

Acoustic event characterization for service robot using convolutional networks

Fernando Martínez, Fredy Martínez, César Hernández

Facultad Tecnológica, Universidad Distrital Francisco José de Caldas, Bogotá, Colombia

Article Info

Article history:

Received Oct 25, 2021

Revised May 25, 2022

Accepted Jun 21, 2022

Keywords:

Acoustic event

Convolutional neural network

Human-machine interaction

Image categorization

Learning process

ABSTRACT

This paper presents and discusses the creation of a sound event classification model using deep learning. In the design of service robots, it is necessary to include routines that improve the response of both the robot and the human being throughout the interaction. These types of tasks are critical when the robot is taking care of children, the elderly, or people in vulnerable situations. Certain dangerous situations are difficult to identify and assess by an autonomous system, and yet, the life of the users may depend on these robots. Acoustic signals correspond to events that can be detected at a great distance, are usually present in risky situations, and can be continuously sensed without incurring privacy risks. For the creation of the model, a customized database is structured with seven categories that allow to categorize a problem, and eventually allow the robot to provide the necessary help. These audio signals are processed to produce graphical representations consistent with human acoustic identification. These images are then used to train three convolutional models identified as high-performing in this type of problem. The three models are evaluated with specific metrics to identify the best-performing model. Finally, the results of this evaluation are discussed and analyzed.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Fredy Martínez

Facultad Tecnológica, Universidad Distrital Francisco José de Caldas

Carrera 7 No. 40B-53, Bogotá, Colombia

Email: fhmartinezs@udistrital.edu.co

1. INTRODUCTION

The technological development of today has made life easier for human beings, particularly for those individuals with limitations and/or special needs. Among the first technologies designed by men to improve their living conditions were prescription lenses and glasses. Thanks to them, a person with reduced visual capacity was able to interact socially without restrictions. Today this technology has evolved to incorporate ultralight and intelligent materials that adapt not only to the user but also to the conditions of the environment. This is an example of current smart hardware, but hand in hand with this hardware has also evolved software tools, which in the case of glasses allow not only the design but also to estimate the future performance of the lenses and even their acceptance among users. These developments are strongly marked by their ability to integrate with the human being, they must not only be able to solve a problem but they must also be accepted by people [1].

Here is where service robotics research has found a broad niche with significant research problems to solve. Today's society has new challenges that demand the provision of specialized services, such as the care of people (elderly and children), or special training processes (children in their homes, or adults in industrial

environments) [2]. Such services require an artificial system that beyond providing content, tracks the processes according to the particular behavior of the individual [3]. There is a lot of research on human-machine integration systems, in some cases trying to identify emotional states from people's faces, but this involves processing the image in familiar environments including children, which for many is a privacy risk. A less intrusive strategy is based on the same principle of estimating emotions and special events but from acoustic signals, which is the area of interest of this research.

However, the question is whether it is possible for a small robot to autonomously identify acoustic events, particularly using convolutional networks. The identification of sound events by an autonomous robot is a desirable feature because it can improve integration in task execution, particularly in human-machine interaction [4]–[6]. The level of interaction, and thus task success, is strongly related to the robot's ability to anticipate human attitudes, which is particularly true for robots that provide care to people [7], [8]. Examples of acoustic events may include screaming people, crying babies, or an audible alarm. However, these events are characterized by complex parameters, which, while recognizable to a person, are not recognizable to a computerized system [9], [10]. Two cries of people turn out to be completely different, parameters such as volume, length, frequency, or acoustic power are not enough to characterize them, moreover, the robot should identify the group of any person, which infers the need for a learning process with a large dataset [11], [12]. Although the characterization of these complex events is complex, convolutional networks have demonstrated great success in the categorization of other problems with similar complexity [13], [14].

Additionally, convolutional models can be trained for specific needs. The databases, and the categories defined within them, can be adjusted to the needs of the system [15]. For example, it is possible to identify whether a set of images belong to a jungle landscape or a beach, but the same images can also be used to determine whether there are people or animals in them, or whether there are vehicles or houses in them [16]. A dataset can contain a lot of information, and it is up to the use and designer of the classifier to define which features he wants to identify and use in his model. In addition, while most of the known applications focus on image categorization, in many applications the signals of a given system have been modified into a visual representation, which has allowed its use in multiple situations with various input parameters [17], [18].

Finally, it is possible to autonomously identify acoustic events by a small robot using convolutional neural networks (CNNs) because convolutional models can run in real-time after they have been trained correctly [19], [20]. Service robots must operate in human environments, interacting with humans, so they must be able to respond in real-time to immediate needs [21]–[23]. If a service robot detects a person's scream, it must be able to autonomously give and request help at the same instant it detects the event [24]. Even so, robots include a large number of parallel systems that allow them not only to interact but also to guarantee their functionality and the safety of the users [25]. Therefore, the performance demands, and therefore on hardware, are high. Each of these systems, including the sound event identification event, must be capable of running on the robot hardware, considering the possibility of restricted communication in cases of emergency [26]. CNNs can create models that can be deployed on small embedded systems without excessive resource consumption [27], [28]. The current frameworks have compatibility for their implementation on different systems, which also facilitates their continuous updating. Thus, while neural networks are black boxes, and their performance is strongly dependent on their training, it is possible to autonomously identify acoustic events by a small robot using convolutional neural networks. Beyond the existing acoustic event detection (AED) systems, our application requires in addition to high performance the ability to operate in real-time, and the possibility of running continuously as a parallel task in a small service robot. This is precisely the gap in the current research that we seek to solve, since the reported solutions, to the best of our knowledge, have not reached a point of development that allows their massive use on commercial hardware (embedded processors for these robots, audio digitization systems, and cloud processing platforms).

The structure of this paper is as follows: section 2 provides the general details of the robotic development platform, its function, and the objective of the acoustic event characterization scheme on it. It discusses the desired characteristics of such a scheme, and how it should be integrated with the current robot schemes. Section 3 provides the design features of the convolutional models, all the tasks developed for their implementation and evaluation, and the corresponding analysis in each case. All the characteristics of each test are detailed, as well as the information required for its duplication. Section 4 focuses on the results of each model concerning the performance metrics applied on them with an unknown data set. Section 5 provides the limitation of this research. Finally, section 6 presents the interpretation of the performance results of each model, and conclusions.

2. PROBLEM STATEMENT AND STATE OF THE ART

The objective of this research is to evaluate whether a model based on deep networks can correctly identify acoustic events that may indicate high risk to users. This model is intended to be implemented on a service robot designed for home environments, performing tasks such as the care and surveillance of children, the elderly, or vulnerable people. Therefore, the aim is to establish the real performance of convolutional models trained to classify acoustic events from a specific database. If the robot can identify, for example, screams or gunshots, it is expected to be able to communicate autonomously to request immediate help. This idea can be extended to the care of sick people, or people with disorders that require special monitoring, or even for assistance in disaster situations.

