acoustic features engineering and audio descriptors for smart environments, which will sustain our approach to occupancy and activity estimation.

3.1. Acoustic Features

In order to work with audio datasets, it is important to understand basic acoustic features that will enhance the use of the information that can be obtained from audio files. According to Tahir et al. [16], there are acoustic features that can be mainly useful for automatic speech recognition, but for this research, the focus will be on three of these basic features that can be extracted in an audio file: dominant frequency, loudness and reverberation time.

The *dominant frequency* is the frequency that carries the maximum energy among all frequencies found in an acoustic spectrum [17]. It is expressed in Hertz (Hz). It is a similar term to the fundamental frequency, which is the smallest frequency having a peak among all frequencies in a spectrum. Its determination allows the understanding of the structure of time series, and recognizing the dominant frequency in data analysis, which leads to more accurate predictions and better engineering designs.

As explained by Schmidt et al. [18], *loudness* is an auditory measure of perception, and it is the perceived intensity of a sound. It is usually expressed in decibels (dB). It depends mainly on the sound acoustic pressure, but also on its frequency, wavelength and duration [19].

The *reverberation time* is the time (T_{60}) takes an acoustic impulse response energy to decay 60 dB [20]. It is measured in seconds (sec). In an ideal case, there is a linear relationship between the sound pressure level and time, being background noise low enough [19].

These concepts will be used with our estimation approach to determine different activities that take place in a given smart environment.

3.2. Audio Features Engineering Procedures for Smart Environments

This methodology is based on [1] and is as follows:

3.2.1. Input

- Audio datasets: an audio dataset of smart environments with different occupancy levels.
- Descriptors: There are three groups of descriptors/variables. The first one with general audio features (acoustic descriptors), such as acoustic complexity index, resonant quality factor, etc. The second one with statistics based on short-term Fourier transform (STFT) time and frequency contours; and finally, the third one with statistical descriptors of time series. Six variables will be measured over time: instant frequency, fundamental frequency, dominant frequency, Hilbert amplitude, spectral entropy and entropy flux. For each of these variables are obtained time series descriptors such as: trend force, seasonality force, linearity, nonlinearity, entropy. For more details see [1].

3.2.2. Process

- Audio features extraction: in this stage, the descriptors are extracted from each audio file from the dataset, and this information is stored in a data frame with a row per audio and a column per feature.
- Selection of the best feature combination: Here is identified the subset of descriptors most suitable for estimating occupancy. This selection is made by means of a search algorithm, which is guided by a prediction quality metric of the predictive model used. The best combination of descriptors is determined by optimizing a mathematical function, which corresponds to the occupancy prediction quality metric.

3.2.3. Output

- Best subset of features. The subset of descriptors with the best predictive quality is obtained.

A summary of the approach is presented in Figure 1. For more details of the feature engineering methodology see [1].

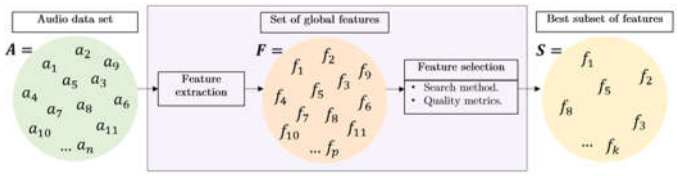

Figure 1. General process of our approach.

4. Approaches to Occupancy and Activity Estimation

In this section is presented the response module that indicates the occupancy and activity estimation in a given space based on acoustic features of dominant frequency, loudness and reverberation time. The smart environments are divided into academics, administrative and commercials. Given the particular activities that take place in the different scenarios, it is not correct to assume that the occupancy and activities are relative.

4.1. Occupancy Estimation

The occupancy on all three smart environments is going to be divided into low, medium, high and overcrowded (see in Tables 1–3 the occupation definition for three different smart environments, respectively). Our approach will indicate if the occupancy is low, medium, high or if it is overcrowded based on the following data:

- Academic environment.

