Framework for Acoustic Comfort Analysis in Digital Twins

Hossein Pouresmaeeliasiabar[1] and Ali Motamedi[2]

[1] École de technologie supérieure, Université du Québec, Canada
[2] École de technologie supérieure, Université du Québec, Canada
hossein.pouresmaeeliasiabar.1@ens.etsmtl.ca, ali.motamedi@etsmtl.ca

Abstract –

Acoustic comfort, a critical yet often overlooked aspect of indoor environmental quality, plays a significant role in occupant health and productivity. Unlike other comfort dimensions, such as thermal or lighting, measuring acoustic comfort remains challenging due to its subjective nature and the complex interplay of physiological and psychological factors. Current approaches to assessing acoustic comfort in indoor environments often overlook the content of sound, despite its potential to be a decisive factor. For instance, the perceived comfort of listening to music at high sound levels differs significantly from that of hearing construction noise, even at lower sound levels. This research proposes a novel framework for integrating acoustic comfort analysis in the digital twin environment, which comprises psycho-acoustic metrics, sound event classification, and predictive analytics. The implemented system leverages sensor data, a sound event classification neural network, and advanced visualization methods to enable real-time and historical acoustic analysis. Privacy concerns are addressed through a privacy-by-design approach, ensuring data security by processing audio on the edge devices without storing raw sound. A case study in an office environment demonstrates the framework's effectiveness in monitoring and improving acoustic conditions. Microphones connected to edge devices classify sound events and calculate soundwave parameters such as relative sound pressure levels while integrating results into the digital twin.

Keywords –Digital Twins; Acoustic comfort; Psycho-acoustic metrics; Sound event classification; Acoustic monitoring; Comfort analysis

1 Introduction

There exist numerous definitions for the term “comfort” in various domains such as AEC-FM (architecture, engineering, construction, and facilities management) and healthcare. The common theme in the definitions of comfort in the built environment refers to it as a state of physical and psychological ease, characterized by the absence of discomfort or distress. Although this high-level definition understandably describes comfort, it fails to adequately pinpoint its defining measurable characteristics. The notion of comfort in buildings can be viewed in multiple distinct but interrelated aspects, such as thermal comfort, acoustic comfort, indoor air quality, and lighting comfort; each of which depends on various variables. However, the level of perceived comfort differs for each person, depending on the physiological and psychological state at that very moment in time. Maintaining comfort is a fundamental need in buildings; the indoor environment quality (IEQ) performance of a building, is a decisive factor in the health, productivity, and well-being of its occupants, and can be a determinant factor in lifecycle costs and energy consumption [1].

Among the different aspects of comfort, acoustic comfort received less attention [2]. A simple search on academic databases related to indoor comfort reveals a substantial disparity in the number of studies focusing on different aspects of comfort, with significantly fewer studies focusing on acoustic comfort compared to other aspects, such as lighting, thermal, and air quality. Acoustic comfort is generally defined as “a state of contentment with acoustic conditions” [3], and it can play a pivotal role in determining the overall comfort of an environment. Furthermore, it is interrelated to other aspects of comfort and the building's overall energy consumption. For instance, occupants may face the trade-off between the generated HVAC noise by the cooling system and thermal discomfort [4].

Similar to other aspects of comfort, measuring acoustic comfort is complex. Sound perception is subjective and varies among individuals based on their experiences, cultural backgrounds, and personal preferences [5]. Additionally, it highly depends on the psychological state of the occupants, which makes its objective assessment more difficult or impossible. Although the physical characteristics of sound waves can be adequately measured, there is no measurable characteristic showing the overall psychological state of a human being and how comfortable a person is while being exposed to acoustic stimuli. As a result, a certain

sound pattern may seem uncomfortable or annoying to a person and can be ignored or even considered pleasant by another person [6].