By definition, an acoustic event corresponds to an audio segment that is identifiable by the human being to be related to a specific context [29]. In this way, a human being can identify the proximity of a vehicle by listening to its engine or establish the aggressivity of a dog by listening to its bark. AED corresponds to the identification of specific parameters in the audio signal that allow classifying the event that produces the sound. This technique has been successfully applied in acoustic surveillance applications, audio signal labeling, and environmental sound identification [30], [31].

The analysis strategies used in AED are of two types, those applied to monophonic sounds and those used with polyphonic sounds, in which different events appear in the same audio signal. Most of the recent research focuses on monophonic sounds under the assumption that the predominant acoustic event in the audio sample is identified and characterized. This predominant event corresponds to an anomaly (a rare event) and therefore should be easily separable from the rest of the sample. This approach goes hand in hand with the development of systems for real-life tasks, since acoustic events are usually mixed with other sounds. In addition, a good AED must be able to isolate and identify the event of interest regardless of the environment (other sounds in the audio), which becomes irrelevant in the identification.

From this perspective, AED systems correspond to one of the strategies of computational auditory scene analysis (CASA). These strategies are based on the development of artificial systems from computational tools that can similarly isolate sounds as human beings do. This type of processing is performed by identifying parameters and characteristics in the audio that allow classification. The audio is divided into segments, and then the characteristics of interest are extracted from each segment. Something similar is done in the development of trainable models, in which these features are paired with event labels.

Machine learning techniques have brought more tools to the process, including mel-frequency cepstral coefficients (MFCCs), and perceptual linear prediction coefficients [32], [33]. The mel-scale is of particular importance since it allows to express of digital acoustic signals on a scale following the human perception of sounds. It corresponds to a perceptual scale of tones defined by human observers, which allows an artificial system to respond to stimuli close to those experienced by humans. This scale has as a reference point a 1,000 Hz tone at 40 dBs above the auditory threshold, which is leveled with a 1,000 mels tone. In addition, above 500 Hz the frequency intervals are exponentially separated, which is perceived by the human ear as a more linear spacing. To convert a frequency f in Hertz to a value m in mels, (1) is used.

$$m = 1127.01048 \log_e \left(1 + \frac{f}{700} \right) \quad (1)$$

Convolutional networks have improved the performance of automatic classification systems, particularly in applications with images [34]. This is convenient since musicians and psychologists often choose to generate a two-dimensional representation of the mel-scale. To construct these representations, tone color or chroma is assigned, as well as a pitch, which generates an image with the audio information itself. Many deep convolutional network models have been proposed, which have been evaluated both with public datasets and in proprietary applications. We propose to select some of these classification models to determine the actual performance of a real-time event identification task.

3. METHODS AND MATERIALS

Our AED system is supported on deep networks, so the training database and its pre-processing are critical. The goal is that the system will be robust to the presence of ambient noise, so the audio samples used in training the models must contain background noise, as expected from the real-world operation. We selected a

set of public databases that meet these characteristics. Not all samples in our dataset come from a single public database, to avoid bias we mix audios from different databases, and some were even recorded by us in the lab. In addition, in the cases where the human voice was required, we tried to make the database gender-balanced (males comprise 41%). This database was separated into a training set and a validation set. We used 80% of the audios in each category for training, and the remaining 20% for validation. The selection of the training and validation sets was completely random, but the same set was used to train each convolutional model. In total, we selected three deep architectures according to previous laboratory performance evaluation: residual neural network (ResNet), dense convolutional network (DenseNet), and neural architecture search network (NASNet). The fitted models were evaluated with the same metrics on the same validation set. Figure 1 shows the details of our framework.

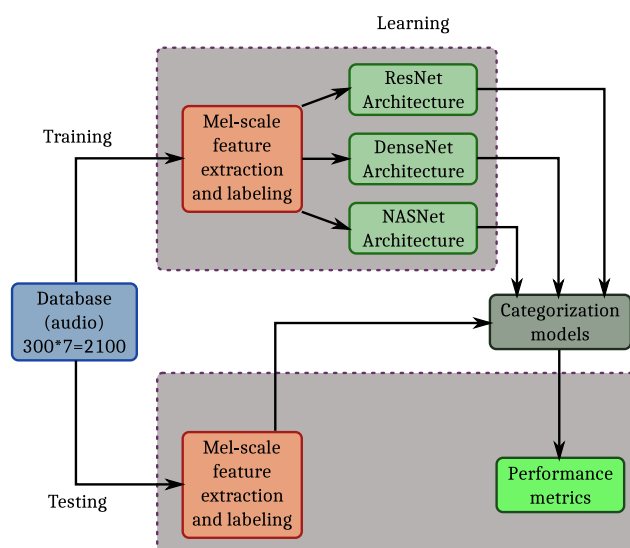


Figure 1. Framework to evaluate the robustness of CNN based AED

We examined different public databases to build the dataset used in our experiments. We used the common voice dataset to build the base category corresponding to the human voice in a natural and quiet (normal) state. This dataset contains the voice of people between 19 and 89 years old, corresponding to reading blog posts, books, and movies mainly [35]. We also used the UrbanSound Dataset from which we extracted the category corresponding to the gunshot. This database is composed of labeled sounds corresponding to urban events such as children playing, dogs barking, vehicle sirens or gunshots [36]. Another key database in the construction of our dataset is neural information processing group GENERAL sounds (NIGENS), this database also provides isolated urban acoustic events. From NIGENS we extracted the categories for crying baby (crying baby), burning fire (burning fire), human screams of men and women (scream), and dog barking (dog) [37]. The last database used was TUT Rare Sound Events 2017, which was developed for the DCASE challenge 2017 Task 2, and is composed of events corresponding to babies crying, gunshots, and glass breaking. All images from this dataset were used either to complement the already defined categories or to create the new category corresponding to glass breakage (glass) [38]. Some of these databases provide the audio files in MP3 format, others in WAV format. All files were initially unified to WAV format. In the case of MP3 files, the FFmpeg library was used to convert them to WAV format.

Our dataset is made up of 2,100 sounds (300 per category) with durations between 2 and 40 s. The Mel spectrogram for the audio files was calculated using the librosa 0.8.1 library. The audio signals in WAV format were sampled as a time-series input at a sampling rate of 22,050, the magnitude of the spectrogram was calculated and then mapped onto the Mel scale. An FFT window of 1,024 was used, and 100 samples were between successive frames. Figure 2 shows some images resulting from this conversion process, Figure 2(a) shows an audio corresponding to the normal base category, Figure 2(b) audio from the gunshot category, Figure 2(c) audio from the baby crying category, Figure 2(d) audio from the burning fire category, Figure 2(e) audio from the human scream category, Figure 2(f) audio from the dog barking category, and in Figure 2(g) audio from the glass breaking category.

