Measurement: Sensors 38 (2025) 101711

Contents lists available at ScienceDirect

Measurement: Sensors

journal homepage: www.sciencedirect.com/journal/measurement-sensors

Implementing transfer learning for sound event classification using the realised audio database

ARTICLE INFO

Keywords:
Smart spaces
Acoustic sensors
Audio detection
Transfer learning

ABSTRACT

Currently, smart spaces utilize diverse sensors to gather real-time data on activities within environments, enabling systems to activate alarms or execute tasks. These environments support various applications integrating advanced sensory systems like cameras, microphones, and motion sensors. Acoustic sensors are crucial, extracting valuable information from sound waves, including source details and environment characteristics. This article focuses on detecting and classifying common household sounds using the ReaLISED database, comprising 2479 recordings of 18 events. Audio data were collected in real homes with everyday objects and activities, unlike reverberation models. Artificial Intelligence techniques, including Machine Learning classifiers, yielded satisfactory outcomes. Transfer Learning methods were explored, leveraging pre-trained models like YAMNet for sound event recognition by Google, potentially outperforming previous classifiers. Preliminary results reveal a relative reduction of around 25 % in the probability error rate when comparing the obtained results with previous studies, providing valuable insights into potential possibilities and enhancements.

1. Introduction

The surge in interest surrounding sensor utilization within smart homes marks a significant transition towards integrating technology into residential settings, aiming to elevate convenience, efficiency, and safety [1]. Sensors play a pivotal role in smart home ecosystems by capturing real-time data on diverse environmental aspects and user behaviors, encompassing motion, temperature, humidity, and light sensors, among others. One driving factor behind this increasing interest is the rise of the Internet of Things (IoT) technology [2], which enables seamless communication and interaction between devices and systems within the home [3]. By connecting sensors to IoT platforms, homeowners can remotely monitor and control their home environment via smartphone apps or voice assistants. Another factor contributing to the growing adoption of sensors in smart homes is the desire for energy efficiency and sustainability [4]. Sensors can help optimize energy usage by automatically adjusting lighting, heating, and cooling systems based on occupancy patterns and environmental conditions. This not only reduces energy waste but also lowers utility bills for homeowners.

Moreover, the paramount emphasis on home security and safety drives the uptake of sensor-centric smart home solutions [5]. Motion sensors, door/window sensors, and surveillance cameras are instrumental in detecting and notifying homeowners of potential security breaches, bolstering overall peace of mind.

Sound processing in smart environments often aims to address the challenge of Sound Event Classification (SEC). This involves identifying the collection of active sounds within a provided audio signal. An expanding area where acoustic recognition finds application is in speech emotion recognition [6–10].

Numerous algorithms of various types can be employed to identify different sounds. Additionally, the choice of learning method can significantly impact the results. Some of the most well-known methods include: machine learning, deep learning, transfer learning, and deep learning, among others.

Machine Learning, a subfield of artificial intelligence (AI), is dedicated to crafting algorithms and statistical models empowering computers to execute tasks sans explicit programming [11]. These algorithms iteratively enhance their performance through experience or data. Machine learning finds applications in diverse fields, including spam filtering, recommendation systems, image recognition, and medical diagnosis.

Deep Learning, a subset of machine learning, utilizes neural networks consisting of multiple layers (known as deep neural networks) to extract representations from data. These networks possess the capability to autonomously identify and comprehend features from raw data, enabling them to undertake intricate tasks like image recognition, natural language understanding, and speech recognition [12]. Deep learning methodologies, such as convolutional neural networks (CNNs) utilized for image analysis and recurrent neural networks (RNNs) employed for sequential data, have demonstrated exceptional efficacy across diverse domains. These techniques have exhibited impressive performance levels in fields like computer vision, speech recognition, and autonomous driving.

Transfer Learning is a machine learning approach in which a model trained for one task is repurposed as a foundation for a related task. By capitalizing on the insights acquired from extensive datasets and pre-trained models, this method enables more efficient handling of new tasks, especially when data scarcity is a concern [13]. Transfer learning is widely applied in domains such as computer vision, natural language processing, and speech recognition to improve performance on target tasks.