To estimate the acoustic comfort level of a space, traditionally, some of the basic variables of the acoustic waveform are measured. Such variables include the sound pressure levels (dB), the frequency (Hz), and the time that a person is exposed to the sound. Although these variables are essential for estimating acoustic comfort, and some regulations and guidelines are already developed recommending their acceptable ranges, "acoustic comfort" is by far more complex to measure. For example, the human hearing system does not respond to all the frequencies, and the sound pressure levels equally [7]; hence, the sounds with some frequencies may be perceived as louder than others. This shows the inadequacy of assessing acoustic comfort using only the sound (acoustic) pressure level.

Consequently, psycho-acoustic indicators such as loudness, A-weighted sound pressure level, sharpness, roughness, fluctuation strength, articulation index, and impulsiveness were introduced[8, 9]. These psycho-acoustic indicators consider the human hearing system in perceiving sounds, resulting in a better estimation of acoustic comfort. For example, to have a better understanding of "how loud a sound is", loudness is defined which is the property of the sound that can be "ordered on a scale extending from soft to loud" [10].

However, even the existing psycho-acoustic indicators would not be sufficient to adequately assess the acoustic comfort for a specific space. For example, although there might be a similarity between calculated psycho-acoustic indicators of audio patterns of certain musical pieces and construction tools (e.g., jackhammer), the perceived sensation for the occupants would be completely different. The feeling of acoustic comfort for occupants who are exposed to a musical piece heavily depends on their musical taste, whereas the noise of the construction tool can be unpleasant for all occupants.

The inadequacy of such metrics and indicators can be linked to their exclusive reliance on sound wave parameters, without accounting for the content or type of the sound. Consequently, researchers have sought to develop new psychoacoustic indicators and sound perception descriptors, such as satisfaction, dissatisfaction, noisiness, and pleasantness, as highlighted by Hossain et al. [11]. The recent advancement of the Deep-learning-based models allows a near real-time analysis and classification of sound events. Combining the classic metrics, psycho-acoustic indicators, and sound event categorization provides better monitoring and estimation of acoustic comfort for the built environment. Additionally, such a combinatory effect allows for defining customized indications of the comfort level considering individual occupants' preferences.

This research aims to provide a framework to perform acoustic comfort analysis by combining various metrics and indicators together with sound event category identification within a digital twin. The proposed framework comprises the integration of sensor data within a digital twin solution, the calculation of psycho-acoustic indicators, the near-real-time identification of sound event categories, and the predictive analytics capability. Additionally, the framework adopts the privacy-by-design approach in which the stored data cannot be used to recreate the original captured sound.

The remainder of this paper presents a brief review of the literature on acoustic environmental assessment, the application of psychoacoustic indicators, and the role of Sound Event Detection (SED) in such evaluations. This is followed by a detailed explanation of the proposed framework and its potential applications. A case study implementation is then described, along with the corresponding results and conclusions.

2 Related Work

Literature on assessing the acoustic suitability of environments reveals two interrelated yet distinct concepts: soundscape and acoustic comfort. While both address the auditory experience, their focus and applications differ. Acoustic comfort primarily pertains to enclosed spaces, emphasizing the reduction of annoyance and enhancement of auditory satisfaction within confined environments. In contrast, the concept of soundscape is broader, encompassing urban, natural, and built environments. It explores the dynamic interaction between environmental sounds and human experiences.

Both soundscape and acoustic comfort assessments rely on a combination of objective and subjective parameters. For example, Aletta et al. [12] emphasize the role of psychoacoustic indicators, such as loudness, sharpness, roughness, and fluctuation strength, in soundscape studies. Similarly, Engel et al. [13] conducted a systematic review of 46 peer-reviewed studies, demonstrating the widespread application of psychoacoustic indicators in soundscape research. These findings highlight the critical role of such indicators in capturing human perceptions of acoustic environments, thereby enhancing the accuracy and relevance of soundscape evaluations. Likewise, psychoacoustic indicators are integral to assessing acoustic comfort. For instance, the relationship between objective acoustic measurements and subjective responses to evaluate acoustic comfort in residential settings is investigated in [14]. Their study underscores the importance of indicators addressing low-frequency noise to better correlate technical data with perceived annoyance, thereby emphasizing the need to integrate human

perception into acoustic comfort assessments.