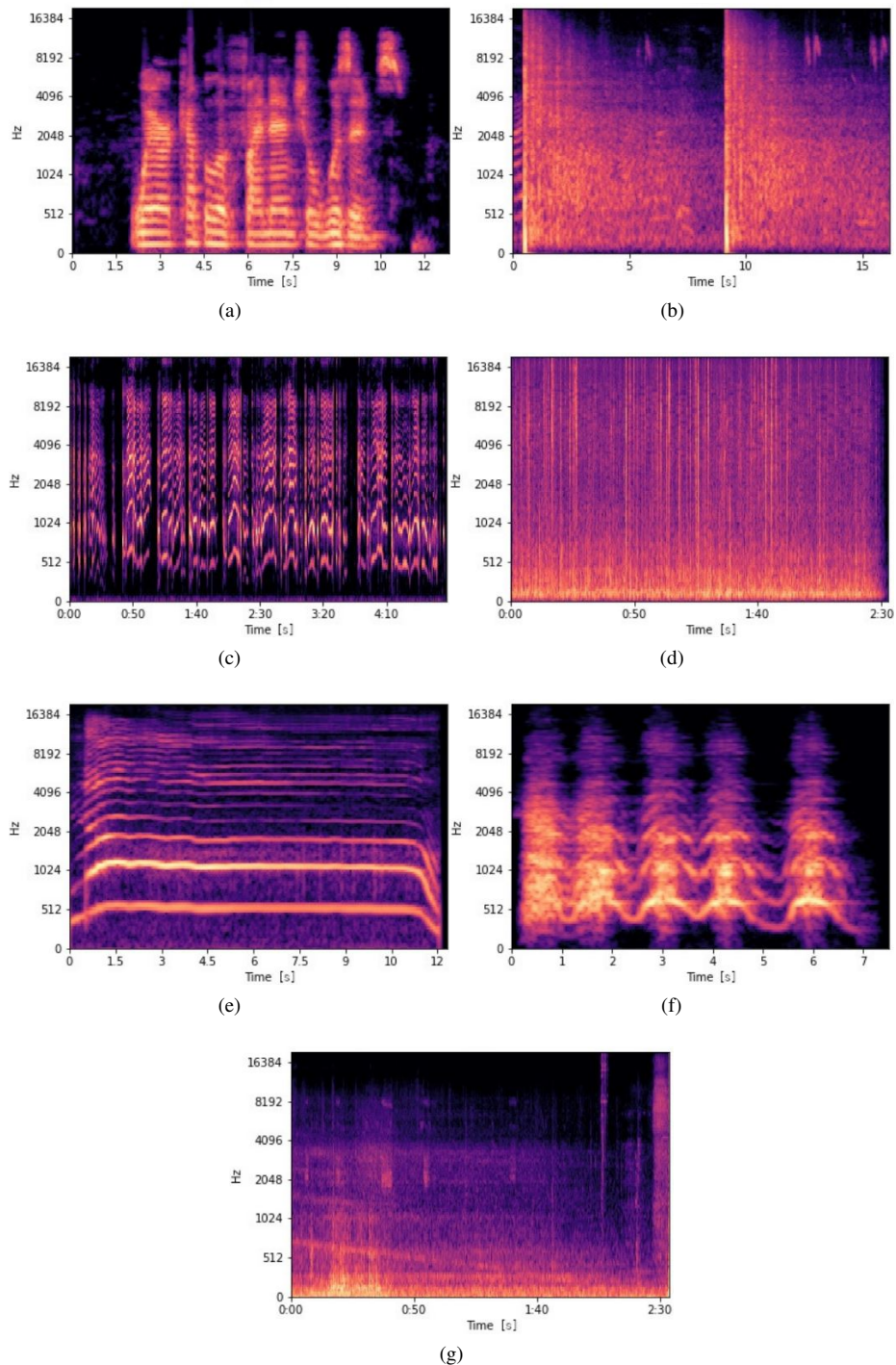


Figure 2. Sample of some Mel scale images used for model training (a) normal category, (b) gunshot category, (c) baby crying category, (d) burning fire category, (e) human scream category, (f) dog barking category, and (g) glass breaking category

The images resulting from this process originally had a size of 432×288 pixels. For ease of testing, this size was modified to a square structure of 256×256 pixels using OpenCV 4.1.2. For this process, the images were initially randomly shuffled by changing the seed for each category, and then stored in folders labeled with the name of the corresponding category. These files had different characteristics such as length

(some were 1 s, but others were as long as 2 min), number of channels, and frequencies. We pre-processed these audios to normalize the frequency (they were adjusted to the bandwidth of the human ear), unified the number of channels to monaural, and adjusted the length of the audios from 2 to 40 s. The seven categories or labels are:

- Category 1: normal
- Category 2: gunshot
- Category 3: cryingbaby
- Category 4: burningfire
- Category 5: scream
- Category 6: dog
- Category 7: glass

The normal category corresponds to the voice of people in a quiet state reading texts, the second category corresponds to the sounds of firearms in different scenarios, the third is the sound of small babies crying, the fourth corresponds to the sound of flames burning different materials, the scream category again has the voice of men and women but producing screams, category six includes barking dogs, and category seven corresponds to sounds produced by breaking glass. The content of the audios in the public databases was verified and edited to ensure the content of the event as the protagonist in each file. Most of the audios correspond to polyphonic events.

We selected three convolutional architectures for the models. The selection was made according to their high performance on similar problems, and the small size of the models that make them suitable for implementation on the robot. The research group had previously used these architectures in other modules of our robotic platform to solve problems such as automatic emotion recognition and movement strategies in unknown and dynamic environments [19], [20]. The selected architectures are ResNet, DenseNet, and NASNet. For the ResNet architecture, we selected the ResNet-50 model with 50 depth layers for a total of 23,602,055 parameters. For the DenseNet architecture, we selected DenseNet-121 for a total of 7,044,679 parameters, and for the NASNet architecture, we selected NASNet-Mobile for a total of 4,277,115 parameters. All models were fit for 20 epochs under the same criteria, i.e., categorical cross-entropy was used as the loss function, and stochastic gradient descent (SGD) was used as the optimizer. In all cases, the optimization process was monitored by calculating both the accuracy and the mean squared error (MSE) of the model with the training and validation samples. Table 1 summarizes the number of parameters related to each model.

