International Journal of Applied Science and Engineering

Special Issue –Convergent Technologies for Innovative Learning Environment and Information Systems

Sound detection for study room monitoring and evaluation

Ricardo Catanghal Jr*

College of Computer Studies, University of Antique, Antique, Philippines

ABSTRACT

In the field of sound recognition, the research and study of sound event detection are still active, with the vast majority of the papers focusing mostly on the domain of speech and music. This paper presents and discusses a framework for a study room event detection system. Feature extraction techniques are utilized and discussed to obtain the parametric type representation for the analysis of the sound for intelligent homes machine listening systems specifically for study room. The conduct of sound analysis within the category of the sounds the least accurate was the door knock, but the accuracy of 95.00%, currently in the field is acknowledged as good, making the parameters fit for detecting surrounding sounds. The performance of the CNN in detecting environmental sounds was analyzed using the parameters that were defined, with an overall accuracy of 96.8%. The result was promising for machine learning that detects sounds that can be applied as technology for an innovative learning environment.

Keywords: Smart homes study room, Innovative learning technologies, Machine learning, Artificial neural network (ANN).

OPEN ACCESS

Received: February 1, 2021
Revised: April 11, 2021
Accepted: May 11, 2021

Corresponding Author:
Ricardo Catanghal Jr
racatanghal@antiquespride.edu.ph

Publisher:
Chaoyang University of Technology
ISSN: 1727-2394 (Print)
ISSN: 1727-7841 (Online)

1. INTRODUCTION

Several international interests and working groups have looked at the potential usefulness and advancement of the acoustic classification. Research suggests that developing signal processing methods to extract this information has enormous potential in several applications automatically, for example searching for multimedia based on its audio content, making context-aware mobile devices, robots, cars, and many others, and intelligent monitoring systems to recognize activities in their environments using acoustic information (Yamakawa et al., 2011). However, a significant amount of research is still needed to reliably recognize sound scenes and individual sound sources in realistic soundscapes.

The convolutional neural network (CNN) gained its popularity when it achieves an error rate of only 16% in the ImageNet Large Scale Visual Recognition Challenge, a significant improvement from previous years' performance. Since then, researchers have tried to apply in different fields and areas to see the results and measure the performance of the CNN. The area of acoustic classification and recognition is one of the newly explored areas (Catanghal, 2020).

The predominantly research on audio recognition has concentrated basically on music and speech. In the past consideration, was taken into the area of research in sound pattern recognition as the beginning or stagging point for the reason that if we know more about how humans understand or interpret those sound, we could make intelligent machines understand better of what they hear, in the sense of being able to extract and analyze purposive and meaningful information from it (Catanghal et al., 2019). For the research to encourage and arouse in machine listening for general audio environments, in 2012–2013 IEEE Audio and Acoustic Signal Processing Technical Committee organized a research challenge on Detection and Classification of Acoustic Scenes and Events, that

sounds from the environment or no-speech/non-music (Stowell et al., 2015).

There are various attempts to utilize the information extracted from different kinds of sounds to enhance various types of applicability, for example, lifelogging automated systems, systems for monitoring infants or elderly persons, automated surveillance systems, and multimedia sound retrieval systems. The methodology or approach used in these smart systems includes acoustic scene analysis that analyzes scenes regarding the places, situations, and user activities they depict and acoustic event analysis that examines various sounds, for example, traffic rumbles, footsteps, shattering of glass, thunder, and scream (Catanghal et al., 2019).

In this study, several sound detection techniques, and algorithms were combined to create a system that can identify the sounds in the surroundings that would be useful in the study room. Sounds carry a large amount of information about our everyday environment and physical events that take place in it, being able to detect which are the sound sources present in the signals would further increase the usefulness of any audio recording. Several acoustic properties were also combined to form the features, Mel-frequency cepstrum (MFCC), Chroma, Mel Spectrogram, and Contrast. The surrounding sounds are the most abundant, yet very few studies were conducted, this will be another contribution in the field of acoustic sound environment classification and innovative learning environment.

