DOI: 10.61782/fa.2023.1274

forum acusticum 2023

FuSA: APPLICATION OF A MACHINE LEARNING SYSTEM FOR NOISE MITIGATION ACTION PLANS IN URBAN ENVIRONMENTS

J.P. Arenas[1,*] **E. Suárez**[1] **P. Huijse**[2] **V. Poblete**[1] **M. Vernier**[2]
R. Viveros-Muñoz[1] **D. Espejo**[2] **V. Vargas**[2] **D. Vergara**[1]
[1] Institute of Acoustics, Univ. Austral of Chile, Valdivia, Chile
[2] Institute of Informatics, Univ. Austral of Chile, Valdivia, Chile

ABSTRACT

The urban noise environment comprises many sources, some of which are regulated by local legislation setting maximum permitted noise levels, which are vital in implementing the noise action plans. A multidisciplinary project funded by the Chilean R+D Agency has resulted in a machine learning-based system called FuSA that automatically recognizes sound sources in audio files recorded in the urban environment to assist in their analysis. FuSA (Integrated System for the Analysis of Environmental Sound Sources) incorporates a deep neural model transferred to a dataset of urban sound events compiled from public sources and recordings. The target dataset follows a customized taxonomy of urban sounds. The system also uses a public API so potential users can post audio files to determine the overall presence of noise sources contributing to environmental noise pollution. This work provides examples of how stakeholders can use FuSA to address urban noise problems and contribute to city noise abatement policies.

Keywords: *environmental noise, urban noise, machine learning, legislation.*

1. INTRODUCTION

Recently, the application of neural networks to solve different problems in engineering has received renewed attention. The main reason is that advances in computational resources, storage capability, and the availability of massive datasets have made the practical implementation of deep learning techniques possible. Reviews on the use of deep learning in acoustics and vibration have pointed out the advantages and limitations of such methods [1,2].

Deep learning (DL) is a branch of machine learning focused on training deep artificial neural networks (ANNs) to solve complex pattern recognition problems. ANNs are mathematical models inspired by biology. They consist of layers of artificial neurons that perform simple processing tasks when combined. In machine learning, models with more layers have greater flexibility to fit the training data accurately. This fact means that the deeper a model is, the better it can perform. DL models have achieved outstanding results, making them the top choice for many perception-related issues, including computer vision, speech recognition, and natural language processing [3-5]. A significant challenge is that training deeper models requires more labeled data. Therefore, having large, high-quality datasets is essential for practical training.

Transfer learning (TL) is a methodology that utilizes deep neural networks trained with a large amount of data to address a related and more specific problem [6]. In TL, the parameters of the final layers of the original model are adapted using a target dataset that may have a different class taxonomy than the source dataset for a specific task. This process is called fine-tuning. Therefore, it is possible to train a very deep and complex model with TL using a small target dataset, provided the source dataset is diverse enough.

Earlier neural network architectures have been replaced by processing layers with neurons organized as convolutional filters inspired by the response of a neuron in the visual

[*] ***Corresponding author***: *jparenas@uach.cl.*

cortex to a specific stimulus. Convolutional neural networks (CNNs) create deeper ANNs for larger inputs, like high-resolution data [7,8]. Thus, currently, CNNs are alternatives to conventional, fully connected ANNs for temporally or spatially correlated signals, such as advanced sound event detection (SED).

It is well-known that urban noise is a severe environmental problem growing over the years. Although the total urban noise comprises several sources, traffic noise is the most prevalent. However, due to their high sound levels, construction noise, entertainment, and leisure activities are commonly reported as sources of nuisance in the community.

Modern cities have addressed environmental noise by primarily elaborating strategic noise maps [9] followed by noise action plans. Not only the average sound levels are essential for implementing noise control action plans. Identifying the prevalent sources composing the soundscape is also critical, especially when enforcing laws regarding acceptable noise levels in urban settings. Therefore, DL appears to be an excellent alternative for creating support tools for monitoring and mitigating the adverse effects of environmental noise.