Table 1. Occupancy in an academic environment.

Occupancy	Low	Medium	High	Overcrowded
People	0–10	10–50	50–200	+200

- Administrative environment

Table 2. Occupancy in an administrative environment.

Occupancy	Low	Medium	High	Overcrowded
People	0–25	25–50	50–100	+100

- Commercial environment

Table 3. Occupancy in a commercial environment.

Occupancy	Low	Medium	High	Overcrowded
People	0–50	50–500	+500	Non-considered

In our approach, in an academic environment is considered that a low occupancy is between 0 and 10 people (see Table 1), while in an administrative environment this number goes from 0 to 25 (see Table 2), and in a commercial environment from 0 to 50 (see Table 3). A medium occupancy in an academic environment is between 10 and 50, in administrative 25 to 50, and in commercial 50 to 500. The high occupancy in academic places is between 50 and 200, in administrative is between 50 and 100, and in commercial is at more than 500 people.

An academic place would be overcrowded with more than 200 people, an administrative place with more than 100, but a commercial place might not be considered as overcrowded.

This conception corresponds to the known uses that these places offer, which are going to be specified in the next section.

4.2. Activity Estimation

For the activity estimation, it is necessary to consider the acoustic features described in Section 3: dominant frequency, loudness and reverberation time. The following tables show how our approach will indicate the possible activities that are taking place according to the level of occupancy and acoustic features that can be separately extracted (see in Tables 4–6 the activities for academic, administrative and commercial environments, respectively).

- Academic environment

Table 4. Activity in an academic environment.

Occupancy	Acoustic Feature	Activity
Low	Normal level of loudness	Academic tutoring/private classes
Medium	Long reverberation time and low level of loudness	Plenary lecture
High	High dominant frequencies and high level of loudness	Guided visit from kids or teenagers in a university for special events
Overcrowded	Medium level of loudness	Congress/symposium

- Administrative environment

Table 5. Activity in an administrative environment.

Occupancy	Acoustic Feature	Activity
Low	High dominant frequencies and high level of loudness	Child-related activities (i.e., special schedules to obtain the ID or passport of children)
Medium	Medium dominant frequencies and high level of loudness	Normal business operations
High	Medium dominant frequencies and high level of loudness	Business operations after holidays or long weekends
Overcrowded	High level of loudness	Special schedule (i.e., for renewal of documents in immigration office)

- Commercial environment

Table 6. Activity in a commercial environment.

Occupancy	Acoustic Feature	Activity
Low	Long reverberation time, low levels of loudness and low dominant frequencies	Cleaning and maintenance (i.e., the shopping mall is closed)
Medium	High dominant frequencies and high level of loudness	Circus/child related fun activities
High	Low levels of loudness	Movie theater (a movie about to start)
High	High levels of loudness	Food fair/lunch time/dinner Time

As shown in Figure 2, the occupancy low, medium, high or overcrowded varies depending on the activities, which can be determined from the acoustic features in each environment. At the same time, the different uses of academic, administrative or commercial environments show that the same occupancy and the same acoustic features can lead to completely different activities.

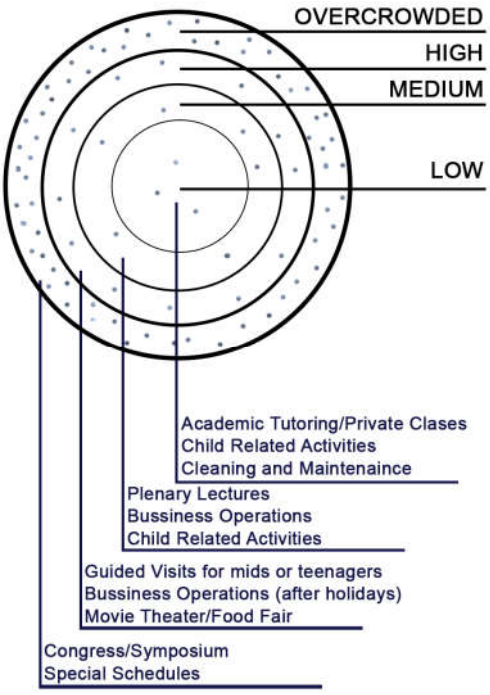

Figure 2. Relation between occupancy and activities.