2. BACKGROUND AND METHODS

In the field of pattern recognition, the sound processing term refers to the way how sound wave is transformed to emphasize different features. Under the term sound transformation, one understands a wide field of different sound manipulation techniques for different purposes. Signals perceived by humans through their hearing systems are called audio signals. Those signals come from a sound source that vibrates in the audible frequency range of approximately between 20 Hz and 20,000 Hz. The resulting vibrations are causing different pressure (amplitude) in a medium (usually air) which causes the human eardrum to vibrate and send the information to the brain for interpretation (Catanghal, 2019).

What we can learn from the signal processing subfield has very similar processing algorithms despite having different purposes. Examples include the invention of wavelets and their immediate translation into different areas, or the development of the fast Fourier algorithm. In general, all signal recognizers share the same workflow consisting of three main components: signal acquiring, signal processing, and signal recognition as depicted in Fig. 1. The signal acquiring step deals with the raw sounds and the preparation for processing. The signal processing step involves the different transformation of the raw signal for the machine learning to process. The final step, signal recognition, deals with machine learning recognizing the sound.

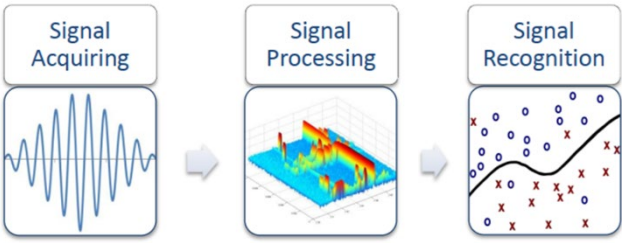

Fig. 1. General steps of signal processing (Catanghal, 2020)

Environmental sound classification is an exciting field and is now a growing research problem with many applications. The environmental sounds are a very diverse group of everyday audio events that cannot be described as only speech or music (Agrawal et al., 2017).

The top five environmental sounds for the smart homes from the survey conducted by Audio Analytica (2018) are dog, crying baby, door knock, alarm, and glass breaking. These five sounds play an essential role in smart homes, as suggested by the consumer survey. Consumers see that smart homes with an AI for acoustics are of considerable value in providing them assistance in the specifics in the smart homes such as security, safety, child monitoring, assisted living, and pet monitoring.

The convolutional neural networks (ConvNets or CNN's) are analogous to traditional artificial neural networks (ANN) in that they are comprised of neurons that self-optimize through learning. Each neuron will still receive input and operate (such as a scalar product followed by a non-linear function) - the basis of countless ANNs. From the input raw image vectors to the final output of the class score, the entire network will still express a single perceptive score function (the weight). The last layer will contain loss functions associated with the classes, and all of the regular tips and tricks developed for traditional ANNs still apply (O'Shea and Nash, 2015).

The sound datasets from this study were gathered following the methods of Salamon et al. (2014) with modifications to fit in this study. Five sounds were chosen based on the ranked most important for smart homes from the survey conducted by Audio Analytic (2018), namely: dog, crying baby, door knock, alarm, and glass breaking. These five sounds play an essential role in smart homes, and as the survey suggests that consumers or people surveyed agree with this (Audio Analytic, 2018).

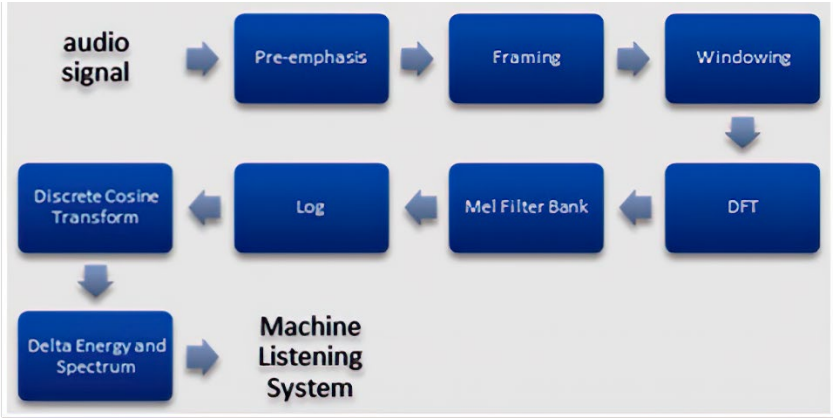

Fig. 2. Block diagrm of MFCC (Catanghal, 2019)

