Unsupervised Neural Network approach for the Identification of Anomaly in Speech Signal from Spectrogram Images

M. Joshwin Darington, V. Sathiesh Kumar.
Department of Electronics Engineering, M.I.T campus, Anna University, Chennai-44, Tamilnadu, India.

E-mail: m.joshwin@gmail.com.

Abstract. In this paper, anomaly identification in speech signal is carried out using an unsupervised Neural Network approach. Input audio signal is divided into small segments of information. Initial segment of the samples are used in the training process. The later portion of the sample segments are used in the test process. The features (Spectral Roll-off, Spectral Centroid, Mel Frequency Cepstral Coefficient (MFCC), Pitch (PH) and Energy Density (ED)) are extracted from the input data. The extracted 1D features are converted into spectrogram images. Then, the images are fed as an input to Neural Network for the prediction of feature values. If three or more feature value does not exceed the threshold value of 75% then the input signal is considered to be free from anomaly. The proposed model results in an accuracy of 97.50% in the detection of anomaly in input speech signal.

1. Introduction
In recent years, numerous activities on the identification of anomaly is carried out by different research groups across the world. It is applied in the fault identification process of machineries, system health monitoring, disturbance identification in medical records, fraud detection etc. [1]. The supervised and unsupervised learning methods are used to find the anomalies in the system. Implementing a supervised learning method by utilizing a standard dataset, might fail in some real-time circumstances due to the presence of physical disturbances in the machines. [2]. It results in a faulty system. In order to minimize this problem, we adopted the principle of dynamic dataset generation and anomaly identification using Neural Networks. Any sound signal can be represented in spectrogram format by using a spectrum analyser or by some computational tools [2-3]. The operation of the machinery is analysed with the aid of spectrogram images [4]. Any deviation from the pattern of normal spectral images is identified and termed as anomaly in the input data. In this paper, an unsupervised learning method is used to identify the anomalies in the speech signal. The dynamic dataset is collected in a self-generated manner. The features (five features namely Spectral centroid, Mel frequency cepstral coefficient, pitch, Spectral energy density and Spectral roll-off) are obtained from the real-time speech signal. These features are used in the process of identifying the anomalies. The normal sound (without anomalies) in the training process, to set the flags associated with each feature. The scoring technique is used to eliminate the false positive rate. To make the technique to be robust in nature, the maximum passing rate of the input data is also considered in addition to the threshold-based analysis.
2. Conventional method

Existing anomaly detection systems utilize an auto encoder, numerical based and single format of datasets. The conventional method focuses on the detection of outliners, where the true positive rate and the false positive rate are used to find the anomalies. On the other hand, the supervised method has few important criteria to be followed like fundamental frequency, pitches, mean period, mean autocorrelation, etc. Identification of anomalies from the standard datasets may fail in certain conditions like abrupt change in volume, speakers’ tone, white noise addition, etc. The cardiac disorder is identified from the patient's heart sound signal by utilizing the approximated Wavelet mean method [5]. It is used to build a classifying system to determine the anomalies. This method is tested on standard dataset. Real time testing of this method is not reported by the authors. Another interesting approach in detecting the anomaly is by outliner-detection method based on numerical method (setting up the threshold). The main drawback of this method is in the usage of threshold value for the identification of anomalies. Fixing a threshold value may result in an inappropriate prediction if there is an abrupt change in signal strength [6-7]. The pitch of a person may vary depending on the situation. It results in the abrupt change in signal strength [8].

2.1. Sample preparation:

The audio sample with a time limit of five second is read by the system. It is then sampled at a user defined rate. The audio file has 10948 number of sample points. It is represented in the numerical data (array) as shown in Figure 1. Tempo information (beats per minute) in the audio signal is used to identify the presence or absence (null data) of signal. It is also used in the early partitioning of the signal. The information on the sample audio file is represented in Table 1. Analog signal is sampled in a desired rate to attain the array of data which in turn given as an input to the anomaly identification system. Figure 2 shows the plot on input audio signal, generated White noise and noise added input signal. The equations relating to the generation of White Noise is given below.

\[ S_X(f) = \frac{N_0}{2}, \text{ for all } f. \] (1)

\[ E[X(t)^2] = \int_{-\infty}^{\infty} S_x(f) \, df. \] (2)

f: Frequency range.