The Institutes of Acoustics and Informatics at the College of Engineering Sciences of the University Austral of Chile have completed a joint project titled "Integrated System for the Analysis of Environmental Sound Sources: FuSA System." The Chilean R+D Agency funded the project to create a machine learning-based system that automatically recognizes sound sources in audio files recorded in the urban environment. Other cities, like New York and Lorient in France, have also conducted research on categorizing urban sound sources [10,11].

The developed FuSA tools have aimed to be used by different stakeholders to address urban noise problems and contribute to city noise abatement policies. This work presents some examples of FuSA applications.

2. THE FUSA SYSTEM

2.1 Deep Learning model and dataset of urban sound events

The FuSA system uses a deep neural model introduced in 2020 by Kong et al. [12] called Pretrained Audio Neural Network (PANN), a powerful deep neural network model for audio tagging. This model has surpassed the performance of previous systems documented in the literature. PANN was trained using Google AudioSet, a dataset with over 2 million 10-second audio clips collected from YouTube and classified into 632 categories.

In their work, Kong et al. conducted experiments where they used knowledge from Google AudioSet to solve audio tagging problems in different target datasets. Notably, the fine-tuned PANN model has performed better than the state-of-the-art in some general urban sound event datasets.

The PANNs architecture consists of convolutional layers, i.e., processing layers with neurons organized as convolutional filters (see Fig. 1). An audio signal is processed within PANN through two paths. The first path utilizes a 1D convolutional neural network to analyze the waveform in the time domain. The network generates a WaveGram, a variation of the Fourier transform using neural networks. On the other hand, the second pathway converts the information into a log-scale mel spectrogram and enters a 2D convolutional neural network. The results from both paths are combined and further processed by a final 2D convolutional neural network. This network produces a probability estimate of whether various labels are present in the audio file.

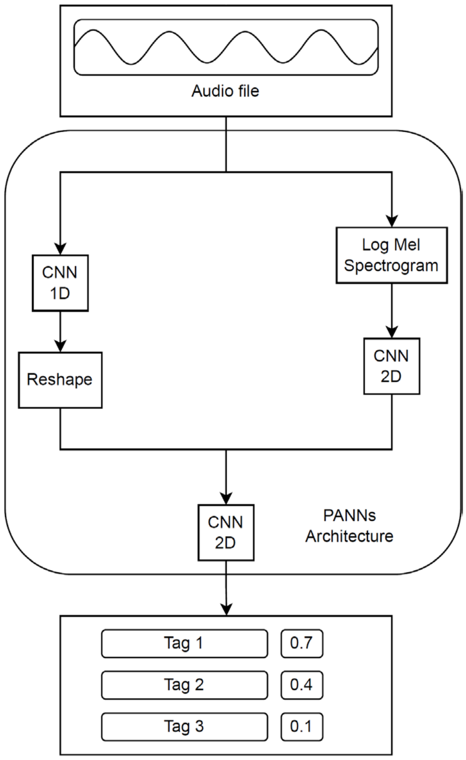

Figure 1. Diagram of the architecture of pre-trained audio neural networks (PANNs, see [12]).

The architecture of PANN has been applied to a dataset of urban sound events that includes recordings from public sources [13] as well as those taken by FuSA in Valdivia, Chile. The target dataset follows a customized taxonomy shown in Table 1.

Table 1. Customized urban sound event taxonomy used in the FuSA system.

Categories	Subcategories
Humans	Talking, screaming, crowd, others
Music	Music
Animals	Dog, bird, others
Environment	Rain, wind, waterfall, thunder, others
Mechanical	Impact, cutting, explosion, drilling, others
Vehicles	Motorcycle, car, bus and truck, helicopter and plane, others
Alerts	Siren, alarm, horn, bell, others

Users can post audio files of up to 60 seconds in various formats through the system's API through an HTTP request (public.labacam.org), as shown schematically in Fig. 2.