The listening machine concept is termed as extracting valuable or useful information from the sound that it acquires as described in Fig. 2. The fundamental architectural design of the machine listening system is summarized and characterized into a general model of three steps: acquisition of signal or capturing through a device, signal processing, and recognition of the signal. In an experiment with regards to human hearing perception, it is found out that the ears of human beings function as a bank of subband filters or filter banks. The pre-emphasis is useful to avoid numerical problems during the Fourier transform operation and improvement of the signal to noise ratio (SNR). The signal needs to be divided into short time frames after the pre-emphasis because frequencies change over time in a signal. One of the reasons of windowing is to avoid or minimize the spectral leakage and to cancel out the inference by the fast Fourier transform (FFT) that our signal is infinite.

The model that will be utilized in the concept for the CNN is the sequential and a 1D convolutional neural network. The sequential model allows us to create or establish layer by layer of the model. The following parameters were utilized during the training of the dataset or the machine learning: the epoch was set to 1500, the batch size is 64, the learning rate was set to 0.01, stochastic gradient descent was used as an optimizer, the rectified linear unit or ReLU was used as supplementary activation.

The layers will be arranged sequentially as described earlier which comprises the following: the neural network layer, max-pooling layer, dropout layer. The first two layers which consist of nodes define a filter or sometimes referred to as a feature detector of the height of the kernel size. On the other hand, the max-pooling layer will serve as a deterrent to the data from overfitting and curtail the complicatedness of the data. The dropout rate was set to 50% or 0.5, in doing this, the neural network (NN) will less responsive to any small amount of changes in the data.

As seen in Fig. 3, the sound detected and recognized will be feed to the innovative learning environment which is an overall system. This main system will control other subsystems, for instance, there is a knocking on the door, the camera will be enabled to detect a person. Take for another instance, a dog barking or broken window, the camera will be enabled on the area to detect any person before informing the person in the room.

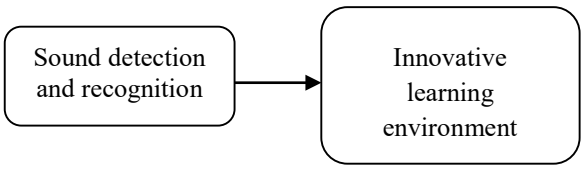

Fig. 3. Block diagram of sound detection to innovative learning environment

3. RESULTS AND DISCUSSION

We have demonstrated that a machine learning method via the CNN model can be useful for detecting sounds which can be applied as technology for an innovative learning environment. To extract the data from sound, the first feature that we extract is the spectrogram, by performing some exploratory visual analysis, we can see the different pattern compared to that of the waveform, it reveals some pattern or characteristics (Foggia, et al., 2016). Some very distinct sound sources can be easily distinguished from other classes by quick visual inspection and some are quite similar in their spectral content. Part of this comes from their actual perceptual closeness, but most of it is due to the sub-optimal representation of the data.

Analyzing, the result of the essential tool for evaluating models, to have a better understanding of the performance (Catanghal, 2020): the confusion matrix. The confusion matrix presents the summary of the prediction results and where it was confused when making a prediction. Let us start with the analysis on the “baby crying” it has an accuracy of 99.00%, the machine learning model that we have correctly labeled every “baby crying” as it is, with a confidence of 99.73%. Although some of the broken glass is classified as baby crying, bringing the precision to only 95%, see Fig. 4.

The “broken glass” has a true positive value of 36, which has an accuracy of 96.00% is considered still in a good

performance. The model can correctly identify 38 different broken glass sounds out of the forty samples. Although the model was confused with broken glass as baby crying, alarm, and door knock having a false positive rate of 2.50%, it still has a precision of 90.00%. The machine learning model is confused with other sounds such as door knock, alarm clock, and dog barking as broken glass. In the confusion matrix out of the four confused, 50% is identified as the dog.