\( S_X(f) \): White noise.

\( N_0 / 2 \): Spectral densities.

\( E[X(t)^2] \): Expected power. (3)

![Figure 1. Representation of an audio sample.](image)
Table 1. Specifications of the sample audio.

| Specification       | Value                                      |
|---------------------|--------------------------------------------|
| Tempo               | 215.33203125 BPM                           |
| Range               | -0.75 to 0.99331045 V                      |
| Peak amplitude      | 0.99331045 V                               |
| No of frames        | 10948 (No of counts)                       |

Figure 2. White noise added with the sound for testing.

3. Methodology

Figure 3 shows the model block diagram representation of anomaly detection in speech signals. The model consists of two important sections, namely, pre-processing block and Neural Network block. The pre-processing block performs the sample splitting process, feature extraction and data-split process. The Neural Network section performs the actual learning process on input data and also identifies the anomaly in the speech signal.
Algorithm will take a thirty-minute audio file and segment them into n-number of samples. Each sample will have about 20000 sample points (equivalent to sample signal of 5 seconds). The system performs two distinct operations which is coined as post and pre-processing the temporal data which is in one dimensional format as shown in Figure.4.

Figure. 3. Block diagram representing the anomaly detection in speech signal.

Figure. 4. Flowchart representing the combined work of training and testing process on data samples.
The temporal data is converted into spectrum images. The initial samples (five seconds) in the input signal is used in the training process. The later samples (after five seconds of information) in the input audio file is used for the testing process. From the spectral image, the features are extracted. The features extracted are Spectral Roll-off (SRO), Spectral Centroid (SC), Mel Frequency Cepstral Coefficient (MFFC), Pitch (PH) and Energy Density (ED) as shown in figure 5 and 6. The above said features are used in the training and test processes. The feature split ratio for the training and test process is listed in Table 2.

![Figure 5](image1.png)  
**Figure 5.** Extracted features from the audio signal. a) Pitch, b) Spectral roll-off and c) Spectral Energy density.

![Figure 6](image2.png)  
**Figure 6.** Extracted features from the audio signal. a) MFFC and b) Spectral centroid.

| FEATURES         | COUNT                             |
|------------------|-----------------------------------|
| Spectral centroid| 500(100 test and 400 train)       |
| Spectral bandwidth| 500(100 test and 400 train)       |
| MFFC             | 500(100 test and 400 train)       |
| Spectral Roll-off| 500(100 test and 400 train)       |
| Pitches          | 500(100 test and 400 train)       |

The spectrum is a graphical representation of the frequency content in a sound or an audio signal. Spectral centroid is the centre of mass for an audio located at certain time period. The Spectral
Roll-off denotes the total spectral power percentage in the audio signal. Mel Frequency Cepstral Coefficient (MFFC) describes the overall shape of a spectral envelope and pitch are the scientific unit of sound wave vibration. The equations related to the extracted features are represented in Algorithm-1.

