Intelligent Systems and Computer Technology
D.J. Hemanth et al. (Eds.)

doi:10.3233/APC200135

A Deep Learning Based System to Predict the Noise (Disturbance) in Audio Files

Kartik P V S M S [a,1], and Jeyakumar G[b]
[a,b]*UG scholar, Dept of CSE, Amrita School of Engineering, Coimbatore*

Abstract. Generally, people prefer their audio to be with very good clarity. They want no disturbances during any interaction and while listening audio files. Automated systems to remove disturbances in an audio file to bring good clarity real time audio communications are in high demand. In this paper a deep learning model to detect the noises in a given audio file is proposed and it working principle is explained. The proposed model was trained, first, to predict the places of noise in the audio file by a well-defined training set which consists of set of audio files with the interval of clear audio and noise. After training, the proposed model predicts the area of disturbance (noise) in any given audio file using the integrated techniques of deep learning and audio processing, and the results are reported. The prediction accuracy of the model was found 90.50 %.

Keyword. Audio Processing, Noise Detection, Deep Learning, Prediction Accuracy.

1. Introduction

Sounds are one of the properties of environment. Accessing various aspects of sounds from the environment, studying their patterns and proposing ideas for reducing the sound noises are the upcoming research areas in the environmental studies. The environmental noise is defined as unwanted or harmful outdoor sound created by human activities. The sources for this noise are traffic, industry, airport and activities such as construction, rock-crushing and recreating activities etc. The noise level in environment is increasing year after the year, which is harmful for health of people. It is experienced by people that noise exposure not only affects their health, but can also affect their social and economic aspects. Hence, modern environmental noise monitoring systems that can continuously measure the sound levels and different noise sources are essential for present environmental conditions. Monitoring environmental noise, over a range of time and space, and making a comprehensive report which can be used further for designing novel noise detection algorithms is a challenging task. Because it involves high cost equipment and man power. However, investigating and implementing suitable automated systems for this purpose will save considerable amount of manual work and effort. The primary objective of this paper is to propose a system which can automatically detect noise in the original audio source. The Deep Learning (*DL*) methods have become popular nowadays for the conventional data analysis techniques, especially with video, image and audio dataset. As well as, these

[1]Kartik P V S M S ,UG scholar,Dept of CSE ,Amrita school of Engineering, India.
Email: cb.en.u4cse17033@cb.amrita.students.edu

methods enables faster assess of large set of attributes from the source data. This paper proposes to use a *DL* model to present a simple but elegant noise detect algorithm from given audio files.

2. Related works

Evaluation of sound event detection, classification and localization of hazardous acoustic events in the presence of background noise of different types and restoring the original audio are the challenging and interesting research problems associated with acoustic research. This section summarizes various research articles in the literature presenting above acoustic research problems.

A work describing different approaches for anomalous noise detection with low-cost sensors is presented in [1]. This study includes both the synthetic and real-life acoustic data. In [2], the authors have proposed an acoustic pattern classification algorithm which can automatically detect different noise sources. This algorithm was trained using manually annotated recordings and was used to detect a target noise source in the presence of interfering noise sources. An automatic noise recognition system to monitor the noises due to aircraft is presented in [3]. This system does feature extractor and training-recognition using Hidden Markov Model. An approach deals with the problem of localization of impulsive disturbances in non-stationary multivariate signal is presented in [4]. The proposed approach was tested in two practical applications – elimination of impulsive disturbances from audio files and robust parametric spectrum estimation. A sound source localization algorithm modeled based on the analysis of multichannel signals from the Acoustic Vector Sensor is presented in [5]

[6] presents an Anomalous Noise Event Detector to differentiate Road Traffic Noise and Anomalous Noise Events. This detector was developed using a distributed low-cost acoustic sensors of Wireless Acoustic Sensor Networks. This detector uses a two-class audio event detection and classification approach.

A non-negative matrix factorization based approach to adaptive noise reduction for sound events is presented in [7]. This approach first employs a noise dictionary learning process which supports the source separation framework construction. Then, a separation process to filter the target sound event from the noise is carried out.

The work investigating the problem of detecting hazardous events on roads by designing an audio surveillance system that automatically detects perilous situations such as car crashes and tire skidding, is presented in [8]. This work combines time-domain, frequency-domain, and joint time-frequency features to detect events on roads.

A two stage approach which includes detection and interpolation is presented in [9]. This approach is to restore the audio signals corrupted by noises. In detection stage the degraded audio signal is taken as input and detects the locations of the degraded samples. In the interpolations stage the degraded locations are replaced with appropriate values.