A review of the literature reveals the extensive use of psychoacoustic indicators in assessing acoustic environments. Zhang et al. [6] developed a high-precision model for predicting acoustic comfort in electric buses by integrating psychoacoustic indicators with machine learning techniques. Similarly, Herranz-Pascual [15] introduced the Acoustic Comfort Assessment Scale (ACAS-12), a psychometrically validated tool designed to assess acoustic comfort in urban environments. This scale incorporates indicators, such as pleasantness, eventfulness, familiarity, informational capacity, and congruence, demonstrating its reliability and effectiveness in capturing subjective perceptions of acoustic environments. Additionally, Kang et al. [16] proposed a framework for creating soundscape maps in smart cities. Their model predicts perceptual attributes, such as pleasantness and calmness, using sound profiling and linear regression analysis, offering a valuable tool for enhancing urban soundscape design.

Advancements in artificial intelligence have enabled the application of SED in soundscape assessment. For example, Espejo et al. [17] explored the use of short-time acoustic indices in combination with artificial neural networks (ANNs) to monitor and analyze soundscapes in urban-natural environments. Similarly, Bonet-Solà et al. [5] developed a predictive model for assessing acoustic comfort in urban settings. Their approach integrates SED using convolutional neural networks with noise data collected from wireless acoustic sensor networks (WASN), offering a sophisticated method for evaluating and improving urban acoustic environments.

To effectively visualize the acoustic status of a large environment, a 2D color-coded heat map based on soundscape evaluation is a practical approach, as demonstrated by Yue et al. [18]. They developed a visual soundscape prediction model for urban park design, which integrates sound pressure levels, sound source perception, and soundscape evaluation using machine learning techniques. This model provides an intuitive representation of acoustic conditions, facilitating the design and optimization of urban soundscapes.

3 The Proposed Framework

Potential Applications and Use Cases

Having a framework that allows monitoring the present and past values for a collection of psycho-acoustic indicators as well as sound event categories for each room of an indoor environment, facilitates the comfort analysis, prediction, and visualizations through the digital twin web interface. It allows facility managers to better plan for acoustic insulation and activity planning. The occupants can consult the monitoring data to decide if a certain room is suitable in terms of acoustic comfort for their intended use. The stored data allows the application of predictive algorithms to recommend rooms for certain activity types or to predict the acoustic comfort of a certain room in a selected time period in the future. The integrated assessment algorithms can be further enhanced by including other sources of data, such as activity schedules for each room.

The data can be used to identify sources of unwanted noise (such as background noise or noise pollution) and their effect on the neighboring rooms by cross-checking metrics of various neighboring spaces. Additionally, the system can accommodate user inputs to provide personalized metrics and Key Performance Indicators (KPIs) for individuals. Moreover, the system can communicate information about the acoustic comfort preference of neighboring occupants to the occupants of a certain room to adjust their noise-generating activities accordingly. Finally, it can give insights and predictions regarding outdoor noise.

Framework

The framework consists of a five-layered architecture, as depicted in Figure 1 (left). Starting from the bottom of the diagram:

1. In the data acquisition layer, sound is captured by microphones and converted to a digital signal via an analog-to-digital converter (ADC). Using the device-specific interface, the audio signal, in a binary stream format, along with the device-specific metadata such as sampling rate, microphone sensitivity, microphone type, and the location coordinates of the microphone, is prepared for subsequent processing. The data acquisition and data processing layers are distinct; however, they can be implemented on a single device, enabling high-speed data transfer through shared memory or inter-process communication (IPC) interface.
2. In the data processing layer, the audio data stream and device specifications from the data acquisition layer are utilized. This layer processes the input signal according to the requirements of its three modules, based on parameters such as the sampling rate and device specifications. The audio signal is resampled and formatted to meet the input requirements of the selected SED model. Additionally, soundwave parameters and psychoacoustic indicators are computed using the audio signal and accompanying metadata. The processed data is then sent to a cloud-based database to be stored as time-series data, utilizing either REST-based or SQL-based communication interfaces. Depending on the storage layer type,

data transfer can be conducted using GET and POST methods in the REST-based approach or directly via SQL queries.