**Algorithm - 1:**

**Input:** D:\sample.wav:
- \(A = \text{len (D:\sample.wav)}\);
- for i in range (0: A/2:100):
  - \(Y = \text{load (D:/sample.wav)}\);
  - \(\text{Write.Y(samples+i)}\);

\(S\) indicates the (time\((x)\), amplitude\((y)\));

**Generate S \forall Samples:**

1. Extract \((S)\): \(\text{Features:1 } \rightarrow \text{spectral centroid: } C_t = \frac{\sum_{n=1}^{\infty} M_t[n] \cdot n}{\sum_{n=1}^{\infty} M_t[n]}\).
2. \(\rightarrow \text{spectral roll off: } \sum_{n=1}^{\infty} M_t[n] = 0.85 \cdot \sum_{n=1}^{\infty} M_t[n]\).
3. \(\rightarrow \text{spectral MFFC: } \text{Mel}(f) = 2595 \times \log_{10}(1+f/700)\).
4. \(\rightarrow \text{spectral Pitches: } f_0 = \sum (1 / T)\).
5. \(\rightarrow \text{energy density: } \int_{-\infty}^{\infty} x(t)^2 \, dt\).

**Features Write file \(\rightarrow \text{r"D:/spectrums/train"}\).**

**Training** (split data(train/test)):
- Configure: [Hidden layer 1-neurons = 128(relu), Hidden layer 2, neurons = 128(relu),
- SoftMax, output nodes = 5];
- Run training \(\rightarrow\) Epoch=150, batch size=32;
- Print(accuracy);
- Model.save(\(\text{r"D:/model/anomaly")}\)

**Testing an audio file:**

**Input \(\rightarrow\) (D:\test.wav):**

- \(Y = \text{load (D:\test.wav)}\);
- While(Y \(\rightarrow\) Feature):
  - Allocate (P1: SRO; P2: SC; P3: MFFC; P4: PH; P5: ED);
  - \(A = P1+P2+P3+P4+P5\);
  - if(A>3):
    - Print ("anomalies not found");
  - Else
    - Print ("anomalies found");
  - end while
- Update \(\rightarrow\) Data frame.

The extracted features values are converted into a 1D vector using the flattening layer and fed as an input to the neurons of first hidden layer of Neural Network. The number of neurons in the first hidden layer is of 128. Rectified Linear Unit (RELU) activation function is used in this layer. The output from the first hidden layer neurons are sent to 128 neurons of second hidden layer. This layer also utilizes the RELU activation function. The output from the second hidden layer neurons are sent to five output layer neurons. The five output layer neurons estimate the probability score on the feature’s values (SRO, SC, MFFC, PH, ED).
The number of probabilities score above 75% from the five output layer neurons are counted. If the count value is above or equal to 3, then the input signal is said to be free from anomalies. If the count is less than or equal to 2, then the input signal is mean to have anomalies.

The algorithm is implemented using an Intel(R) Core (TM) i-7-8550U, 1.80GHz CPU with 8192MB RAM. Programming is carried out using the Python environment with supporting packages such as pydub and sound card libraries. To support Neural Network implementation librosa package with TensorFlow as backend is used.

4. Results and discussion

Figures 7(a) and 7(b) shows the self-generated audio signal (without and with anomalies). It is used in the training and test process of the model. White noise added input signal is considered as an anomaly appended signal. Total number of samples used in the training and test process are 400 and 100. From the input signal, five features (Spectral Roll-off (SRO), Spectral Centroid (SC), Mel Frequency Cepstral Coefficient (MFFC), Pitch (PH) and Energy Density (ED)) are extracted and utilized in the training cum test process. The extracted feature spectrum with and without anomaly is shown in Figures 7(c) to (l).

The extracted features from the input signal without anomaly is used in the training process of the model. During the test process, the input signal with and without anomaly is fed in to the trained
model. The extracted feature spectrum from the test samples are sent to the Neural Network for the prediction of feature values (SRO, SC, MFFC, PH, ED). A threshold value of 75% per feature is selected to determine whether the sample or input data contains anomaly or not. If the three or more number of feature values with a probability score of >75%, then the input data is considered to be free from anomalies. If two or less number of features values obtain a probability score of >75%, then the input signal is affected by the anomalies. The predicted output for samples with and without anomalies are shown in Figure. 8. The loss and accuracy plot for the proposed model is shown in Figure. 9.