There are attempts in the literature to detect and restore the audio signals corrupted by random-valued impulse noise. The work presented in [10] is an example. The authors have analyzed the noisy samples with a cascading of stages and compared

the similarity degree between the samples and their neighbors. Detecting sound events in audio streams is a widely used application. A solution for sound event detection in mobile platform is presented in [11]. This detection process included acquisition of sensor data, processing of audio signals, and detecting and recording of sound events.

Using neural networks for above said research works are also widely tried out and are still in further developments. In such cases finding suitable training sets and test tests of audio files is the challenging task. The issue of acoustic mismatch between the noisy training set and the test set, due to reason that they are different sources, is addressed in [12]. The authors have proposed a dataset and a base line system for foster label noise research. In [13], a study on creating Artificial Neural Networks (*ANN*) and Recurrent Neural Networks (*RNN*) models to classify the sound sources from a pool of sound clips collected at various streets is presented. An existing audio set is used as training data set in this work. The trained model is tested on classifying the sound classes present in the urban streets. Considering the challenges in the video surveillance systems, the audio surveillance systems are in development to render suitable supports to the video surveillance systems. Monitoring and detecting dangerous audio events is very important in audio surveillance systems. A significant task is to detect the impulsive noises. Computationally efficient methods for detecting non-Gaussian impulsive noise in digital speech and audio signals are presenting in [14]. The authors in [15] presented a review report for the impulsive sound detection algorithms. The sounds from dangerous events such as gunshots, explosion and human screaming are classified as impulse sounds. Algorithms relating the impulsive noise detection with impulsive sound detection are also highlighted. A median filter based detection system for automatic detection and recognition of impulsive sounds is proposed in [16]. The proposed system was evaluated with a sound database with more than 800 signals recorded from noise environment. Indoor audio monitoring software to detect impulse sounds is proposed in [17]. This software has three stages of implementation (1) audio acquisition (2) preprocessing and (3) sound detector. The deep learning and machine learning approaches have become common and most popular problem solving tools nowadays. They are showing their superiority in solving complicated problems around us, to cite but few references are given in [18], [19], [20] and [21][24]. In spite of all the above related works in acoustic research, this paper proposes a system with a deep learning neural network model to predict the presence or absence of disturbances (noises) in the given audio files. The working structure of the proposed system and the results obtained implementing it are explained in the subsequence sections.

3. Proposed System

The proposed system works as follow.

1. A huge audio file in '.wav' format is given as input to the system. s
2. The audio file is broken into multiple small audio files using a given specified interval.
3. The small audio files are labelled as 'Speech' or 'Not speech'.
4. Each audio file is converted into Numpy array by using MFCCS library.
5. The converted arrays are passed to the model for training. The MFCCS by default returns 26 features, each row holds 1 feature vector.

6. The model is trained to get the prediction (Binary classification).

3.1. Model description

A sequential model is used in the proposed system, as sequential models are generally good for deep learning models.

a) The first layer is a dense layer with a total of 25 neurons. The dense layer is a classic fully connected neural network layer, in this each input node is connected to each output node. A dense layer represents a matrix vector multiplication (assuming the batch size is 1). The values in the matrix are the trainable parameters which get updated during back propagation.

$$(U\text{^}\mathrm{T}).\mathrm{W},\ \mathrm{W} \in (R\text{^}(\mathrm{n} \times \mathrm{m}) \qquad (1)$$

The first layer generates a *m* dimensional vector as output. It changes the dimension of the given vector, mathematically, applying rotation, scaling and translation transformation to it.

(a) The second layer is again another dense layer with 15 neurons.

(b) The third layer is the final output layer with 1 neuron, as the system is doing a binary classification. The classes are class A – Audio/Speech and class B – Disturbance/Non Speech. They are represented with '1' and '0', respectively.

The configuration of the model is described in Table 1.

Table 1. Model Configuration.

Model Content	Details
First fully connected Layer	25 nodes, ReLU
Second fully connected Layer	15 nodes,ReLU
Output Layer	Binary classification and Sigmoid Activation
Optimization function	Adam
Learning Rate	0.01
Metrics	Accuracy

4. Results and discussions

The performance of the proposed system which predicts the disturbances in the given audio file is measured using the accuracy of prediction. In initial stage, during the training process, the data is split in to training data and validation data. In this experiment, 30% of the data is kept for validation. At the training stage, after each epoch, the model is evaluated on the validation data. After each epoch, the validation accuracy and validation loss are also measured along with training accuracy and training loss. The values measured are presented in Table 2.