Table 1. Parameters related to each model

| Model | ResNet 50 | DenseNet 121 | NASNet-Mobile |
|---------------------------|------------|--------------|---------------|
| Non-trainable parameters: | 53,120 | 83,648 | 36,738 |
| Trainable parameters: | 23,548,935 | 6,961,031 | 4,240,377 |
| Total parameters: | 23,602,055 | 7,044,679 | 4,277,115 |

In all three cases, the training was optimized according to the behavior of its accuracy with both training and validation data (although the latter were not directly considered for the optimization process). In the case of the ResNet network the accuracy was increased from 32.9% to 90.2% for the training data (considering the training process, i.e., epoch 1), and from 14.8% to 85.0% for the validation data. The DenseNet network achieved an increase from 49.8% to 90.1% in the accuracy of the training data, and from 12.6% to 81.7% for the validation data. Finally, for the NASNet network, an increase in accuracy from 56.1% to 91.9% was achieved for the training data, but 16.2% for the validation data was not achieved. All code was developed in Python 3.7 with support for Keras 2.6.0 and Tensorflow 2.6.0.

4. RESULT AND DISCUSSION

For the evaluation of the ability of the convolutional models to identify the parameters of each event correctly, standard machine learning metrics were used in the evaluation of classification models. These metrics were calculated from the behavior of the models against 20% of the dataset unknown to the model and initially separated for validation. As mentioned in the methodological development, during model training, accuracy and MSE were tracked at each epoch for both training and validation data, the results are summarized in

Figure 3, Figure 3(a) shows the performance for the ResNet model, Figure 3(b) shows the performance for the DenseNet model, and Figure 3(c) shows the performance for the NASNet model.

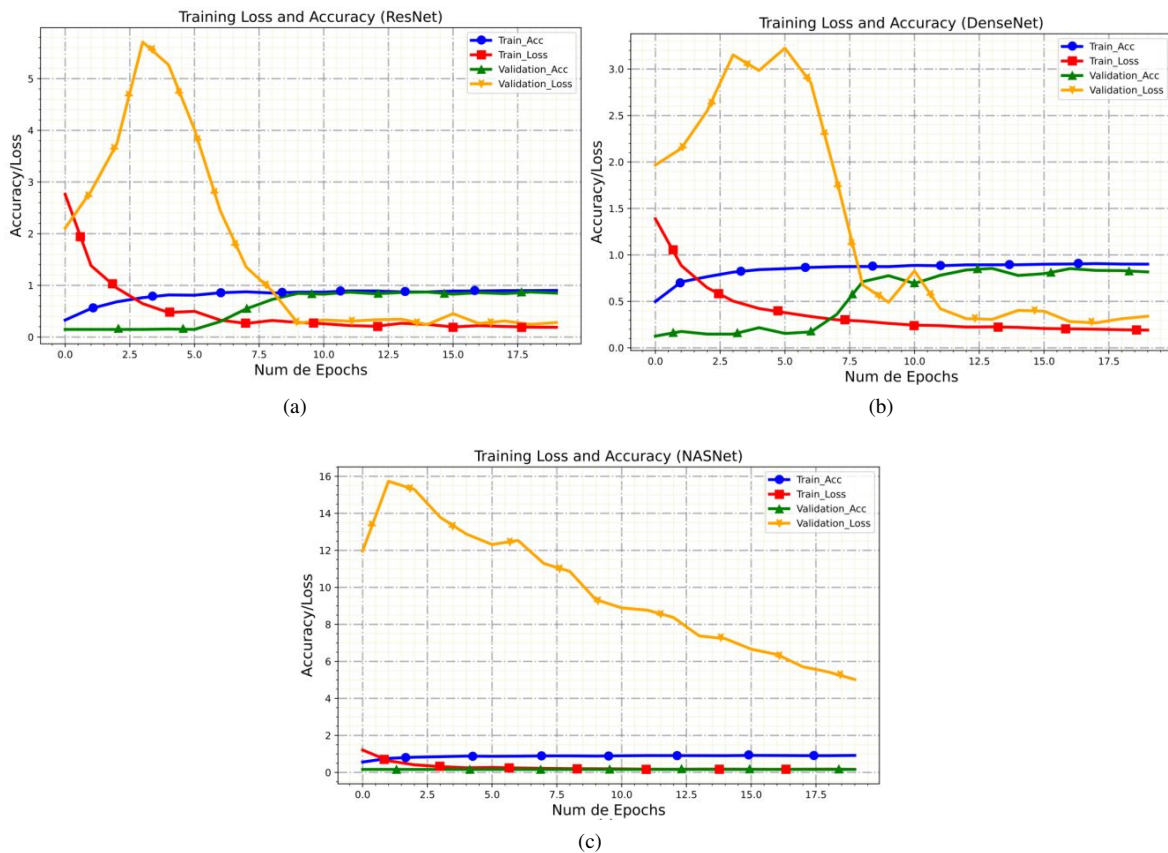


Figure 3. Summary of accuracy and error of the three models throughout training (a) ResNet, (b) DenseNet, and (c) NASNet

The smallest model is NASNet, however, it is also the model with the worst generalization ability. All three models managed to considerably reduce the error against the training data during the first five epochs; ResNet achieved an error of 0.48 and an accuracy of 81.5%, DenseNet showed an error of 0.42 and an accuracy of 84.2%, and NASNet an error of 0.25 with an accuracy of 88.2%. Although the best values for the training dataset were achieved by NASNet, this model suffered overfitting that prevented it from generalizing its categorization capability to unknown data. From the fifth epoch onwards, ResNet and DenseNet considerably reduce the error against validation data without reducing their performance on the training data, while NASNet reduces it but without achieving good performance. At the end of training ResNet and DenseNet achieve an error on validation data of 0.03, while NASNet only achieves a value of 0.23. Similar behavior is observed in the accuracy of the validation data.

The metrics used to evaluate the models are precision, Recall, F1-score, confusion matrix, and receiver operating characteristic (ROC) curves. Precision evaluates how many elements assigned to a category really belong to it. It corresponds to the ratio of true positives to the total number of elements identified in a category (true positives plus false positives). The Recall metric allows establishing how many of the elements assigned to a category by the model really belong to that category. It is calculated as the ratio between true positives concerning the elements that actually belong to the category (true positives plus false negatives). These two metrics can be visually summarized in the confusion matrix, since the main diagonal of this matrix corresponds to the true positives, the elements above it correspond to the false positives, and the elements below it correspond to the false negatives. To build this matrix, a table is assembled in which the rows correspond to the real categories of the elements, and the columns to the elements classified according to the model, values with

which the matrix is filled. The F1-score value corresponds to the weighted average of precision and Recall, so it somehow summarizes these two metrics. Finally, the ROC curve corresponds to a graphical representation of the sensitivity versus the specificity of the classifier. It uses changes in the threshold of the decision to evaluate the ratio of true positives to false positives. This graph is constructed with a diagonal curve, and all points above it (towards the upper left corner) are considered good classification results.

