Enhancing Environmental Noise Management Through Siamese Convolutional Neural Network with Triplet Lost Function for Identification of Principal Sound Sources

Jean-Pierre Côté [*1], **Marc-André Gaudreau** [†1] **et Sousso Kelouwani** [‡1]
[1]Département de Génie mécanique, Université du Québec à Trois-Rivières

1 Introduction

In the realm of industrial environmental noise management, correctly identifying the source of the disturbance is paramount. This initial step not only sets the foundation for subsequent noise mitigation strategies but also aids in preventing further escalation of the noise problem. Furthermore, to optimize effectiveness and reduce costs, this process of identifying the noise source should be automated [1].

In a multi-source environment, a major obstacle in identifying primary noise sources is the convergence of multiple sound sources from inside as well as outside of the industrial zone, coupled with the degradation of acoustic signals over distance. Additionally, sources can be mobile, which adds to the intricacy of identifying the dominant noise sources from a distant field where the disturbance is measured.

This research aims to automatically link noise sources to environmental measurements from distant fields, providing a proof of concept for this complex task. While our larger project focuses on real-time feedback (Figure 1), this paper details our algorithm's development, results, and their significance.

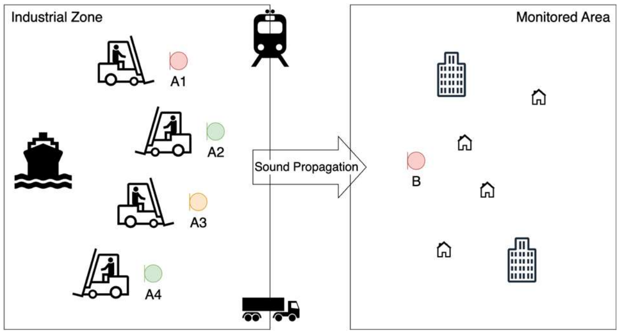

Figure 1: Schematic diagram of the studied scenario. A1-4 represents sound captures from potentially noisy mobile equipment. B is an environmental monitoring station. Colors at A indicate feedback to the user as to which source triggers B to surpass regulatory levels.

2 Method

The experiment we describe below involves the use of in the field recorded data to simulate real-time captures of close and far field versions of sounds. Those simulations are fed to an Artificial Neural Network (ANN) for machine learning.

2.1 Database

Our research project is sponsored and commissioned by a medium-sized port on the Saint-Lawrence River, active in the bulk and general cargo sector. At this port, 3 noise monitoring stations are set up to automatically capture 20-second sound clips when a specified noise level is exceeded. Over the past three years, these stations have recorded over 150,000 sound snippets, each stored in WAV file format. After an automatic processing to isolate loudest parts and eliminate irrelevancy, we randomly selected 5000, 2-second snippets from those files to serve as instances of sound A. The sounds, predominantly created by the movement of large metal beams and plates by lift trucks in an open space, are marked by metallic impacts, scrapes, intermittent backup alarms, engine noise, and passing trains. The database can be considered as varied.

2.2 Simulations

In order to train the ANN, we also require far-field captures (B) to be compared with the original sounds (A). We simulate these captures by generating new audio files that blend three original sounds with pink noise. While A1 is the reference, A2 and A3 are selected randomly for each twin set. For each of the individual source extracts, we generate 20 positive and 20 negative twin pairs. In positive twins, A1 is dominant, while in negative twins, A2 or A3 dominates.

2.3 Network

We propose a Siamese Convolutional Neural Network (SCNN) with triplet lost function, as described below (Figure 2).

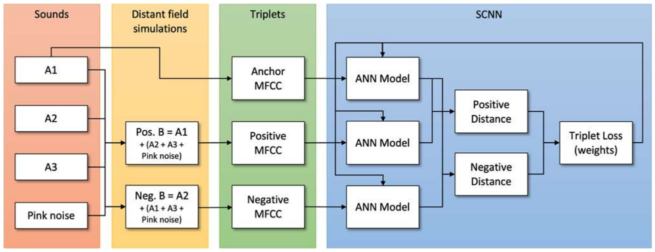

Figure 2: Schematic diagram of the learning process using a SCNN with triplet lost function.

Feature extraction

Based on its successes in automatic Environmental Sound Classification (ESC) [2], we use Mel Frequency Cepstral Coefficients (MFCCs) for feature extraction. The 2 second sound extracts are applied a 512 FFT pt (Hamming) with

[*] Jean.Pierre.Cote@uqtr.ca
[†] Marc-Andre.Gaudreau@uqtr.ca
[‡] Sousso.Kelouwani@uqtr.ca

hop size = 256 pt on 32 Mel filters to generate 20 - 1 cepstral coefficients, which produce a matrix of dimension 19 × 171 for each instance to feed the network.