3. The storage layer manages both historical and real-time data, ensuring continuous data availability for further analysis and processing. Notably, the recorded data includes metadata essential for the digital twin, such as microphone location coordinates, device identification, and room designation. This data can also be accessed through REST-based or SQL-based communication interfaces.
4. The calculated parameters and metadata are retrieved from the cloud-based database in the logic layer. Based on the nature of the environment, the defined KPIs, and user- or environment-specific settings, this layer processes the data and prepares it for visualization. The logic layer can be implemented at any host platform, whether on a local or a cloud platform. It uses user inputs and commands gathered in the presentation layer and the data managed in the storage layer to perform various calculations and predictions. Additionally, it can integrate external data sources, such as Building Information Modeling (BIM) models, to allow 3D visualizations as well as access metadata related to the built environment that is used for comfort analysis. Depending on the presentation layer and whether it is remotely connected to this layer or operates on the same device, the processed data can be accessed through various communication interfaces, including REST-based, file-based, or shared memory communication methods.
5. The presentation layer provides users with various visualization tools, including graphs, charts, and 2D or 3D heat maps. It also accepts user-defined settings and, through continuous communication with the logic layer, ensures that the system displays the desired outputs. This layer can be implemented either on an online digital twin platform or a local computer.

Notably, in a real-world scenario, collecting data from multiple rooms requires additional instances of the implemented data acquisition and data processing layers. This scalable approach facilitates the integration of multiple rooms without significant computational constraints. The performance on each edge device may vary depending on the complexity of the Sound Event Detection (SED) model and the computational power of the edge device, but it is not dependent on the number of rooms.

The proposed framework was adopted for the use case of sound event classification in an office building. A sound classification system that is integrated with the digital twin platform is developed, and a case study is performed to assess the applicability of the developed system.

3.1 Case Study

The case study was conducted in an office space within a public building. The room, measuring 106 square meters, has an L-shaped layout with a ceiling height of 2.96 meters. It is typically used for meetings, classes, and various gatherings, but it also occasionally serves as a study or group work area for students.

The objectives of this implementation are threefold:

1. To develop a system capable of measuring a collection of soundwave parameters, classifying sound events, and transmitting the collected data to a cloud-based platform
2. To develop a method for visualizing the collected data effectively
3. To integrate the developed acoustic comfort measurement module into the existing digital twin

The digital twin of the environment is implemented using Autodesk Platform Services. Time-series data is stored in Microsoft Azure Data Explorer, and the existing dashboard provides historical data visualization for parameters such as temperature, air pressure, air quality, and occupancy (Figure 2).

In the case study implementation, a sound event classification neural network was utilized to categorize audio events. YAMNet [19], [20] is a pre-trained deep learning model with 521 output classes. For the case study, the model needs to be customized. For example, the existing classes are grouped into a smaller number of categories, tailored to the case study environment. Additionally, the model is required to be fine-tuned to enhance its accuracy for the given application. For effective grouping, it is important to consider the environment's characteristics (e.g., a wild animal sound is unlikely to occur in an office, despite the existence of such a class in YAMNet sound categories). Based on the types of activities expected in the case study environment (an office), five main categories were defined (Table 1). Each YAMNet class was then reviewed and assigned to one of these categories.