In this paper Transfer Learning is used, specifically YAMNet (Yet Another Multitask Network) which is a deep learning model designed for environmental sound analysis and classification. It is specifically

https://doi.org/10.1016/j.measen.2024.101711

Available online 4 January 2025

engineered to recognize and classify audio signals into a wide range of sound categories, such as musical instruments, human sounds, animals, and various environmental noises. YAMNet leverages deep neural networks, particularly convolutional neural networks (CNNs), to extract high-level features from audio spectrograms and make predictions about the content of the audio. This model has been trained on a large dataset of labelled audio clips and has demonstrated impressive performance in sound event detection and classification tasks [14].

The database used is named ReaLISED (Real-Life Sound Event Database) [15]. The Real-Life Indoor Sound Event Dataset (ReaLISED) represents a newly developed database aimed at advancing scientific research by furnishing a substantial collection of real labelled indoor audio event recordings. This resource provides the scientific community with an opportunity to assess Sound Event Classification (SEC) algorithms. Comprising 2479 sound clips encompassing 18 distinct events, the dataset adheres to a meticulous recording process outlined in the proposal. Additionally, a method for evaluating the similarity of new audio data has been described, rendering the dataset scalable and poised for potential expansion as desired by researchers. With a well-balanced distribution of recordings across different event types, the dataset exhibits a desirable characteristic. However, it is important to note that all audio recordings are confined to indoor environments, aligning with the dataset's primary objective.

2. Methods and procedures

In this section, the database, the algorithms used to perform the experiments, are described in detail below.

Fig. 1 depicts the system described in the paper. In the initial phase, the YAMNet network, a deep neural network detailed later, is trained using 527 audio event classes and 2,084,320 human-labelled 10-s sound clips sourced from YouTube videos. Following training, a model is obtained, which is provided by YAMNet, and subsequently used to test the ReaLISED database, which is transfer Learning.

2.1. Description of the database

The ReaLISED (Real-Life Indoor Sound Event Dataset) is a new database recorded, labelled and developed by the authors of the present proposal. It is composed of eighteen (18) different events. Some of them are related to daily household tasks, such as cooking, walking or cleaning, and others include sounds produced by humans or other objects present in a house. The event classes are the following: beater, cooking, cupboard/wardrobe, dishwasher, door, drawer, furniture movement, microwave, object falling, smoke extractor, Electronics 2022, 11, 1811 4 of 18 speech, switch, television, vacuum cleaner, walking, washing machine, water tap, and window. There are 2479 clips of isolated sounds, which result in 3624.51 s (slightly more than 1 h).

The recordings were conducted across various locations, encompassing apartments, semi-detached houses, and the Polytechnic School of the University of Alcalá. The rationale behind selecting this diverse range of environments was to promote generalization of room impulse responses. This approach aims to enable the development of classifiers that prioritize event class information over the specific environment where the sound source is situated, thus enhancing the robustness and applicability of the classifiers to real-world scenarios.

The recording process for the dataset followed a robust protocol, incorporating strict guidelines regarding the parameters of the recorder and its proximity to the sound source. The sampling frequency was established at 44.1 kHz with 24 bits per sample, providing ample capability to sample signals with a bandwidth of up to 20 kHz and a dynamic range of 144 dB. Although recorded in stereo mode, one channel was subsequently discarded. A medium sensitivity setting of the microphone, corresponding to 94 dBSPL, was chosen to prevent sound saturation. However, prior to recording, it was ensured that saturation did not occur in the recorder sound. Nonetheless, a post-recording review was conducted, and any instances of saturation were identified and eliminated. Additionally, the distance between the recorder and the sound source was maintained at approximately 30–40 cm.