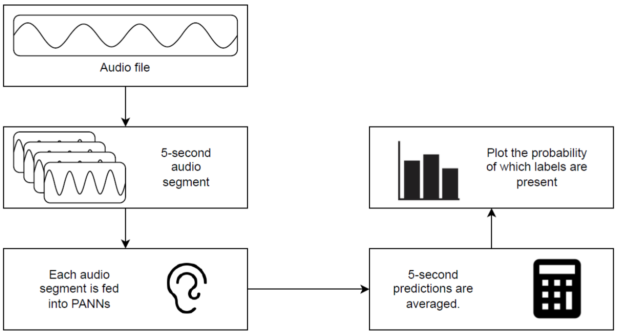
Audio file
5-second audio segment
Each audio segment is fed into PANNs
5-second predictions are averaged.
Plot the probability of which labels are present

Figure 2. Workflow diagram of the public.labacam.org application.

The PANNs model implemented in the web page is configured to receive 5-second audio segments so that audio signals that are longer are divided into 5-second pieces. The model predicts for each 5-second piece and calculates an average value to determine the overall presence of each tag in the audio input file. Finally, these values are displayed as a horizontal bar graph (see Figs. 3 and 4).

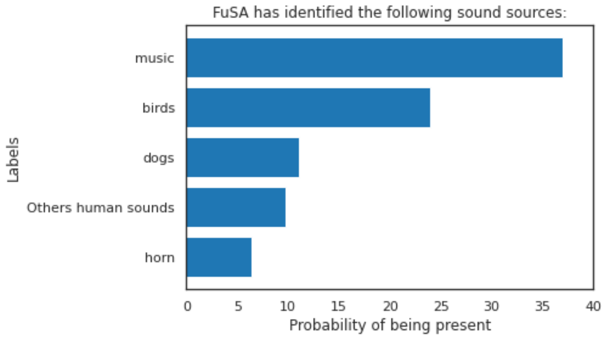

Figure 3. An example of the display of the overall presence of different sources in an audio input file where the sourced labelled as music and birds are prevalent.

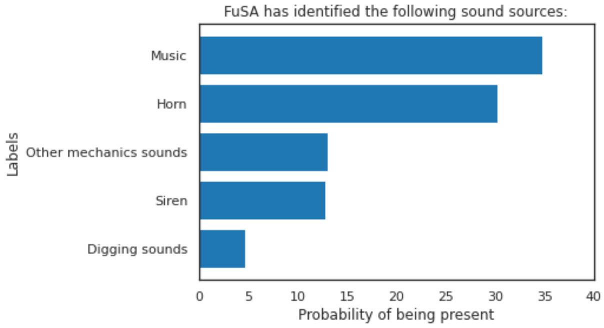

Figure 4. Another example of the display of the overall presence of different sources in an audio input file where the sources labelled as music and horn are predominant.

3. MAIN USERS OF THE FUSA SYSTEM

The FuSA system is designed to serve various purposes, a complex challenge due to its numerous applications. Therefore, three main user profiles have been created to address this issue:

- *Citizens:* individuals, neighborhood councils, students, NGOs, foundations, etc.
- *Companies:* acoustics consultants, technology developers, consultants in other engineering fields, etc.

- *Data analysts:* researchers, academic institutions, ministries, public services, policy makers, law enforcers, scholars, etc.

We have defined a protocol for interacting with the FuSA system for each type of user. Regardless of the user type, the maximum audio recording length that can be uploaded is 1 minute per file.

3.1 Case 1: Citizen

Example: "*I want to send acoustic information through my computer or phone to raise awareness about the acoustic problems in my sound environment.*"

The first use case corresponds to the ordinary citizen's ability to identify sounds of different natures. The user can upload an audio file to the FuSA system through an interactive web interface (FuSA Public)[1]. It is compatible with various devices such as computers or smartphones, being able to recognize GPS location, make recordings directly from the platform, listen to audio files to be uploaded, and indicate relevant parameters such as the date and time of recording, latitude, and longitude, etc.