5. Experiments and Results

5.1. Estimation Experimental Protocols

An experiment was conducted to evaluate the performance of the proposed feature engineering methodology in smart environment occupancy estimation tasks. The dataset used was the audiovisual crowd counting (DISCO), which was presented in [5]. The DISCO dataset contains 1935 images and their corresponding audio clips, with a person count that goes from 0 to 709. The dataset is divided into three sets: 1435 files for training, 200 files for validation and 300 files for testing. In this work, we only use audio information.

The estimation of occupancy is performed by fitting a regression model in which the response variable is a count, and the predictor variables are the audio descriptors. Three different techniques were implemented for the prediction of occupancy, namely: generalized linear model with Poison response (GLM), random forests (RF) and XGBoost. The GLM is considered in this work because it is a technique specifically designed for cases where the response variable is a count. RF and XGBoost are algorithm ensemble techniques that are included in the comparison because they are widely used techniques in recent literature due to their good results.

The metric used to evaluate and compare the occupancy prediction quality of the fitted models was the mean absolute error (MAE), which is calculated using Equation (1).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, \tag{1}$$

where N is the total number of audios in the evaluated dataset, y_i and $\hat{y}_i$ are the actual number of people and the number predicted by the model, respectively.

To take a first look at the performance of the algorithms, each algorithm was trained with the training dataset using all the descriptors previously extracted from the audio dataset. The quality of the predictions was then evaluated with the training, validation and test datasets.

Subsequently, for each model, the selection process of the best descriptors was executed using a genetic algorithm (GA), which used the MAE of the validation set as the evaluation function for each chromosome (combination of descriptors).

5.2. Estimation Result Analysis

After the extraction of the features, a data frame of 1935 rows and 324 columns (the count of persons and 323 descriptors) was defined. A brief exploratory analysis of the data was performed afterward to identify those descriptors that are very homogeneous throughout the training dataset. These features with very low variability do not make a significant contribution to the characterization of the audios, so they are not very useful for predicting the occupancy of the environment.

Table 7 presents the mean, standard deviation (SD) and coefficient of variation (CV) for six of the descriptors. As can be seen, the descriptors t.p2 and f.pre have a rather low variability, so they should be removed from the data frame.

Table 7. Descriptive statistics of six descriptors.

Feature	Mean	SD	CV (%)
Aci	152.926	4.621	3.022
t.p1	0.046	0.010	21.296
t.m	0.500	0.038	7.660
t.p2	0.954	0.010	1.040
f.m	1.409	0.675	47.895
f.preci	93.750	0.000	0.000

In addition, the descriptors have notable differences in scale. Therefore, each of the techniques was evaluated for three different types of scaling: no scaling, min–max scaling and standard scaling, and for different CV thresholds to remove descriptors with very low variability, i.e., descriptors with a CV less than or equal to the CV threshold were removed. Table 8 shows the MAE for the random forest algorithm using all descriptors.

Table 8. MAE for different scales and CV thresholds for RF.

Scale	CV.Thre	MAE		
		Training	Validation	Test
No Scaling	1	21.23	56.62	54.71
	2	21.17	57.15	54.33
	3	21.37	56.56	54.51
	4	21.17	56.33	55.05
	5	21.17	56.33	55.05
Min–Max Scaling	1	21.21	57.00	54.80
	2	21.14	56.41	54.16
	3	21.39	55.98	55.20
	4	21.15	57.59	55.63
	5	21.15	57.59	55.63
Standard Scaling	1	20.99	57.18	54.28
	2	21.41	57.35	54.83
	3	21.38	56.54	54.75
	4	21.34	57.69	55.18
	5	21.34	57.69	55.18

As presented in Table 8, the type of scaling and the CV threshold have no significant influence in the MAE of any of the data subsets. This situation is also present for GLM and XGBoost.