Table 1 summarizes the description of the dataset, providing details such as the event class recorded in each audio, the number of recordings (audios) for each class, the average duration of each class (in seconds), and whether the class is characterized by impulsivity or not. The impulsivity of an audio is defined by its brief duration, sudden onset, and rapid decay [16]. The number of events in each class ranges from 104 for the "Window" class to 190 for the "Speech" class, with a mean value of 138 events and a standard deviation of 25. The average duration varies depending on the type of sound, with the shortest durations observed in impulsive sound classes such as "Object falling" and "Switch," while the longest durations are found in classes where the sound source is a person, such as "Speech" and "Walking."

2.2. Transfer learning

Transfer learning is a machine learning technique where a model trained for a specific task is repurposed as a starting point for a related task, typically when there is a scarcity of data for the second task or when the second task is very different but shares certain characteristics with the first. This approach allows leveraging knowledge gained from large datasets and pre-trained models to tackle new tasks more efficiently [13].

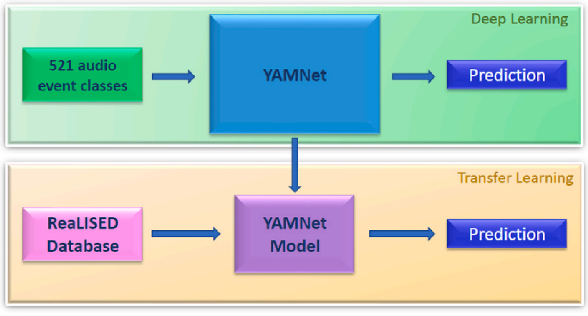


Fig. 1. Scheme of the system.

Table 1
Summary of the dataset characteristics, including the number of audios, the average duration and the impulsivity or non-impulsivity of the different classes.).

Event Class	Number of Audios	Average Duration (s)	Impulsive Sound (Yes/No)
Beater	126	1.02	No
Cooking	176	1.01	No
Cupboad/ Wardrobe	156	0.96	Yes
Dishwaher	130	1.01	No
Door	109	1.75	Yes
Drawer	158	1.34	Yes
Smoke extractor	142	1.01	No
Furniture movement	152	1.18	No
Microwave	124	1.03	No
Object falling	119	0.49	Yes
Speech	190	4.03	No
Switch	108	0.14	Yes
Water tap	140	1.40	No
Television	143	2.18	No
Vacuum cleaner	166	1.26	No
Walking	119	2.67	No
Washing machine	117	1.01	No
Window	104	1.74	Yes

2.3. YAMNet

YAMNet, which stands for "Yet Another Multitask Network," is a deep learning model developed by Google Research specifically for sound event detection and classification. It is designed to analyze environmental audio data and categorize it into a wide range of sound classes.

Is primarily used for sound event detection and classification tasks. It can recognize various types of sounds in audio recordings, such as musical instruments, human sounds, animal sounds, and environmental noises.

YAMNet, is primarily used for sound event detection and classification tasks. It can recognize various types of sounds in audio recordings, such as musical instruments, human sounds, animal sounds, and environmental noises.

YAMNet employs a deep convolutional neural network (CNN) architecture to extract features and perform classification tasks. Trained on a substantial dataset of labelled audio clips, the model can effectively discern patterns and characteristics within audio spectrograms. Utilizing a multitask learning paradigm, it simultaneously learns to classify multiple sound events. This approach enables the model to exploit correlations and interdependencies between different sound classes, enhancing its overall performance and versatility. Google provides a pre-trained version of this network, which has been trained on a dataset called AudioSet. This pre-trained model can be fine-tuned or used directly for various sound classification tasks. YAMNet is engineered with scalability in mind, enabling efficient handling of large volumes of audio data. It can seamlessly process audio recordings in real-time or batch mode, catering to both online and offline applications. YAMNet's applications span a broad spectrum, encompassing environmental monitoring, surveillance systems, audio-based activity recognition, and assistive technologies for the hearing impaired.

YAMNet is a pretrained deep net that predicts 521 audio event classes based on the AudioSet-YouTube corpus, and employing the Mobilenet_v1 depthwise-separable convolution architecture.