Furthermore, the user can specify sound sources previously detected in the file through a drop-down menu with different boxes, which show the specific taxonomy of the FuSA system. Once loaded, the FuSA system delivers a window indicating the probability of confidence in recognition of the sound sources existing in the analyzed audio file.

3.2 Case 2: Specialized company

Example: "*I want to use artificial intelligence models to identify existing sound sources in audio data and records coming from my company's sensors.*"

Unlike the citizen case, companies can establish direct communication with the FuSA system to upload and consult audio files, replacing the graphical interface with an API (FuSA API)[2] that can be used through instructions called by commands configurable with any programming language. Through a terminal, a company can upload an audio file indicating relevant information such as format, location path, and specific characteristics related to the recognition models. Currently, the FuSA system has tested several models (PANN – SED - CRNN) and recognition schemes (PSED - TAG), uploaded directly to the project server and accessible through the API. The response of the FuSA system corresponds to the prediction of each of the sources, the confidence probability, and their start and end times.

3.3 Case 3: Data analyst

Example: "*I want to download a report from data uploaded to the FuSA system to analyze an environmental acoustics problem in a specific period and geographic location.*"

The final application case of the FuSA system involves a user with a researcher or data analyst profile whose interest is to explore audio data sets generated by acoustic sensors in a specific period and location. The user may draw conclusions about the acoustic behavior of the sound environment studied and generate relevant inputs such as noise maps, sound quality maps, environmental prevention and decontamination plans, welfare indicators, environmental education methodologies, etc.

After defining relevant parameters such as start and end dates of the period to be studied, latitude and longitude of the sensor used, and expected area of application, the FuSA system transforms the analyzed audio dataset into a CSV spreadsheet containing several indicators associated with the measurements, such as sound levels, probability of the presence of sound sources of interest, etc.

This last function of the system is only available upon request[3], given the specificity of the procedures to be assigned and the complexity of each potential application.

An example of the FuSA application has been presented in [14] for environmental noise analysis. Cities currently use 24-h noise measuring stations to monitor the noise levels in urban areas. The station audio recordings are essential for assessing source compliance with noise regulations in a specific city area when the noise exceed the permitted levels at sensitive receivers. However, the process is quite challenging because trained experts must examine a large amount of data by listening to each audio file individually. Additionally, external sources like animals and weather might have surpassed the allowed limits and could be unrelated to the primary noise source under monitoring. This fact leads to practical difficulty when enforcing noise regulations.

In this scenario, the stations can be the origin of the audio files fed into the FuSA system. Consequently, the FuSA system can deliver a prediction matrix reporting the

[1] https://public.labacam.org

[2] https://api.labacam.org

[3] https://www.acusticauach.cl/fusa/

presence of noise sources contributing to environmental noise pollution. An example of such a prediction matrix is shown in Fig. 5. The vertical axis displays the class probabilities for each 5-second segment of the 1-minute recording, whereas the horizontal axis shows the time. On the vertical axis, the probabilities sum up to one. The likelihood is higher when the color is darker. This outcome helps experts to promptly evaluate the acoustic environment's dynamics by identifying its most prominent sound events.

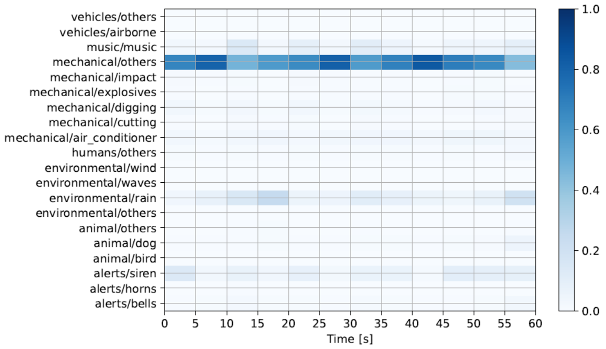

Figure 5. Prediction matrix for a recording made by a noise measuring station. Trained experts confirmed that this audio file corresponded to construction-related noise [14].