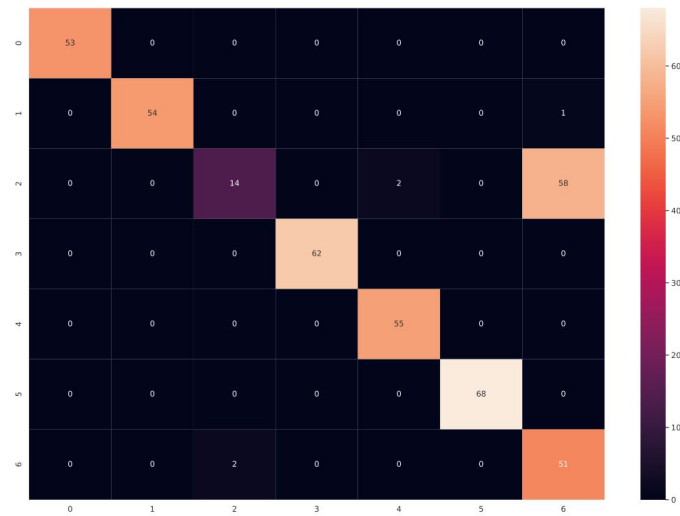
Table 2 outlines the results of the first three metrics. The values in the table are related to the previous curves. The ResNet and DenseNet models perform much better than the NASNet model, which performed poorly for the validation data. The only interesting value of the NASNet model corresponds to the Recall of the sixth category, with a value of 100%. This contrasts with the values for the other metrics in all categories but shows that the model was unable to categorize the elements of the validation group, assigning them all to a single category, the one corresponding to dog barking. Leaving the NASNet model aside, the performance of the first two models is excellent. In both cases, the average values of the three metrics were in the worst case at 79%. These values indicate that these models classify most of the validation values in the correct category. In both cases, however, problems are noted in two categories, in the third category (cryingbaby) very low Recall values are obtained, and in the seventh category (glass) very low precision values are obtained. In the first case, it is observed that many elements of this category were classified in other categories, i.e., the crying baby is not adequately differentiated from other sounds. In the second case, it means that many of the elements of this category do not really belong to it, i.e. it confuses as glass breakage many of the sounds of other categories.

Table 2. Precision, Recall, and F1-score values for the three models with validation data

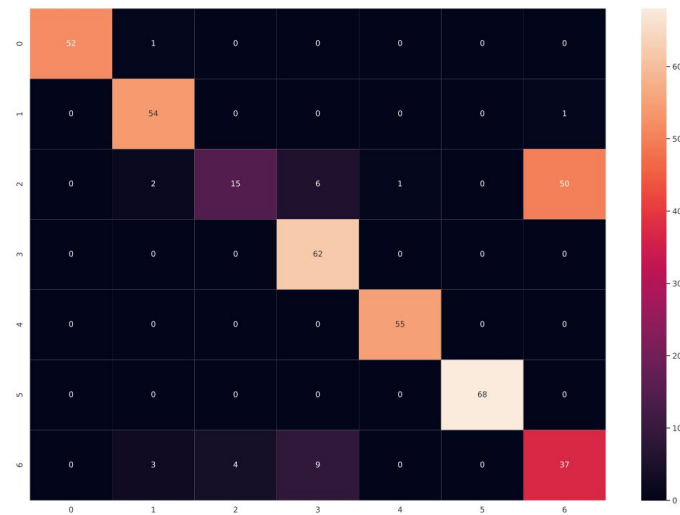
| Category | ResNet | | | DenseNet | | | NasNet | | |
|---------------|-----------|--------|----------|-----------|--------|----------|-----------|--------|----------|
| | Precision | Recall | F1-Score | Precision | Recall | F1-Score | Precision | Recall | F1-Score |
| normal | 1.00 | 1.00 | 1.00 | 1.00 | 0.98 | 0.99 | 00.00 | 00.00 | 00.00 |
| gunshot | 1.00 | 0.98 | 0.99 | 0.90 | 0.98 | 0.94 | 00.00 | 00.00 | 00.00 |
| crying baby | 0.88 | 0.19 | 0.31 | 0.79 | 0.20 | 0.32 | 00.00 | 00.00 | 00.00 |
| burningfire | 1.00 | 1.00 | 1.00 | 0.81 | 1.00 | 0.89 | 00.00 | 00.00 | 00.00 |
| scream | 0.96 | 1.00 | 0.98 | 0.98 | 1.00 | 0.99 | 00.00 | 00.00 | 00.00 |
| dog | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 00.16 | 1.00 | 00.28 |
| glass | 0.46 | 0.96 | 0.63 | 0.42 | 0.70 | 0.52 | 00.00 | 00.00 | 00.00 |
| Weighted avr. | 0.91 | 0.85 | 0.83 | 0.85 | 0.82 | 0.79 | 00.03 | 00.16 | 00.05 |

The problem identified in the metrics can be observed and analyzed in the confusion matrix in Figure 4, Figure 4(a) shows the confusion matrix for the ResNet model, Figure 4(b) shows the confusion matrix for the DenseNet model, and Figure 4(c) shows the confusion matrix for the NASNet model. The confusion matrices of the ResNet and DenseNet models show that in fact the problems in the third and seventh categories are related. The plots of these matrices use color-coding that makes it easy to identify the behavior of the diagonal of the matrix, and the classification model it represents. Dark colors show a lower amount of elements in the box, while light colors show a high concentration (scale on the right of each matrix). This coding should show in light colors the diagonal of a good classifier. In the ResNet and DenseNet models, this diagonal is quite well defined for most categories, but as before, problems are observed in the third and seventh categories. The third category looks in both cases very dark, and in the same row, in the seventh column, a high number of false positives are observed. The matrices show that many of the sounds corresponding to baby cries are erroneously placed in the category glass breakage. While this may be an isolated problem, the fact that both models have the same inconsistency suggests problems in the audio quality of the training dataset. In the third matrix, the one corresponding to the NASNet model, it is observed as before that all items were categorized in the sixth category. This caused the audios that belonged to the category to match, but the overall performance was poor due to the misclassification of all the others.