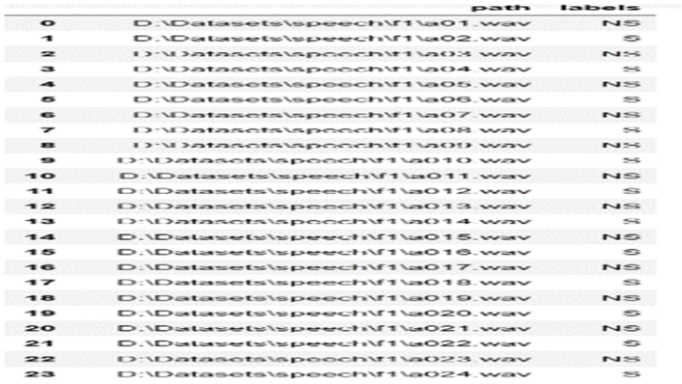
Table 2. The Values Measured For The Performance Metrics

Metrics	Value
Training Accuracy	90.50 %
Training Loss	0.26
Validation Accuracy	83.29%
Validation loss	0.35

The optimizer used in the proposed system is '*Adam'*. It is an adaptive learning rate optimization algorithm designed specifically for training deep neural networks. The loss is calculated on Binary Crossentropy. It is a loss function used on problems involving yes/no (binary) decisions. To use Binary Crossentropy the Sigmoid activation function is to be used in the previous layer.The model in the proposed system is trained with a batch size of 32. In each epoch, 32 nodes get trained at once. It trains 1282239 training files which are in the form of numpy array. At the time of prediction with the test data, the system generates a series of 0s and 1s in the interval of 10 ms, denoting the '*Audio'* or '*Disturbance'* classes (0 refers to Disturbance and 1 refers to audio). The intervals and their classification are depicted, as a sample, in Table 3. A sample output showing the classification after the Audio files are split based on the intervals is depicted in Figure 1. The accuracy scores obtained by the system after each epoch is presented in Figure 2.

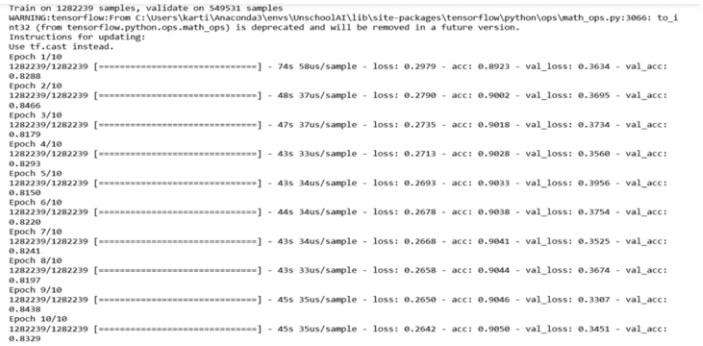
Table 3. The Intervals And Their Classifications.

Sno	Left	Right	Label
0	0.00	2.50	NS
1	2.50	3.31	S
2	3.31	5.47	NS
3	5.47	8.33	S
4	8.33	11.05	NS
5	11.05	11.38	S
6	11.38	15.55	NS
7	15.55	16.24	S
8	16.24	20.23	NS
9	20.23	24.38	S
10	24.38	28.03	NS
11	28.03	31.25	S
12	31.25	34.41	NS
13	34.41	35.50	S
14	35.50	41.59	NS
15	41.59	43.58	S
16	43.58	50.11	NS
17	50.11	51.43	S
18	51.43	51.44	NS



	path	labels
0	D:\Datasets\speech\f1\a01.wav	NS
1	D:\Datasets\speech\f1\a02.wav	S
2	D:\Datasets\speech\f1\a03.wav	NS
3	D:\Datasets\speech\f1\a04.wav	S
4	D:\Datasets\speech\f1\a05.wav	NS
5	D:\Datasets\speech\f1\a06.wav	S
6	D:\Datasets\speech\f1\a07.wav	NS
7	D:\Datasets\speech\f1\a08.wav	S
8	D:\Datasets\speech\f1\a09.wav	NS
9	D:\Datasets\speech\f1\a010.wav	S
10	D:\Datasets\speech\f1\a011.wav	NS
11	D:\Datasets\speech\f1\a012.wav	S
12	D:\Datasets\speech\f1\a013.wav	NS
13	D:\Datasets\speech\f1\a014.wav	S
14	D:\Datasets\speech\f1\a015.wav	NS
15	D:\Datasets\speech\f1\a016.wav	S
16	D:\Datasets\speech\f1\a017.wav	NS
17	D:\Datasets\speech\f1\a018.wav	S
18	D:\Datasets\speech\f1\a019.wav	NS
19	D:\Datasets\speech\f1\a020.wav	S
20	D:\Datasets\speech\f1\a021.wav	NS
21	D:\Datasets\speech\f1\a022.wav	S
22	D:\Datasets\speech\f1\a023.wav	NS
23	D:\Datasets\speech\f1\a024.wav	S