Therefore, in the experimentation of this work, it was determined to use Min–Max scaling and to remove the descriptors with a CV (CV.thre) less than or equal to 5% to reduce

the dimensionality of the data frame. The CV.thre is a preprocessing parameter that can be adjusted. Thus, 31 descriptors were removed.

Table 9 presents the MAEs for the three partitions of the dataset for the three techniques implemented using all the descriptors. Clearly, the XGBoost algorithm presents far better prediction quality on the training data than the other techniques. However, on the validation dataset, its performance is very similar to that of the RF algorithm, they do not have significantly different MAEs but XGboost is much faster. In the test dataset, XGBoost wins again, and the second best is RF.

Table 9. MAE for the three techniques for the three partitions of the dataset.

Techniques	MAE		
	Training	Validation	Test
GLM (0.55 s)	46.14	63.17	64.62
RF (38.7 s)	21.15	57.59	55.63
XGBoost (1.02 s)	7.00	57.43	49.06

Finally, the best subset of descriptors for each technique was selected using a GA. The tournament selection method, a population size of 100 individuals, a sample size of 10 (population), a crossover rate of 0.92, a mutation rate of 0.1 and a total of 100 generations were used.

The behavior of MAEs over generations of individuals evaluated with the GA is depicted in Figure 3. The XGBoost algorithm shows no improvement as the GA combines descriptors with the goal of minimizing the MAE of the predictions. The best subset of descriptors from the last generation of the GA produces an MAE equal to that obtained with the starting set of descriptors.

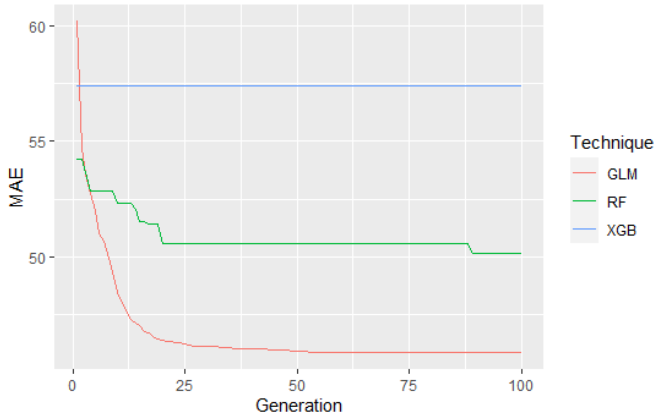

Figure 3. Evolution of MAEs over generations with the GA.

On the other hand, the GLM and RF techniques show a decrease in the mean absolute error over the evolution of the solution population. In particular, GLM shows a higher rate of improvement, even though it starts with the highest MAE, it finds a combination of descriptors that has a lower MAE (45.82) than RF (50.56) and XGBoost (57.43) in the final generation. GLM achieves this MAE by combining only 161 descriptors out of the initial 292.

Considering that the DISCO dataset contains audios from commercial environments, the information in Table 3 was used to classify the actual occupancy and the occupancy estimated by the GLM selected by the GA into low, medium and high categories. Table 10 shows the confusion matrix of the model, obtaining an accuracy of 0.69.

Table 10. Confusion matrix of the GLM technique after applying the GA.

		Actual		
		Low	**Medium**	**High**
Prediction	**Low**	43	19	0
	Medium	44	93	1
	High	0	0	0

Since the occupancy level has three categories, it is necessary to analyze them individually to obtain measures such as sensitivity or precision. Thus, as an example, the binary confusion matrix for the medium category is presented in Table 11.