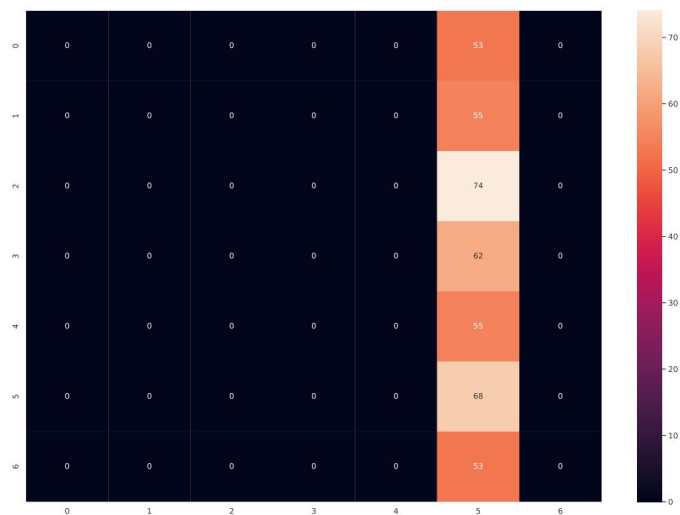
Since the ROC curve considers only the ratio of true positives to false positives, it is to be expected that the behavior for the NASNet model moves close to the diagonal, while for the ResNet and DenseNet models the curve seeks the upper left corner as shown in Figure 5, Figure 5(a) shows the ROC curve for the ResNet model, Figure 5(b) shows the ROC curve for the DenseNet model, and Figure 5(c) shows the ROC curve for the NASNet model. This means that the latter two models can correctly classify the elements that actually belong to a certain category (vertical axis of the ROC curve), something equivalent to the Recall metric. In short, the performance of the ResNet and DenseNet models is very high and similar to each other, while the NASNet model is inadequate for the development of the event characterization system.



(a)



(b)



(c)

Figure 4. Confusion matrix of the three models with validation data (a) ResNet, (b) DenseNet, and (c) NASNet

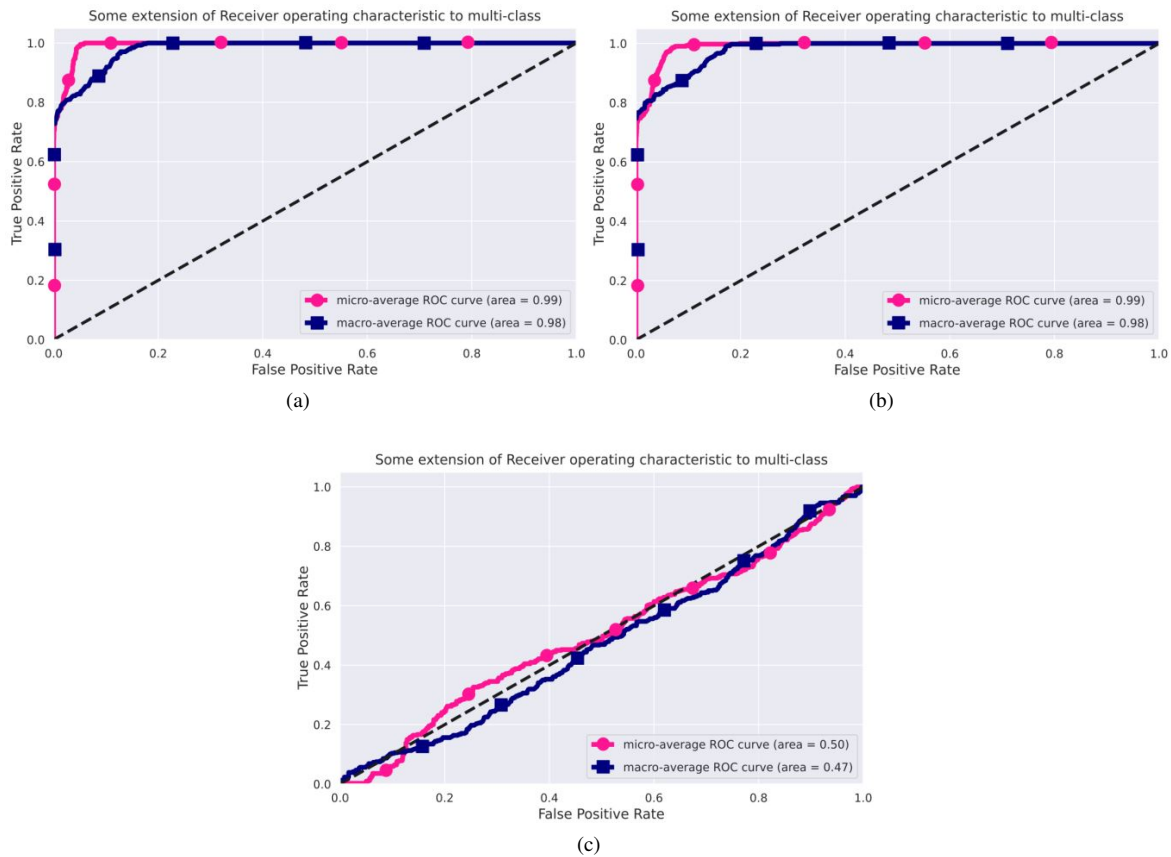


Figure 5. ROC curves of the three models with validation data (a) ResNet, (b) DenseNet, and (c) NASNet

5. LIMITATION

Our experiments relied on audio signals captured by commercial microphones, in some cases installed in low-quality devices. While this approach goes hand in hand with the existing microphones in our robot, and with the hearing ability of an average adult, it is clear that the audio quality and performance capability of the acoustic event characterization system is increased with the use of higher-quality sensors. Future developments of our system contemplate the use of these sensors and the development of a proprietary database that incorporates the acoustic signals captured by them.

The other limitation of our research is its narrow focus, which is primarily on domestic service robots. While this is our field of research, and the scope of our robotic prototypes, this limited perspective also limits the usefulness of our results in other scenarios such as applications in commercial and industrial environments. Still, we believe that the general scheme of work, as well as the results with the evaluated convolutional models, can serve as an initial stage for these application fields.