Figure 1. Sample Output For Classification.



```
Train on 1282239 samples, validate on 549531 samples
WARNING:tensorflow:From C:\Users\karti\Anaconda3\envs\UnschoolAI\lib\site-packages\tensorflow\python\ops\math_ops.py:3066: to_i
nt32 (from tensorflow.python.ops.math_ops) is deprecated and will be removed in a future version.
Instructions for updating:
Use tf.cast instead.
Epoch 1/10
1282239/1282239 [==============================] - 74s 58us/sample - loss: 0.2979 - acc: 0.8923 - val_loss: 0.3634 - val_acc:
0.8288
Epoch 2/10
1282239/1282239 [==============================] - 48s 37us/sample - loss: 0.2790 - acc: 0.9002 - val_loss: 0.3695 - val_acc:
0.8466
Epoch 3/10
1282239/1282239 [==============================] - 47s 37us/sample - loss: 0.2735 - acc: 0.9018 - val_loss: 0.3734 - val_acc:
0.8179
Epoch 4/10
1282239/1282239 [==============================] - 43s 33us/sample - loss: 0.2713 - acc: 0.9028 - val_loss: 0.3560 - val_acc:
0.8293
Epoch 5/10
1282239/1282239 [==============================] - 43s 34us/sample - loss: 0.2693 - acc: 0.9033 - val_loss: 0.3956 - val_acc:
0.8150
Epoch 6/10
1282239/1282239 [==============================] - 44s 34us/sample - loss: 0.2678 - acc: 0.9038 - val_loss: 0.3754 - val_acc:
0.8220
Epoch 7/10
1282239/1282239 [==============================] - 43s 34us/sample - loss: 0.2668 - acc: 0.9041 - val_loss: 0.3525 - val_acc:
0.8241
Epoch 8/10
1282239/1282239 [==============================] - 43s 33us/sample - loss: 0.2658 - acc: 0.9044 - val_loss: 0.3674 - val_acc:
0.8197
Epoch 9/10
1282239/1282239 [==============================] - 45s 35us/sample - loss: 0.2650 - acc: 0.9046 - val_loss: 0.3307 - val_acc:
0.8438
Epoch 10/10
1282239/1282239 [==============================] - 45s 35us/sample - loss: 0.2642 - acc: 0.9050 - val_loss: 0.3451 - val_acc:
0.8329
```

Figure 2. The Accuracy Score Measured.

5. Conclusions

A system to detect noises in the given audio file, which was proposed, is presented in this paper. The proposed system detects and predicts the intervals in the audio file with noise. For the given audio file which is divided in to equal size small wave files with equal interval, the proposed system generates a sequence of 0s and 1s based on the presence or absence of disturbance at each interval. The performance of the proposed system is evaluated using the metrics – training accuracy, training loss, predication accuracy and predication loss.The performance of the proposed system can be further improved employing additional layers and features in the neural network. This system can be extended to delete the noise once it is detected from the source audio file. This is much useful in preserving important audio files with confidential information and listening to real time audio communications without any disturbance.

References

[1] FrancescAlías, Rosa Ma Alsina-Pagès, FerranOrga, and Joan ClaudiSocoró.Detection of Anomalous Noise Events for Real-Time Road-Traflc Noise Mapping: The DYNAMAP's project case study.*Research Article,Noise Mapp*, Vol. 5, p. 71-85, 2018.

[2] PanuMaijala, Zhao Shuyang, Toni Heittola and Tuomas Virtanen.Environmental noise monitoring usingsource classification in sensors.*Applied Acoustics*, Vol. 129, pp. 258-267, 2018.