Convolutional Neural Network model

Convolutional Neural Networks (CNNs) are a class of deep neural networks specifically tailored to process grid-like data where local patterns matter and there is a natural notion of translation invariance [3]. A key benefit of using CNNs in our research is their ability to identify the attributes of a sound notwithstanding the level of asychnonicity between A and B. The ANN architecture that we use is adapted from the SEnv-Net CNN model [4] that was developed and optimized for ESC (Figure 3).

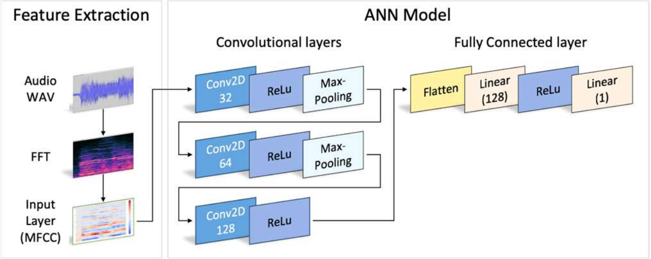

Figure 3: Schematic diagram of the ANN model which uses 3 convolutional layers and 2 linear layers.

Siamese network

For this research, sound A1 needs to be matched to sound B, while also being a component of sound B when another sound, A2 or A3, is the correct match. Therefore, straightforward classification is not the goal; discrimination is.

Siamese networks are a special type of neural network architecture used primarily in tasks involving the comparison of two distinct inputs [5]. They are used to match pairs of data and are a solution when the number of classes is very large and unknown during training, as is our case. The purpose of a Siamese network is to judge how similar the inputs are, rather than classify inputs independently of each other.

The key characteristic of Siamese networks is that they have identical subnetworks, which means they share the exact same parameters and weights. Each of these subnetworks processes one of the inputs. The subnetworks are joined at the top by a layer (or several layers) that combine the outputs of the subnetworks and compute a final output.

Triplet lost function

An extension of the Siamese network is the triplet loss, where a third input (a negative sample) is used along the anchor and the positive sample to improve the discrimination power of the network [6].

The use of the triplet loss function enables the neural network to process triplets of data in which sound A1 serves as the anchor that is compared with a positive instance of sound B and a negative instance of sound B. The goal of the function is to calculate and minimize the distance with positive B and maximize the distance with negative B.

3 Results and Discussion

Performance measured on the test set yield 70% precision for S/N = 0 dB, up to 84% for S/N = 10 dB.

Table 1: Accuracy of the correct identification achieved over signal-to-noise ratios.

S/N (dB)	Accuracy (%)
10	84.29
6	81.20
0	69.68
-6	54.19

The increase in precision with the rise in the signal-to-noise ratio suggests that the algorithm successfully identifies principal individual sounds in simulations of far-field complex signals.

4 Conclusion

Our ongoing research suggests a promising pathway towards balancing the need for maximizing production with adherence to environmental noise standards. If our method proves successful, it could contribute to the automatic identification of the principal sound source in a noisy and desynchronized complex signal. This tool could become useful in different situations and fields.

Acknowledgments

This work was supported by Mitacs through the Mitacs Accelerate program. It is commissioned and sponsored by the Administration du Port de Trois-Rivières and Logistec, with the help of the Cégep de Drummondville.

References

[1] Murovec, J., et al. (2023). Automated identification and assessment of environmental noise sources. Heliyon, 9(1).

[2] This subject is addressed in numerous studies, one of which is Jin, S., et al. (2021). Evaluation and modeling of automotive transmission whine noise quality based on MFCC and CNN. Applied Acoustics, 172, 107562.

[3] A fundamental paper on this topic is LeCun, Y., et al. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

[4] Al-Hattab, Y. A., et al. (2021). Rethinking environmental sound classification using convolutional neural networks: optimized parameter tuning of single feature extraction. Neural Computing and Applications, 33(21), 14495-14506.

[5] A fundamental paper on this topic is Chopra, S., et al. (2005, June). Learning a similarity metric discriminatively, with application to face verification. In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05) (Vol. 1, pp. 539-546). IEEE.

[6] A fundamental paper on this topic is Schroff, F., et al. (2015). Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 815-823).