6. CONCLUSION

In this work, we propose the development of an autonomous system for the characterization of acoustic events that can be implemented in small service robots, particularly dedicated to the development of domestic tasks. In this sense, a scheme based on convolutional networks capable of being trained for a dataset specific to the needs of the problem is proposed. A dataset is built according to seven specific events: human voice in a normal state, gunshot, baby crying, fire in burning processes, human screams, dog barking, and glass breaking. For the construction of these categories, the use of public databases was chosen to evaluate the initial performance of the convolutional models. A database with 300 audios in each category was assembled, which were randomly separated into two groups, a training group with 80% of the data, and a validation group with the

remaining 20%. The audio signals were filtered to produce a homogeneous set of different parameters. To feed the convolutional models with an input signal that reflected the characteristics identifiable by the human ear, the entire dataset was converted to the Mel scale. The code was developed in Python with Keras and Tensorflow support, and the processing of the audios was performed with librosa and then transformed into images with OpenCV. The three convolutional models evaluated for the system were ResNet, DenseNet, and NASNet, these were selected due to their high performance in similar tasks and their compact size compared to other structures. The training was performed in all three cases under the same conditions, which involved the use of categorical cross-entropy as loss function and SGD as optimization function. The training was performed for 20 epochs, during which the error and accuracy of the training and validation data were calculated for model control and adjustment. To evaluate the performance, the Precision, Recall, F1-score, confusion matrix, and ROC curve metrics were used on the models trained with the validation data. The results provided by the metrics confirm a high performance for the ResNet and DenseNet models (average F1-score of 83% for ResNet and 79% for DenseNet), while they show that the NASNet model is unable to generalize the behavior to unknown data. Some problems were observed in two categories which we intend to evaluate with our dataset captured in the laboratory. Finally, although the ResNet model outperforms the DenseNet model in some values, the latter is selected for the development of the prototype system due to its smaller size and memory requirements. The research will continue adjusting the model to the conditions of the robotic platform, which implies the reconstruction of a more complex and higher-quality dataset.

ACKNOWLEDGEMENT

This work was supported by the Universidad Distrital Francisco José de Caldas, in part through CIDC, and partly by the Facultad Tecnológica. The views expressed in this paper are not necessarily endorsed by Universidad Distrital. The authors thank the research group ARMOS for the evaluation carried out on prototypes of ideas and strategies.

REFERENCES





- [1] J. Murphy, J. McVay, P. Mathews, D. A. Carnegie, and A. Kapur, "Expressive robotic guitars: developments in musical robotics for chordophones," *Computer Music Journal*, vol. 39, no. 1, pp. 59–73, Mar. 2015, doi: 10.1162/COMJ_a.00285.
- [2] R. O'Dowd, "Emerging trends and new directions in telecollaborative learning," *CALICO Journal*, vol. 33, no. 3, pp. 291–310, Aug. 2016, doi: 10.1558/cj.v33i3.30747.
- [3] A. Linson, C. Dobbyn, G. E. Lewis, and R. Laney, "A subsumption agent for collaborative free improvisation," *Computer Music Journal*, vol. 39, no. 4, pp. 96–115, Dec. 2015, doi: 10.1162/COMJ_a.00323.
- [4] E. J. Fabris, V. A. Sangalli, L. P. Soares, and M. S. Pinho, "Immersive telepresence on the operation of unmanned vehicles," *International Journal of Advanced Robotic Systems*, vol. 18, no. 1, Jan. 2021, doi: 10.1177/1729881420978544.
- [5] X. Zhao *et al.*, "A smart robotic walker with intelligent close-proximity interaction capabilities for elderly mobility safety," *Frontiers in Neurorobotics*, vol. 14, no. 1, pp. 1–8, Oct. 2020, doi: 10.3389/fnbot.2020.575889.
- [6] C. Kertész and M. Turunen, "Common sounds in bedrooms (CSIBE) corpora for sound event recognition of domestic robots," *Intelligent Service Robotics*, vol. 11, no. 4, pp. 335–346, 2018, doi: 10.1007/s11370-018-0258-9.
- [7] M. Podpora, A. Gardecki, R. Beniak, B. Klin, J. L. Vicario, and A. Kawala-Sterniuk, "Human interaction smart subsystem-extending speech-based human-robot interaction systems with an implementation of external smart sensors," *Sensors*, vol. 20, no. 8, Apr. 2020, doi: 10.3390/s20082376.
- [8] E. Dean-Leon, K. Ramirez-Amaro, F. Bergner, I. Dianov, and G. Cheng, "Integration of robotic technologies for rapidly deployable robots," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 4, pp. 1691–1700, 2018, doi: 10.1109/TII.2017.2766096.
- [9] X. Lin, X. Zou, Z. Ji, T. Huang, S. Wu, and Y. Mi, "A brain-inspired computational model for spatio-temporal information processing," *Neural Networks*, vol. 143, no. 1, pp. 74–87, Nov. 2021, doi: 10.1016/j.neunet.2021.05.015.
- [10] R. M. Alsina-Pagès, R. Benocci, G. Brambilla, and G. Zambon, "Methods for noise event detection and assessment of the sonic environment by the harmonica index," *Applied Sciences*, vol. 11, no. 17, Aug. 2021, doi: 10.3390/app11178031.
- [11] R. Löwe, J. Böhm, D. G. Jensen, J. Leandro, and S. H. Rasmussen, "U-FLOOD-Topographic deep learning for predicting urban pluvial flood water depth," *Journal of Hydrology*, vol. 603, pp. 1–6, 2021, doi: 10.1016/j.jhydrol.2021.126898.
- [12] T. C. Pham, C. M. Luong, V. D. Hoang, and A. Doucet, "AI outperformed every dermatologist in dermoscopic melanoma diagnosis, using an optimized deep-CNN architecture with custom mini-batch logic and loss function,"