Table 11. Confusion matrix for "medium" level occupancy.

		Actual	
		Positive	**Negative**
Prediction	**Positive**	93	45
	Negative	19	43

With the information in Table 11, a precision of 0.674, a recall of 0.83 and an F1 score of 0.744 are calculated for the prediction of the medium class. Calculating the confusion matrix for the low class, the values 0.69, 0.49 and 0.577 are obtained for the precision, recall and F1 scores, respectively. This shows that the model has a better performance detecting the medium class than the low class, i.e., it works better identifying medium occupancy levels of commercial environments than low levels.

Since the high class has only one positive case in the data set evaluated and was not correctly predicted by the model, only the precision can be calculated, which has a value of zero. For this reason, it is not prudent to make claims about the ability of the model to detect high occupancy levels in commercial environments.

5.3. Activity Result Analysis

To link the experimentation to the activity estimation, we are using four examples of audio information in commercial buildings comparing the original value from the DISCO dataset and the estimated value from our model.

Table 12 shows the original value, the estimated value from our model, and the absolute error. Some of the descriptors used to obtain this estimation are the median frequency (f.med), the flatness of a frequency spectrum (sfm), the stability, the nonlinearity, the max level shift, and the normalized difference soundscape index (ndsi).

Table 12. Comparison between original values and the value of our model with estimated occupation and activity.

Original Value	22	44	114	258
Estimated Value	22,847	45,090	113,667	257,357
Absolute Error	0.847	1.090	0.332	0.642
Occupancy Estimation (Table 3)	Low	Low	Medium	Medium
Descriptors	f.med/sfm/ stability/nonlinearity/ max-level-shift/ndsi	f.med/sfm/ stability/nonlinearity/ max-level-shift/ndsi	f.med/sfm/ stability/nonlinearity/ max-level-shift/ndsi	f.med/sfm/ stability/nonlinearity/ max-level-shift/ndsi
Activity Estimation (see Table 6)	Cleaning and Maintenance	Cleaning and Maintenance	Normal Operations/ Child Activities	Normal Operations/ Child Activities

These descriptors are related to the acoustic features explained in Section 3.A, and they play a role in the identification of loudness, dominant frequency and reverberation time that help with the activity estimation.

5.4. Comparison with Previous Works

In this work, it is important to obtain as much information as possible from the buildings only using audio information. Audio descriptors provide a lot of information for estimating occupancy and activities that may be taking place. Using other elements, such as speaker recognition and visual data, among others, can improve the results, but when no other device is available, the audio information also gives a good prediction. Table 13 presents an overview of the different elements used by other systems and our approach for the occupation-activity estimation. Particularly, our approach used building classification and audio information.

Table 13. Comparison with other works.

Work	[4]	[5]	[9]	[10]	This Work
Building Classification		X			X
Speaker recognition	X				
Visual Data		X			
Hybrid Sensors			X	X	
Audio Descriptors Only					X

The main contribution of our work is to determine the occupation and activity in a building having as previous information just the type of building (academic, administrative, commercial) and using only audio descriptors.

In [4], Huang uses also speaker recognition, while in [5] Hu et al. propose mixing image descriptors with audio descriptors to improve the quality of occupancy estimates obtained with image descriptors alone. Their results show that their approaches produce good quality predictions. In [9,10], a mixing of sensors also benefits the predictions.

In this paper, we propose an audio descriptor engineering methodology for occupancy estimation using time series descriptors and compare different prediction techniques. The MAEs obtained are higher than the best recorded with the approach proposed in [5], but our work is innovative because of the way the audio characterization is approached, the possibility to find the best combination of descriptors, and the low computational cost due to the low number of descriptors. This approach can be enriched by introducing concepts such as by mixing resolution techniques with distributed strategies such as agent theory [21] or artificial immune systems [22], or with autonomous cycles of data analysis tasks [23].