The dataset utilized for evaluation is derived from the AudioSet [17] database, which features an extensive ontology comprising 527 audio event classes Six classes were omitted based on the recommendations from reviewers. This database encompasses a vast collection of 2,084, 320 human-labelled 10-s sound clips sourced from YouTube videos. The ontology is structured as a hierarchical graph of event categories, encompassing a diverse array of human and animal sounds, musical instruments and genres, as well as common everyday environmental sounds. On average, each video is associated with 5 event labels. Labels for the videos were generated automatically using a combination of metadata (title, description, comments, etc.), context, and image content.

The MobileNet architecture predominantly employs depthwise separable convolutions, with the exception of the initial layer, which is a full convolution. Each layer is subsequently followed by batch normalization and Rectified Linear Unit (ReLU) activation, except for the final fully connected layer, which lacks nonlinearity and connects to a softmax layer for classification purposes. Down-sampling is achieved through strided convolution in both the depthwise and initial layers. A concluding average pooling step diminishes spatial resolution to 1 before reaching the fully connected layer. With depthwise and pointwise convolutions counted separately, MobileNet comprises a total of 28 layers [18].

YAMNet comprises a total of 86 layers, which can be broken down as follows: 27 layers for convolutional operations (14 for standard convolution and 13 for depth-wise convolution), 27 layers for batch normalization and ReLU activation, 1 average pooling layer, 1 fully connected (FC) layer, 1 softmax layer, and final classification layer [19].

In YAMNet, the initial layer is the convolutional layer, responsible for extracting meaningful features from the input audio waveform. Proper normalization of data is essential to prevent training difficulties and enhance the network's learning speed. To achieve this, batch normalization layers are employed after each convolutional layer. These normalization layers not only expedite the training process but also simplify the learning dynamics. Subsequently, a rectified linear unit (ReLU) activation function is applied to prevent the network's computational complexity from growing exponentially.

During the classification phase, an FC (fully connected) layer and a softmax classifier are utilized. The FC layer consolidates information from each neuron in the preceding layer with every neuron in the subsequent layer, enabling comprehensive analysis of the entire input and facilitating informed decision-making [20].

The network is trained using the Adam algorithm [21] with a learning rate of 5.10^{-4}, a minibatch size of 256, and for 20 epochs. The best-performing network is selected based on validation results. The data is split, with 80 % used for training and 20 % for validation.

3. Results and discussion

This section presents the results obtained using the ReaLISED database described in the previous section, employing the YAMNet neural network. Table 2 displays the accuracy, precision, and F1-Score results for each class, expressed as percentages.

Overall, the results are generally very favourable, indicating proficient performance. However, it is notable that certain classes, such as cupboard, door, and window, exhibit higher error rates compared to others.

To identify areas of confusion between classes, Fig. 2 illustrates the confusion matrix derived from the experiments. It is evident that the most significant confusion occurs between the sound of doors and windows, which is logical considering their similarities in sound characteristics.

In [15], the results are shown using several classifiers of different nature such as, Least Squares Linear Classifiers, Least Squares Quadratic Classifier, k-Nearest Neighbours, Support Vector Machine, Multilayer Perceptron and Deep Neural Networks.

To compare with the previously published study, we averaged the results for all events and obtained an F1-score of 87.66 %. When comparing this with the machine learning methods and two deep learning networks presented in that study, we outperform all cases.

The F1-scores (%) obtained in the previous study are as follows: Least Squares Linear Classifier 77.69 %, Least Squares Quadratic Classifier 82.12 %, k-Nearest Neighbours 76.53 %, Support Vector Machine 76.55 %, Multilayer Perceptron with 32 neurons 76.55 %, Multilayer Perceptron with 64 neurons 79.79 %, Multilayer Perceptron with 128 neurons

Table 2
Summary of the dataset characteristics, including the number of audios, the average duration and the impulsivity or non-impulsivity of the different classes.).