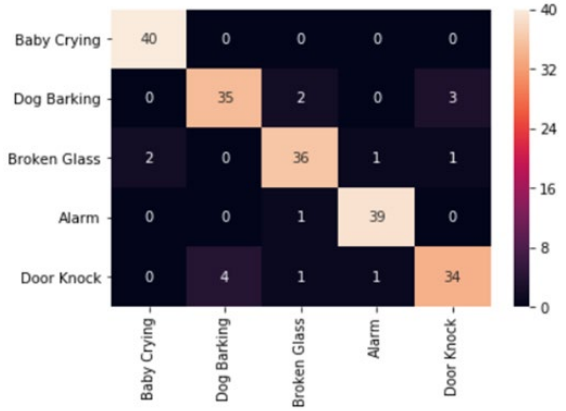

Fig. 4. Confusion matrix (Catanghal, 2020)

The impact of sound detection on innovative learning environments is very promising as it promotes the continuous learning of individuals and thus concentration and focus on the study increase the learning process. The detection of sound in the house then coupled with the smart home system lessens the disturbance of the student or individual, for instance, the knocking of the door some are just only a prank, with this sound detection then focus of the camera on the area to detect the person is very crucial. This sound detection to an innovative learning environment has a significant role in increasing the learning process of the individual.

4. CONCLUSION

Sounds carry a large amount of information about the everyday environment and physical events that take place in it. Developing signal processing methods to extract this information automatically has enormous potential in several applications, specifically smart homes in general. The performance level of the convolutional neural network in identifying the environmental sounds using the parameters that we defined yields an excellent overall accuracy of 96.8%. This gives the model an excellent accurate prediction in identifying the given environmental sounds in the area of machine learning and is useful in different applications.

This will help integrate with the study room as a precautionary measure of any undesirable sounds. Utilizing machine learning in detecting the sound can further increase the concentration of the learner in the study room while the machine filters what needs only to be heard.

ACKNOWLEDGMENT

The researcher would like to thank the support of the University of Antique and for funding this research.

REFERENCES

Agrawal, D.M., Sailor, H.B., Soni, M.H., Patil, H.A. 2017. Novel teobased gammatone features for environmental sound classification. European Signal Processing Conference, 1809–1813.

Ahmad, S., Agrawal, S., Joshi, S., Taran, S., Bajaj, V., Demir, F. 2020. Environmental sound classification using optimum allocation sampling based empirical mode decomposition. Physica A: Statistical Mechanics and its Applications, 537.

Audio Analytica, 2018. The 2018 Smart Home Report: AI attitudes and expectations. audioanalytica.com.

Catanghal, R.A., Palaoag T., Dayagdag, C. 2019. Environmental acoustic transformation and feature extraction for machine hearing. Proc. IOP Conference Series: Materials Science and Engineering, 482, 012007.

Catanghal, R.A. 2019. A framework for home machine listening system with convolutional neural network. International Journal of Simulation: Systems, Science & Technology.

Catanghal, R.A. 2020. Behavior analysis of convolutional neural network for environmental sound. Journal of Innovative Technology Convergence, 2, 103–110.

Foggia, F., Petkov, N., Saggese, A., Strisciuglio, N., Vento, M. 2016. Audio surveillance of roads: a system for detecting anomalous sounds. IEEE Transactions on Intelligent Transportation Systems, 17, 279–288.

O'Shea, K., Nash, R. 2015. An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458

Salamon, J., Jacoby, C., Bello, J.P. 2014. A dataset and taxonomy for urban sound research. in Proceedings of the ACM International Conference on Multimedia. ACM, 1041–1044.

Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M., Plumbley. M.D. 2015. Detection and classification of acoustic scenes and events. IEEE Transactions on Multimedia, 17, 1733–1746.

Yamakawa, N., Takahashi, T., Kitahara, T., Ogata, T., Okuno, H. 2011. Environmental sound recognition for robot audition using matchingpursuit. Modern Approaches in Applied Intelligence, 1–10.