6. Conclusions

This work proposed an approach to estimate occupation and activity in smart buildings, classifying the environments in academic, administrative and commercial. Each environment has a different consideration for the number of people (low, medium, high, overcrowded), and according to this, it is possible to determine the activity taking place in that environment.

While other works use information from different sources (speaker recognition, visual data, hybrid sensors), our approach only uses audio descriptors, which is useful in most cases where the only information available is auditory. That is the main contribution of our work, we propose an estimation approach with acceptable accuracy from audio information.

The comparison with previous works shows the potential of our approach, so it is expected that adding speaker recognition, image descriptors and different sensors will significantly improve the quality of occupancy and activity estimates.

Although the results are encouraging, there is still room for more improvements. This is particularly visible in the results we obtained with the occupancy classes in commercial buildings (see Table 10). In this sense, future works will seek to determine the minimum number of descriptors from sources other than sound, which, when combined with the sound descriptors with which we obtained the best results, allow obtaining high levels of quality in the metrics studied. This will enable minimizing the number of devices, other than the sound devices required.

Future works involve the improvement of the results by reducing the MAE in the predictions and using different scales for the input data, aiming at exploiting all the possibilities of using audio information for a more accurate estimation of occupation and activity in smart buildings.

Author Contributions: Conceptualization, G.S., M.J. and J.A.; methodology, G.S., M.J. and J.A.; software, M.J.; validation, G.S., M.J. and J.A.; investigation, G.S, M.J. and J.A.; resources, J.A. and E.M.; data curation, G.S. and M.J.; writing—original draft preparation, G.S., M.J. and J.A.; writing—review and editing, G.S., M.J. and J.A.; funding acquisition, J.A. and E.M. All authors have read and agreed to the published version of the manuscript.

Funding: This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie Grant Agreement No 754382 GOT ENERGY TALENT.

Conflicts of Interest: The authors declare no conflict of interest. The content of this publication does not reflect the official opinion of the European Union. Responsibility for the information and views expressed herein lies entirely with the author(s).

References

1. Jiménez, M.; Aguilar, J.; Monsalve-Pulido, J.; Montoya, E. An automatic approach of audio feature engineering for the extraction. analysis and selection of descriptors. *Int. J. Multimed. Inf. Retr.* **2021**, *10*, 33–42. [CrossRef]
2. Moffat, D.; Ronan, D.; Reiss, J. An evaluation of audio feature extraction toolboxes. In Proceedings of the 18th International Conference on Digital Audio Effects, Trondheim, Norway, 30 November–3 December 2015.
3. Pearce, A.; Brookes, T.; Mason, R. Timbral attributes for sound effect library searching. In Proceedings of the AES International Conference on Semantic Audio, Erlangen, Germany, 22–24 June 2017.
4. Huang, Q. Occupancy-Driven Energy-Efficient Buildings Using Audio Processing with Background Sound Cancellation. *Buildings* **2018**, *8*, 78. [CrossRef]
5. Hu, D.; Mou, L.; Wang, Q.; Gao, J.; Hua, Y.; Dou, D.; Zhu, X. Ambient Sound Helps: Audiovisual Crowd Counting in Extreme Conditions. *arXiv* **2020**, arXiv:2005.07097.
6. Kang, Y.; Hyndman, R.J.; Li, F. *Efficient Generation of Time Series with Diverse and Controllable Characteristics*; Monash Econometrics and Business Statistics Working Papers; Monash University, Department of Econometrics and Business Statistics: Melbourne, Australia, 2018.
7. Nasir, N.; Palani, K.; Chugh, A.; Prakash, V.; Arote, U.; Krishnan, A.; Ramamritham, K. Fusing sensors for occupancy sensing in smart buildings. In Proceedings of the International Conference on Distributed Computing and Internet Technology, Bhubaneswar, India, 5–8 February 2015; pp. 73–92.
8. Afuosi, M.B.; Zoghi, M. Indoor positioning based on improved weighted KNN for energy management in smart buildings. *Energy Build.* **2020**, *212*, 109754. [CrossRef]
9. Rana, R.; Kusy, B.; Wall, J.; Hu, W. Novel activity classification and occupancy estimation methods for intelligent HVAC (heating, ventilation and air conditioning) systems. *Energy* **2015**, *93*, 245–255. [CrossRef]
10. Huang, Q.; Mao, C. Occupancy estimation in smart building using hybrid CO2/light wireless sensor network. *J. Appl. Sci. Arts* **2017**, *1*, 5.
11. Zou, H.; Zhou, Y.; Yang, J.; Spanos, C. Towards occupant activity driven smart buildings via WiFi-enabled IoT devices and deep learning. *Energy Build.* **2018**, *177*, 12–22. [CrossRef]
12. Zou, H.; Zhou, Y.; Yang, J.; Spanos, C. Device-free occupancy detection and crowd counting in smart buildings with WiFi-enabled IoT. *Energy Build.* **2018**, *174*, 309–322. [CrossRef]
13. Wülfing, J.; Riedmiller, M. Unsupervised learning of local features for music classification. In Proceedings of the 13th International Society for Music Information Retrieval Conference, Porto, Portugal, 8–12 October 2012; pp. 139–144.
14. Costa, Y.; Oliveira, L.; Koerich, A.; Gouyon, F. Comparing textural features for music genre classification. In Proceedings of the International Joint Conference on Neural Networks, Brisbane, Australia, 10–15 June 2012.