- Scientific Reports*, vol. 11, no. 1, pp. 1–12, 2021, doi: 10.1038/s41598-021-96707-8.
- [13] Y. Xu and M. Vaziri-Pashkam, “Limits to visual representational correspondence between convolutional neural networks and the human brain,” *Nature Communications*, vol. 12, no. 1, 2021, doi: 10.1038/s41467-021-22244-7.
- [14] J. Karhade, S. K. Ghosh, P. Gajbhiye, R. K. Tripathy, and U. R. Acharya, “Multichannel multiscale two-stage convolutional neural network for the detection and localization of myocardial infarction using vectorcardiogram signal,” *Applied Sciences*, vol. 11, no. 17, 2021, doi: 10.3390/app11177920.
- [15] Z. Nabulsi *et al.*, “Deep learning for distinguishing normal versus abnormal chest radiographs and generalization to two unseen diseases tuberculosis and COVID-19,” *Scientific Reports*, vol. 11, no. 1, pp. 1–10, 2021, doi: 10.1038/s41598-021-93967-2.
- [16] A. De Cesarei, S. Cavicchi, G. Cristadoro, and M. Lippi, “Do humans and deep convolutional neural networks use visual information similarly for the categorization of natural scenes?,” *Cognitive Science*, vol. 45, no. 6, pp. 1–14, 2021, doi: 10.1111/cogs.13009.
- [17] T. Hachaj, Ł. Bibrzycki, and M. Piekarczyk, “Recognition of cosmic ray images obtained from cmos sensors used in mobile phones by approximation of uncertain class assignment with deep convolutional neural network,” *Sensors*, vol. 21, no. 6, pp. 1–16, 2021, doi: 10.3390/s21061963.
- [18] D. Gibert, C. Mateu, J. Planes, and R. Vicens, “Using convolutional neural networks for classification of malware represented as images,” *Journal of Computer Virology and Hacking Techniques*, vol. 15, no. 1, pp. 15–28, 2019, doi: 10.1007/s11416-018-0323-0.
- [19] F. Martínez, C. Hernández, and A. Rendón, “Identifier of human emotions based on convolutional neural network for assistant robot,” *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 18, no. 3, pp. 1499–1504, Jun. 2020, doi: 10.12928/telkomnika.v18i3.14777.
- [20] F. Martínez, C. Penagos, and L. Pacheco, “Deep regression models for local interaction in multi-agent robot tasks,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10942, 2018, pp. 66–73, doi: 10.1007/978-3-319-93818-9_7.
- [21] M. Arduengo, C. Torras, and L. Sentis, “Robust and adaptive door operation with a mobile robot,” *Intelligent Service Robotics*, vol. 14, no. 3, pp. 409–425, 2021, doi: 10.1007/s11370-021-00366-7.
- [22] J. C. Molina-Molina, M. Salhaoui, A. Guerrero-González, and M. Arioua, “Autonomous marine robot based on ai recognition for permanent surveillance in marine protected areas,” *Sensors*, vol. 21, no. 8, pp. 1–28, 2021, doi: 10.3390/s21082664.
- [23] K. Zheng, F. Wu, and X. Chen, “Laser-based people detection and obstacle avoidance for a hospital transport robot,” *Sensors*, vol. 21, no. 3, pp. 1–24, 2021, doi: 10.3390/s21030961.
- [24] J. Castañeda and Y. Salguero, “Adjustment of visual identification algorithm for use in stand-alone robot navigation applications,” *Tekhnê*, vol. 14, no. 1, pp. 73–86, 2017.
- [25] A. Moreno and D. F. Páez, “Performance evaluation of ROS on the Raspberry Pi platform as OS for small robots,” *Tekhnê*, vol. 14, no. 1, pp. 61–72, 2017.
- [26] A. Mehrotra *et al.*, “Accessible maker-based approaches to educational robotics in online learning,” *IEEE Access*, vol. 9, no. 9471866, pp. 96877–96889, 2021, doi: 10.1109/ACCESS.2021.3094158.
- [27] Q. Ji, C. Dai, C. Hou, and X. Li, “Real-time embedded object detection and tracking system in Zynq SoC,” *EURASIP Journal on Image and Video Processing*, vol. 2021, no. 1, Dec. 2021, doi: 10.1186/s13640-021-00561-7.
- [28] G. H. Kim, E. S. Sung, and K. W. Nam, “Automated laryngeal mass detection algorithm for home-based self-screening test based on convolutional neural network,” *BioMedical Engineering Online*, vol. 20, no. 1, pp. 1–10, 2021, doi: 10.1186/s12938-021-00886-4.
- [29] K. Imoto, “Introduction to acoustic event and scene analysis,” *Acoustical Science and Technology*, vol. 39, no. 3, pp. 182–188, 2018, doi: 10.1250/ast.39.182.
- [30] E. Lieskovska, M. Jakubec, and R. Jarina, “Acoustic surveillance system for children’s emotion detection,” in *42nd International Conference on Telecommunications and Signal Processing (TSP)*, Jul. 2019, pp. 525–528, doi: 10.1109/TSP.2019.8768884.
- [31] S. S. Sethi *et al.*, “Characterizing soundscapes across diverse ecosystems using a universal acoustic feature set,” in *Proceedings of the National Academy of Sciences*, vol. 117, no. 29, pp. 17049–17055, Jul. 2020, doi: 10.1073/pnas.2004702117.
- [32] I. D. S. Miranda, A. H. Diacon, and T. R. Niesler, “A comparative study of features for acoustic cough detection using deep architectures,” in *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul. 2019, pp. 2601–2605, doi: 10.1109/EMBC.2019.8856412.
- [33] K. Feroze and A. R. Maud, “Sound event detection in real life audio using perceptual linear predictive feature with neural network,” in *15th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, Jan. 2018, pp. 377–382, doi: 10.1109/IBCAST.2018.8312252.
- [34] F. Martínez, F. Martínez, and E. Jacinto, “Performance evaluation of the NASNet convolutional network in the automatic identification of COVID-19,” *International Journal on Advanced Science, Engineering and Information Technology*, vol. 10, no. 2, Apr. 2020, doi: 10.18517/ijaseit.10.2.11446.





- [35] R. Ardila *et al.*, “Common voice: a massively-multilingual speech corpus,” *arXiv preprint arXiv:1912.06670*, 2019.
- [36] J. Salamon, C. Jacoby, and J. P. Bello, “A dataset and taxonomy for urban sound research,” in *22nd ACM International Conference on Multimedia*, Nov. 2014, pp. 1041–1044, doi: 10.1145/2647868.2655045.
- [37] I. Trowitzsch, J. Taghia, Y. Kashef, and K. Obermayer, “The NIGENS general sound events database,” *arXiv preprint arXiv:1902.0831*, pp. 1–5, Feb. 2019.
- [38] A. Mesaros *et al.*, “DCASE 2017 Challenge setup : Tasks, datasets and baseline system,” in *Detection and Classification of Acoustic Scenes and Events 2017 Workshop*, 2017, pp. 85–92.

BIOGRAPHIES OF AUTHORS







Fernando Martínez     is a doctoral researcher at the Universidad Distrital Francisco José de Caldas focusing on the development of navigation strategies for autonomous vehicles using hierarchical control schemes. In 2009 he completed his M.Sc. degree in Computer and Electronics Engineering at Universidad de Los Andes, Colombia. He is a researcher of the ARMOS research group (Modern Architectures for Power Systems) supporting the lines of electronic instrumentation, control and robotics. He can be contacted at email: fmartinezs@udistrital.edu.co.



Fredy Martínez     is a professor of control, intelligent systems, and robotics at the Universidad Distrital Francisco José de Caldas (Colombia) and director of the ARMOS research group (Modern Architectures for Power Systems). His research interests are control schemes for autonomous robots, mathematical modeling, electronic instrumentation, pattern recognition, and multi-agent systems. Martínez holds a Ph.D. in Computer and Systems Engineering from the Universidad Nacional de Colombia. He can be contacted at email: fmartinezs@udistrital.edu.co.



César Hernández     is a professor of telecommunications, digital signal processing, and electronics at the Universidad Distrital Francisco José de Caldas (Colombia) and director of the SIREC research group (Systems and Cognitive Networks). His research interests are telecommunications, assistive technology, tele informatics, and advanced digital systems. Hernández holds a Ph.D. in Computer and Systems Engineering from the Universidad Nacional de Colombia. He can be contacted at email: cahernandezs@udistrital.edu.co.