To capture sound, microphones are installed in the room. To achieve reliable results, selecting the right microphone (or an array of microphones) is one of the key design considerations for the implemented system. For example, the microphones should not perform active noise control (ANC) (i.e., noise cancellation), or modify the sound balance. Additionally, to efficiently perform the measurement and calculation of metrics, the microphone should be precise and sensitive to adequately capture sounds within its range of operation. In the case study, multiple identical microphones were installed at

different points of the room, each is connected to an edge device that analyzes the sound. The selection of microphone types depends on their placement location. In this study, omnidirectional microphones were used, featuring a signal-to-noise ratio (SNR) of 80 dB and a maximum sound pressure level of 110 dB.

Table 1. The categorization of the possible sounds in the office and the number of corresponding YAMNet output classes

Category	Number of Classes
Speech, Meeting	14
Silence, Ambient Office Sounds	30
Construction and maintenance work, Interruptions	112
Gathering, Crowded, Music	179
Other Classes	186

Installing microphones in the rooms may raise privacy concerns for occupants. As mentioned, a privacy-by-design approach was followed in this implementation, in which the sound classification is performed on the edge devices in real-time using small snippets of the sound signal collected from microphones that are directly attached to them. The actual captured sound is not recorded or transferred over the network. In other words, the data acquisition layer and the data processing layer are implemented in one device, ensuring that the original sound signal remains secure and eliminating the risk of data leakage over data transmission in the network.

The implemented system accommodates the measurement of the relative sound pressure levels as well as the sound event category identification. The implemented system allows the integration of new customized KPIs, considering the preferences of occupants and their intended activity within the space. The results are integrated and visualized in the developed digital twin platform of the facility, which is the primary source of information and a tool for monitoring and performing predictive analytics. The architecture of this implementation is depicted in Figure 1(right).

3.2 Visualization of Sound Content

Although various types of data visualization methods can be easily created based on the data generated in the logic layer, a combination of bar and line charts is the preliminary visualization method (Figure 3). It allows visualizing classifications with color-coded bars overlaid by various line charts that show calculated metrics. As an example, the relative sound pressure level, defined in Equation (1), is illustrated in Figure 3. The graph has the date and time on the horizontal axes, which allows visual analysis of the changes in acoustic metrics and indicators during a certain day.