Nº	Event Class	Accuracy (%)	Precision (%)	F1_Score (%)
1	Beater	97.85	98.87	98.36
2	Cooking	99.20	99.03	99.12
3	Cupboard/Wardrobe	65.25	66.27	65.76
4	Dishwasher	98.76	97.34	98.05
5	Door	60.45	55.47	57.85
6	Drawer	81.13	75.67	78.31
7	Smoke extractor	91.67	99.49	98.57
8	Furniture movement	72.17	86.03	78.49
9	Microwave	96.93	98.84	97.88
10	Object falling	76.21	78.93	77.55
11	Speech	100	99.73	99.86
12	Switch	88.51	89.34	88.93
13	Water tap	95.14	96.24	95.68
14	Television	98.88	97.71	98.29
15	Vacuum cleaner	99.57	98.74	99.16
16	Walking	95.54	81.97	88.24
17	Washing machine	94.95	97.88	96.39
18	Window	60.38	62.36	61.35

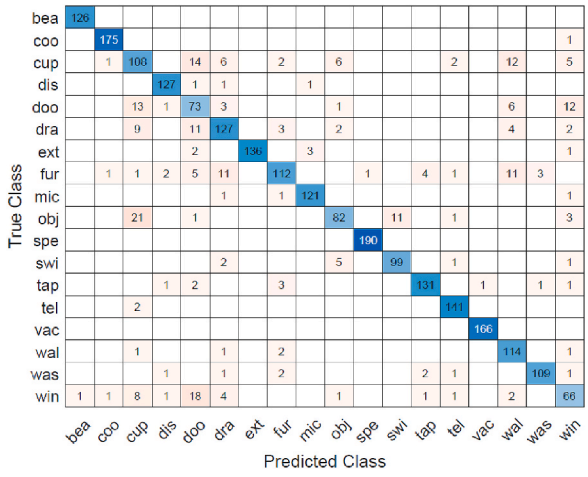


Fig. 2. Confusion matrix of event types.

81.74 %, Deep Neural Network with two hidden layers 84.44 %, and Deep Neural Network with three hidden layers 84.17 %. As we can see, the use of the YamNET network significantly improves the results compared to traditional machine learning algorithms and also shows an improvement, though to a lesser extent, over the deep neural networks.

4. Conclusions

Numerous conclusions can be drawn from the results. The evaluation of the database reaffirms its efficacy, as it functions correctly, yielding highly relevant results in terms of class differentiation. Instances of confusion between events are observed, and upon examination, it becomes evident that these confusions arise between events with very similar origins.

We can observe that the highest F1-scores are associated with sounds such as beater, cooking, dishwasher, smoke extractor, microwave, speech, television, vacuum cleaner, and washing machine, all of which are non-impulsive sounds. This suggests that these sounds are easier to detect.

For example, speech achieves nearly a 100 % success rate. In the confusion matrix example, it has a 100 % success rate, indicating that there are no other sounds with which the system confuses it.

The successful application of transfer learning using YAMNet validates the suitability of the ReaLISED database for training sound event detection systems. This demonstrates the database's utility and effectiveness in training models for sound event detection tasks.

Funding statement

This work is part of the project PID2021-129043OB-I00 funded by MCIN/AEI/10.13039/501100011033/FEDER, UE, and project EPU-INV/2020/003, funded by the Community of Madrid and University of Alcala.

Acknowledgments

We thank Theory of Signal and Communications Department of University of Alcalá, FEDER, and Community of Madrid.