The results obtained using the machine learning tools developed in the FuSA project applying the current Chilean regulation on environmental noise revealed that the number of audio files that required expert analysis was reduced by 97%.

4. CONCLUSIONS

In this work, the main characteristics of a machine learning-based system, FuSA, were presented. The system automatically recognizes sound sources in audio files recorded in the urban environment to assist their analysis. In addition, some examples of users that could benefit from FuSA have been presented. So far, the results obtained by FuSA are pretty promising. However, some issues still need to be addressed, which are related to the robustness of the system and the own concerns associated with using artificial intelligence. These include data traceability to ensure no risks of data tampering and privacy. The latter is particularly important when processing speech-related sounds (human conversations), which may compromise people's privacy. It is expected that the FuSA system will assist in enforcing and monitoring strictly regulated environmental noise sources at sensitive locations to reduce noise pollution.

5. ACKNOWLEDGMENTS

This research was funded by ANID FONDEF grant number ID20I10333. P. Huijse acknowledges support from ANID FONDECYT grant number 1211374.

6. REFERENCES

[1] M.J. Bianco, P. Gerstoft, J. Traer, E. Ozanich, M.A. Roch, S. Gannot, and C.A. Deledalle: "Machine learning in acoustics: Theory and applications," *The Journal of the Acoustical Society of America*, vol. 146, pp. 3590–3628, 2019.

[2] J.P. Arenas: "How contemporary artificial intelligence became a hot topic in acoustics and vibration," *International Journal of Acoustics and Vibration*, vol. 28, no. 1, pp. 2–4, 2023.

[3] I. J. Goodfellow, Y. Bengio, A. Courville: *Deep Learning*. Cambridge, MA, MIT Press, 2016.

[4] Y. LeCun, Y. Bengio, and G. Hinton: "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[5] V. Morfi and D. Stowell: "Deep learning for audio event detection and tagging on low-resource datasets," *Applied Sciences*, vol. 8, 1397, 2018.

[6] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He: "A comprehensive survey on transfer learning," *Proc. of the IEEE*, vol. 109, pp. 43–76, 2020.

[7] R. Venkatesan and B. Li: *Convolutional Neural Networks in Visual Computing: A Concise Guide.* Boca Raton, CRC Press, 2017.

[8] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen: "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1291–1303, 2017.

[9] G. Licitra: *Noise Mapping in the EU: Models and Procedures*. Boca Raton, CRC Press, 2013.

[10] J. P. Bello, C. Silva, O. Nov, R. L. Dubois, A. Arora, J. Salamon, C. Mydlarz, and H. Doraiswamy:

"SONYC: A system for monitoring, analyzing, and mitigating urban noise pollution," *Communications of the ACM* , vol. 62, no. 2, pp. 68–77, 2019.

[11] J. Ardouin, L. Charpentier, M. Lagrange, F. Gontier, N. Fortin, D. Ecotiere, G. Guillaume, J. Picaut, and C. Mietllicky: "An innovative low cost sensor for urban sound monitoring," *Proc. Inter Noise 2018*, pp. 2226–2237, 2018.

[12] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M.D. Plumbley: "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE-ACM Transactions on Audio Speech and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[13] J. Salamon, C. Jacoby, and J.P. Bello: "A dataset and taxonomy for urban sound research," *Proc. of the 22nd ACM International Conference on Multimedia*, pp. 1041–1044, 2014.

[14] V. Carrasco, J. P. Arenas, P. Huijse, D. Espejo, V. Vargas, R. Viveros-Muñoz, V. Poblete, M. Vernier, and E. Suárez: "Application of deep learning to enforce environmental noise regulation in an urban setting," *Sustainability*, vol. 15, no. 4, 3528, 2023.