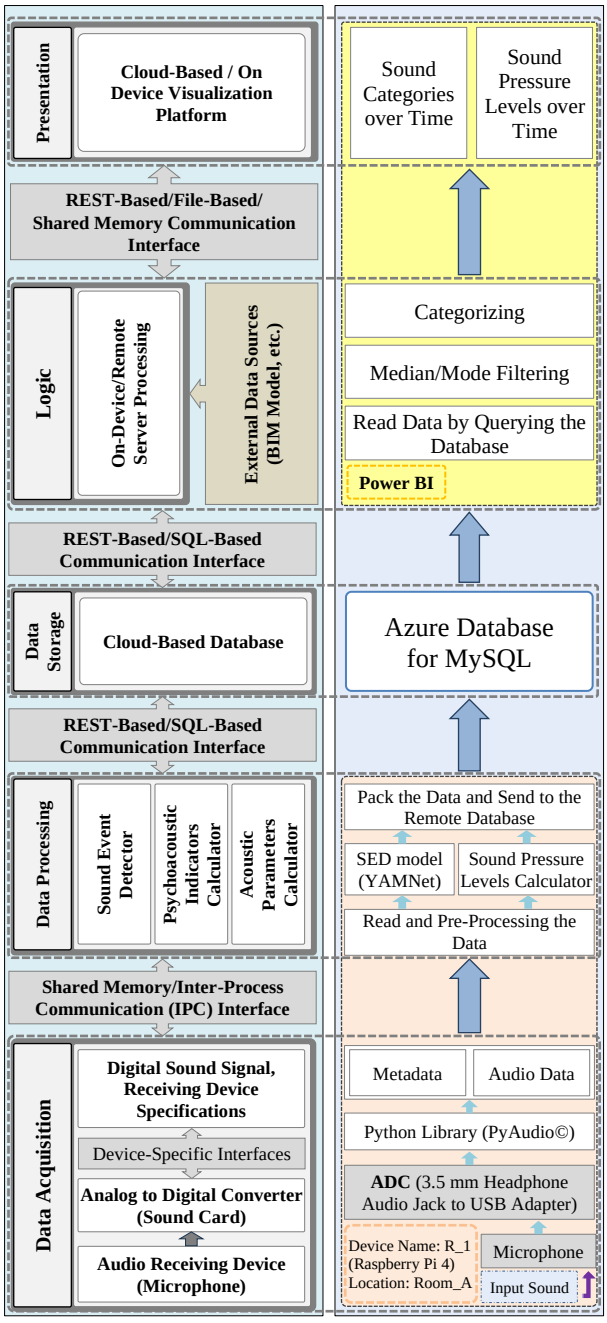

Figure 1. The Proposed Framework Layers (left) and the case study implementation (right)

Equation (1) represents the relative sound pressure level ($SPL_{relative}$). The input to the logarithm function is the Root Mean Square (RMS) value of the microphone's output, obtained using the PyAudio library at a 16 kHz sampling rate, and 16-bit audio depth, continuously

calculated over a defined interval (1-second) within the (-1, 1) range. The output of this equation falls within the (-inf, 0) dB range, and can be a relative measurement of the sound pressure level in an environment. Notably, this value cannot be directly used or compared with established comfort thresholds in the environment unless it is converted to absolute sound pressure levels.

$$SPL_{relative} = 20 * \log(RMS(normalized\ measured\ value)) \quad (1)$$

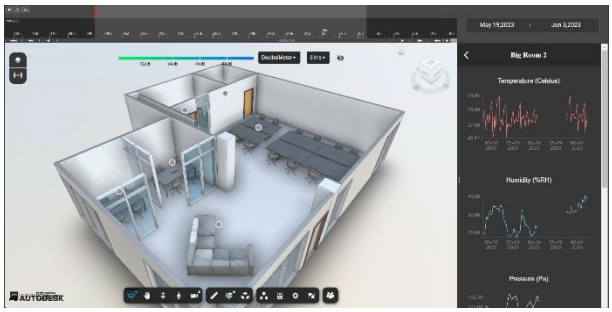

Figure 2. The existing digital twin of the office in Autodesk Platform Services

As depicted in the graph (Figure 3), the sound data over a seven-hour period is classified and visualized. The relative sound pressure level is represented in a dark blue line graph. The Construction and maintenance work/ Interruptions category is highlighted in red. The Gathering/Crowded/Music category is shown in purple, while the Silence/Ambient Office Sounds (such as printer noise) category is displayed in light blue. The "Other Classes" category is represented in dark grey.

Using these color codes and reading the graph from left to right, a meeting is detected from approximately 11:30 a.m. for two hours, represented in yellow. This is followed by a two-and-a-half-hour period of silence, shown in light blue. Finally, starting at 5:00 p.m., another meeting is observed for about two hours.

This chart is accessible through the digital twin platform's web interface by clicking on the acoustic monitoring icon available for each room.

4 Conclusions and Future Work

The focus of this paper was to provide a framework to perform acoustic comfort analysis by combining various metrics and indicators together with sound event category identification within a digital twin. Regarding SED models, YAMNet is used in the case study. However, various other models, such as SoundNet [21], Google's VGGish, HuBERT [22], OpenL3 [23], the CRNN proposed in [24], DENet [25], SincNet [26], and COPE [27] can also be used. YAMNet was selected for this study as it is a lightweight model suitable for implementation on edge devices, and it offers acceptable performance. Additionally, it is pre-trained and is adaptable to various types of environments. However, to get the best results, the performance of these networks should be evaluated and compared based on specific use case and fine-tuning.

The selection of both the microphone and sound interface plays a critical role in ensuring classification accuracy and precise measurement of acoustic parameters. Using a microphone with an inappropriate directional type in relation to its placement or one with a low signal-to-noise ratio (SNR) can limit the effective measurement range, creating blind spots in the system's sound classification capabilities. Furthermore, to fully utilize a microphone's potential, the sound interface must support high bit depth and sample rate, which are essential for capturing maximum acoustic data from the environment for further analyses.