[3] M. Sabri, J. Alirezaie and S. Krishnan.Audio noise detection using hidden Markov model.IEEE Workshop on Statistical Signal Processing, pp. 637-640, 2003.
[4] Maciej Niedźwiecki and Marcin Ciołek. New Semicausal and Noncausal Techniques for Detection of Impulsive Disturbances in Multivariate Signals With Audio Applications. IEEE Transactions on signal Processing, Vol. 65., No. 15., 2017.
[5] Łopatka, Kuba & Kotus, Józef & Czyżewski, Andrzej. Detection, classification and localization of acoustic events in the presence of background noise for acoustic surveillance of hazardous situations. Multimedi Tools and Applications, Vol. 429., No. 17., 2015.
[6] Joan ClaudiSocoró, FrancescAlías and Rosa Ma Alsina-Pagès. An Anomalous Noise Events Detector for Dynamic Road Traffic Noise Mapping in Real-Life Urban and Suburban Environments. Sensors (Basel),Vol. 17., No. 10., 2017.
[7] Qing Zhou, Zuren Feng and Emmanouil Benetos, "Adaptive Noise Reduction for Sound Event Detection Using Subband-Weighted NMF", *Sensors (Basel),* Vol. 19., No. 14., 2019.
[8] Noor Almaadeed, Muhammad Asim, Somaya Al-Maadeed, Ahmed Bouridane and Azeddine Beghdadi. Automatic Detection and Classification of Audio Events for Road Surveillance Applications. *(Basel)*,Vol. 18., No. 6., 2018.
[9] Laurent Oudre. Automatic Detection and Removal of Impulsive Noise in Audio Signals. IPOL Image Processing On Line, Vol. 5., pp. 267-281., 2015.AliAwad.Impulse noise reduction in audio signal through multi-stage technique. Engineering Science andTechnology, an International JournalVol. 22., pp. 629-636., 2019.
[10] AliAwad, "Impulse noise reduction in audio signal through multi-stage technique", *Engineering* Science andTechnology, an International JournalVol. 22., pp. 629-636., 2019.
[11] Yingwei Fu1, Kele Xu1, Haibo Mi1, Huaimin Wang, Dezhi Wang and Boqing Zhu1. A Mobile Applicationfor Sound Event Detection. Proceedings of the Twenty-Eighth International Joint Conference on ArtificialIntelligence IJCAI-19, 2019.
[12] Eduardo Fonseca, ManojPlakal, Frederic Font, Daniel P. W. Ellis and Xavier Serra, "Audio Tagging withNoisy Labels and Minimal Supervision", Detection and Classification of Acoustic Scenes and Events, Vol.25., 2019.
[13] DeepankVerma, Arnab Jana, and KrithiRamamritham, "Classification and mapping of sound sources in local urban streets through AudioSet data and Bayesian optimized Neural Networks", *Research Article, NoiseMapp*, Vol.6, pp. 52-71, 2019.
[14] I. Kauppinen, "Methods for detecting impulsive noise in speech and audio signals", *In Proceedings International Conference on Digital Signal Processing*, 2002.
[15] Arslan, Yuksel, "A New Approach to Real Time Impulsive Sound Detection for Surveillance Applications",*arXiv:1906.06586*, 2019.
[16] A. Dufaux, L. Besacier, M. Ansorge and F. Pellandini. Automatic sound detection and recognition for noise environment. In Proceedings of 10thEuropean Signal Processing Conference, Tampere, pp. 1-4, 2000.
[17] Fayçal Ykhlef, Sarah Ahmed Hamada and Djamel Bouchaffra.Real-Time Detection of Impulsive Sounds for Audio Surveillance Systems. *JERI*, 2019.
[18] J. Balaji, Ram, D. S. Harish, and Dr. Binoy B. Nair. A deep learning approach to electric energy consumption modelling.Journal of Intelligent and Fuzzy Systems, vol. 36, pp. 4049-4055, 2019.
[19] R.Suresh and Prakash, P.Deep learning based image classification on amazon web service. Journal of Advanced Research in Dynamical and Control Systems, vol. 10, pp. 1000-1003, 2018.
[20] R. Vinayakumar, Dr. Soman K. P., and Poornachandran, P..Detecting malicious domain names using deep learning approaches at scale.Journal of Intelligent and Fuzzy Systems, vol. 34, pp. 1355-1367, 2018.
[21] Tadi Aravind Sasidhar , Bhimavarapu Reddy, Sai Avinash and Jeyakumar G.A Comparative Study on Machine Learning Algorithms for Predicting the Placement Information of Under Graduate Students. InProceedings of 3rd International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), 2019.
[22] Petitti DB, Crooks VC, Buckwalter JG, Chiu V. Blood pressure levels before dementia. Arch Neurol. 2005 Jan;62(1):112-6.
[23] Rice AS, Farquhar-Smith WP, Bridges D, Brooks JW. Canabinoids and pain. In: Dostorovsky JO, Carr DB, Koltzenburg M, editors. Proceedings of the 10th World Congress on Pain; 2002 Aug 17-22; San Diego, CA. Seattle (WA): IASP Press; c2003. p. 437-68.
[24] V.D.Ambeth Kumar (2016).Gender Identification Using a Speech Dependent Voice Frequency Estimation System.International Journal of Computer Science and Information Security,, Vol 14, No: 11, PP: 1025-1034, 2016.