References

[1] D. Wahyudin, et al., Smart home system for elderly healthcare, in: International Conference on Computer, Control, Informatics and its Applications (IC3INA), IEEE, 2020, pp. 1–6.
[2] C. Bisdikian, Internet of Things (IoT) for smart homes: challenges and opportunities, in: IEEE 17th International Conference on Ubiquitous Wireless Broadband (ICUWB), IEEE, 2017, pp. 1–4.
[3] A. Al-Fuqaha, et al., Internet of Things: a survey on enabling technologies, protocols, and applications, IEEE Commun. Survey. Tutorial. 17 (4) (2015) 2347–2376.
[4] H. Cappon, et al., The role of IoT sensors in energy efficiency in smart homes, in: IEEE Global Humanitarian Technology Conference (GHTC), IEEE, 2018, pp. 1–8.
[5] S. Kumar, et al., Smart home security system using IoT and machine learning, in: International Conference on Emerging Trends in Information Technology and Engineering (Ic-ETITE), IEEE, 2020, pp. 1–6.
[6] I. Mohino, M. Goni, L. Alvarez, C. Llerena, R. Gil-Pita, Detection of emotions and stress through speech analysis, in: Proceedings of the Signal Processing, Pattern Recognition and Application-2013, 2013, pp. 12–14. Innsbruck, Austria.
[7] M. Engin, A. Demirel, E. Engin, M. Fedakar, Recent developments and trends in biomedical sensors, Measurement 37 (2) (2005) 173–188.
[8] S. Kwon, Att-Net: enhanced emotion recognition system using lightweight self-attention module, Appl. Soft Comput. 102 (2021) 107101.
[9] S. Kwon, MLT-DNet: speech emotion recognition using 1D dilated CNN based on multi-learning trick approach, Expert Syst. Appl. 167 (2021) 114177.
[10] J. García-Gómez, M. Bautista-Durán, R. Gil-Pita, I. Mohino-Herranz, M. Rosa-Zurera, Violence detection in real environments for smart cities, in: Ubiquitous Computing and Ambient Intelligence: 10th International Conference, 2016, Part II vol. 10, 2016, pp. 482–494.
[11] Zhi-Hua Zhou, Machine Learning, Springer nature, 2021.
[12] Ian Goodfellow, Yoshua Bengio, Aaron Courville, Deep Learning, MIT press, 2016.
[13] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2009) 1345–1359.
[14] Eleni Tsalera, Andreas Papadakis, Maria Samarakou, Comparison of Pre-Trained Cnns for Audio classification using transfer learning, J. Sens. Actuator Netw. 10 (4) (2021) 72.
[15] Inma Mohino-Herranz, et al., Introducing the ReaLISED dataset for sound event classification, Electronics 11 (12) (2022) 1811.
[16] T.H. Pedersen, Audibility of impulsive sounds in environmental noise, in: En INTER-NOISE and NOISE-CON Congress and Conference Proceedings. Institute of Noise Control Engineering, 2000, pp. 4158–4164.
[17] Ievgeniia Kuzminykh, et al., Audio interval retrieval using convolutional neural networks. En Internet of Things, smart spaces, and next generation networks and systems, in: 20th International Conference, NEW2AN 2020, and 13th Conference, ruSMART 2020, St. Petersburg, Russia, August 26–28, 2020, Proceedings, Part I 20, Springer International Publishing, 2020, pp. 229–240.
[18] G. Andrew, Z. Menglong, Efficient convolutional neural networks for mobile vision applications, mobilenets, arXiv preprint arXiv:1704.04861 (2017).
[19] Félix Gontier, Romain Serizel, Christophe Cerisara, Automated audio captioning by fine-tuning bart with audioset tags, in: En DCASE 2021-6th Workshop on Detection and Classification of Acoustic Scenes and Events, 2021.
[20] Kamel K. Mohammed, et al., Radio frequency fingerprint-based drone identification and classification using Mel spectrograms and pre-trained YAMNet neural, Internet of Things 23 (2023) 100879.
[21] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

I. Mohino-Herranz[a,*], J. García-Gómez[b], S. Alonso-Diaz[a], J.G. Gallegos[a], F.J. Perez-Sanz[a], M. Aguilar-Ortega[b], R. Gil-Pita[b]
[a] *National Institute of Aerospace Technology (INTA), Ctra. Ajalvir, km. 4, Torrejón de Ardóz, 28850, Madrid, Spain*
[b] *University of Alcalá, Ctra. Madrid-Barcelona, km. 33,6, 28805, Alcalá de Hernares, Spain*

[*] Corresponding author.
E-mail address: mmohher@inta.es (I. Mohino-Herranz).