The framework also included integrating acoustic measurements, such as acoustic waveform variables (e.g., sound pressure levels), psychoacoustic indicators (e.g., loudness), as well as sound event classification, into a digital twin platform. The potential applications of this

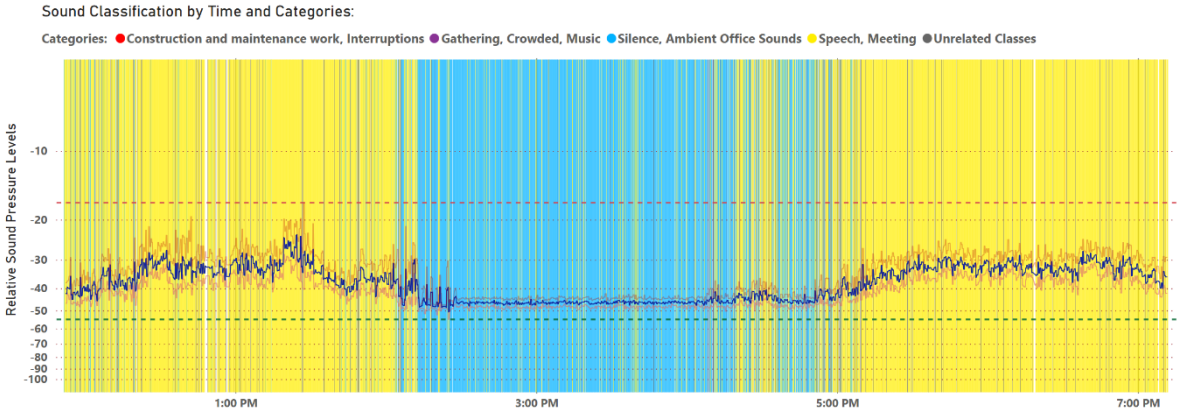

Figure 3. The Output Graph for Monitoring the Acoustic Status of the Case Study

framework were explored, and a case study was conducted in an office space, utilizing edge devices and an SED network to monitor both the current and historical status of the environment in a digital twin platform.

References

[1] Heinzerling, D., Schiavon, S., Webster, T., and Arens, E. Indoor environmental quality assessment models: A literature review and a proposed weighting and classification scheme. *Building and Environment*. 70 :210–222. 2013.

[2] Fasano, S., Fissore, V. I., Puglisi, G. E., Shtrepi, L., and Astolfi, A. Acoustic comfort contribution to the overall indoor environmental quality in workplaces.

[3] Navai, M. and Veitch, J. Acoustic Satisfaction in Open-Plan Offices: Review and Recommendations. 2003.

[4] Zhao, Y., Zhao, Q., Xia, L., Cheng, Z., Wang, F., and Song, F. A unified control framework of HVAC system for thermal and acoustic comforts in office building. *2013 IEEE International Conference on Automation Science and Engineering (CASE)*. pages 416–421. 2013.

[5] Bonet-Solà, D., Vidaña-Vila, E., and Alsina-Pagès, R. M. Acoustic Comfort Prediction: Integrating Sound Event Detection and Noise Levels from a Wireless Acoustic Sensor Network. *Sensors*. 24 (13) :4400. 2024.

[6] Zhang, E., Chen, Y., Chen, X., Zhang, J., Xu, P., and Zhuo, J. High-Precision Modeling and Prediction of Acoustic Comfort for Electric Bus Based on BPNN and XGBoost. *The International Journal of Acoustics and Vibration*. 28 (2) :158–164. 2023.

[7] Dobie, R. A. and Hemel, S. V. Hearing Loss: Determining Eligibility for Social Security Benefits. *Hearing Loss: Determining Eligibility for Social Security Benefits*. 2004.