![Graphical representation of Output](image1)

**Figure. 8.** Graphical representation of Output:
a) input signal without anomalies, b) input signal with anomalies – sample1 and c) input signal with anomalies - sample 2.

![Graphical representation of Loss and Accuracy](image2)

**Figure. 9.** Graphical representation of (a) loss and (b)accuracy.

**Table 3.** Shows the number of samples tested during the test process.

| Prediction rate          | Number of samples | Number of right predictions |
|-------------------------|-------------------|----------------------------|
| Samples without anomaly | 48                | 45                         |
| Samples with anomaly    | 30                | 30                         |
| Total                   | 78                | 75(96.15 %)                |

From the table, it is observed that the proposed model efficiently identifies the anomaly in the input signal. The overall prediction accuracy of the model is about 96.15%. In this paper, white noise is added with the input signal is considered as an anomaly.
As a future work, many other noise sources much related to real time signals will be considered as anomaly. Also, the number of features used in the training cum test process will also be increased to improve the prediction rate of anomalies in the input signal.

5. Conclusion
Anomaly detection in speech signal is carried out using an unsupervised Neural Network approach. The initial portion of the audio signal is used in the training process. The later sequences are considered for test process. 1D features extracted from the input signal is converted into spectrogram images. Later, it is trained by the model. The test spectrogram images are classified using an unsupervised Neural Network. The model results in the probability score for individual feature values. The number of feature values exceeding the threshold (75%) is counted. If the number of count value is greater than or equal to 3, then the input signal is free from anomalies. Otherwise, the anomaly has been identified in the input signal. By using the proposed model an overall prediction accuracy of 96.15% is achieved in the context of anomaly detection.

References
[1] Kaori Suefusa, Tomoya Nishida, Harsh Purohit, Ryo Tanabe, Takashi Endo, and Yohei Kawaguchi, Higashi-koigakubo, Kokubunji-shi, Tokyo, Japan. Anomalous sound detection based on interpolation deep neural network, research and development group, hitachi, ltd. IEEE conference on Acoustic, Speech, and signal processing, pp. 185-8601, 2017.
[2] Yuma Koizumi, Shoichiro Saito, Hisashi Uematsu, and Noboru Harada, NTT corporation, tokyo, Japan. Optimizing acoustic feature extractor for anomalous sound detection based on neyman-pearson lemma, ntt media intelligence laboratories, pp. 1-28 in 2019.
[3] Kenji Ichige, and Koichi Hamada Higashi-koigakubo. China. Anomaly detection based on an ensemble of dereverberation and anomalous sound extraction. research and development group, hitachi, ltd,vol no 1, pp. 102-280, 2018.
[4] Mattia Antonini, Massimo Vecchio, Fabio Antonelli, Pietro Ducange, and Charith Perera, Italy. Smart audio sensors in the internet of things edge for anomaly detection. ieee research centre, ecampus university, via isimbardi , pp.60-220, 2017.
[5] Mohan Mishra, Anushikha Singh, Malay Kishore Dutta, Department of electronics and communication engineering, amity university, noida, India. Automatic screening of cardiac disorders using wavelet analysis of heart sound., 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON), pp.320, 2017.
[6] Yuma Koizumi, Shoichiro Saito, Hisashi Uematsu, Yuta Kawachi, and Noboru Harada, senior member, ieee. Japan.Unsupervised detection of anomalous sound based on deep learning and the neyman–pearson lemma, in proc. ieee,2019, pp.253-270-acoustic signal processing.
[7] A. Yamashita, T. Hara, and T. Kaneko, Japan. Inspection of visible and invisible features of objects with image and sound signal processing, in proc. ieee/rsj int. conf. intellig. robots syst., 2006, pp. 3837–3842.
[8] Y. Koizumi, S. Saito, H. Uematsu, and N. Harada, Japan. Feature extractor for anomalous sound detection based on neyman–pearson lemma. in proc. 25th eur. signal process. conf., 2017, pp. 698–optimizing acoustic.