15. Muaidi, H.; Al-Ahmad, A.; Khdoor, T.; Alqrainy, S.; Alkoffash, M. Arabic audio news retrieval system using dependent speaker mode, mel frequency cepstral coefficient and dynamic time warping techniques. *Res. J. Appl. Sci. Eng. Technol.* **2014**, *7*, 5082–5097. [CrossRef]
16. Tahir, M.; Huang, H.; Zeyer, A.; Schlüter, R.; Ney, H. Training of reduced-rank linear transformations for multi-layer polynomial acoustic features for speech recognition. *Speech Commun.* **2019**, *110*, 56–63. [CrossRef]
17. Chen, Y.; Li, H.; Hou, L.; Bu, X. Feature extraction using dominant frequency bands and time-frequency image analysis for chatter detection in milling. *Precis. Eng.* **2019**, *56*, 235–245. [CrossRef]
18. Schmidt, F.H.; Mauermann, M.; Kollmeier, B. Neural representation of loudness: Cortical evoked potentials in an induced loudness reduction experiment. *Trends Hear.* **2020**, *24*, 2331216519900595. [CrossRef] [PubMed]
19. Sociedad Española de Acústica (Ed.) *Glosario de Términos Acústicos*; Sociedad Española de Acústica: Madrid, Spain, 2012.
20. Gamper, H.; Tashev, I. Blind reverberation time estimation using a convolutional neural network. In Proceedings of the 16th International Workshop on Acoustic Signal Enhancement, Tokyo, Japan, 17–20 September 2018; pp. 136–140.
21. Aguilar, J.; Cerrada, M.; Mousalli, G.; Rivas, F.; Hidrobo, F. A Multiagent Model for Intelligent Distributed Control Systems. *Comput. Vis.* **2005**, *3681*, 191–197. [CrossRef]
22. Araújo, M.; Aguilar, J.; Aponte, H. Fault detection system in gas lift well based on artificial immune system. In Proceedings of the International Joint Conference on Neural Networks, Jantzen Beachm, Portland, OR, USA, 20–24 July 2003; pp. 1673–1677.
23. Sanchez, M.; Exposito, E.; Aguilar, J. Autonomic computing in manufacturing process coordination in industry 4.0 context. *J. Ind. Inf. Integr.* **2020**, *19*, 100159. [CrossRef]