[8] Kang, J., Aletta, F., Gjestland, T. T., Brown, L. A., Botteldooren, D., Schulte-Fortkamp, B., Lercher, P., Van Kamp, I., Genuit, K., Fiebig, A., Bento Coelho, J. L., Maffei, L., and Lavia, L. Ten questions on the soundscapes of the built environment. *Building and Environment*. 108 :284–294. 2016.

[9] Fastl, H. and Zwicker, E. *Psychoacoustics*. Berlin, Heidelberg, Springer Berlin Heidelberg. 2007.

[10] Rossing, T. D. *Springer handbook of acoustics*. New York, Springer. 2007.

[11] Hossain, M. R., Manohare, M., and King, E. A. Systematic review of indoor soundscape assessments: Activity-based psycho-acoustics analysis. *Building Acoustics*. 32 (1) :123–141. 2025.

[12] Aletta, F., Kang, J., and Axelsson, Ö. Soundscape descriptors and a conceptual framework for developing predictive soundscape models. *Landscape and Urban Planning*. 149 :65–74. 2016.

[13] Engel, M. S., Fiebig, A., Pfaffenbach, C., and Fels, J. A Review of the Use of Psychoacoustic Indicators on Soundscape Studies. *Current Pollution Reports*. 7 (3) :359–378. 2021.

[14] Vardaxis, N.-G., Bard, D., and Persson Waye, K. Review of acoustic comfort evaluation in dwellings—part I: Associations of acoustic field data to subjective responses from building surveys. *Building Acoustics*. 25 (2) :151–170. 2018.

[15] Herranz-Pascual, K., Iraurgi, I., Aspuru, I., Garcia-Pérez, I., Eguiguren, J. L., and Santander, Á. Development of the Acoustic Comfort Assessment Scale (ACAS-12): Psychometric properties, validity evidence and back-translation between Spanish and English. *PLOS ONE*. 18 (2) :e0281534. 2023.

[16] Kang, J., Aletta, F., Margaritis, E., and Yang, M. A model for implementing soundscape maps in smart cities. *Noise Mapping*. 5 (1) :46–59. 2018.

[17] Espejo, D., Vargas, V., Viveros-Muñoz, R., Labra, F. A., Huijse, P., and Poblete, V. Short-time acoustic indices for monitoring urban-natural environments using artificial neural networks. *Ecological Indicators*. 160 :111775. 2024.

[18] Yue, R., Meng, Q., Yang, D., Wu, Y., Liu, F., and Yan, W. A visualized soundscape prediction model for design processes in urban parks. *Building Simulation*. 16 (3) :337–356. 2023.

[19] Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. Audio Set: An ontology and human-labeled dataset for audio events. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pages 776–780. 2017.

[20] Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., and Wilson, K. CNN architectures for large-scale audio classification. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pages 131–135. 2017.

[21] Aytar, Y., Vondrick, C., and Torralba, A. SoundNet: Learning Sound Representations from Unlabeled Video. 2016.

[22] Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. 2021.

[23] Cramer, A. L., Wu, H.-H., Salamon, J., and Bello, J. P. Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics,*

Speech and Signal Processing (ICASSP). pages 3852–3856. 2019.

[24] Çakır, E., Parascandolo, G., Heittola, T., Huttunen, H., and Virtanen, T. Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 25 (6) :1291–1303. 2017.

[25] Greco, A., Roberto, A., Saggese, A., and Vento, M. DENet: a deep architecture for audio surveillance applications. *Neural Computing and Applications*. 33 (17) :11273–11284. 2021.

[26] Roberto, A., Saggese, A., and Vento, M. A deep convolutionary network for automatic detection of audio events. *Proceedings of the 3rd International Conference on Applications of Intelligent Systems*. In Las Palmas de Gran Canaria Spain. pages 1–6. 2020.

[27] Strisciuglio, N., Vento, M., and Petkov, N. Learning representations of sound using trainable COPE feature extractors. *Pattern Recognition*. 92 :25–36. 2019.