Audio Classification using Machine Learning

Shivanshi*
Saakshi Joshi**
Richa Singh***

ABSTRACT

The present work has been carried out on audio classification for outdoor events by making use of artificial neural network (ANN) The main aim of using ANN is to figure out a way of mapping the input with the corresponding output class present in the dataset. It is quite interesting to explore about how much machines are capable to recognize the audio and to determine what type of sound it is. In this research authors use audio clips taken from UrbanSound8K dataset are processed and converted to spectrograms that are further fed to the ANN classifier which identifies the type of sound based on the frequency data captured in the spectrogram.

Keywords: *Exploratory Data Analysis, Mel Frequency Cepstral Coefficient, Artificial Neural Network*

1. Introduction

The process of analysing audio recordings is referred to as audio classification. [1]Audio classification has numerous applications in the field of artificial intelligence (AI) and data science, like chatbots, gender identification, speech recognition and virtual assistants. [2]There are four different types of audio classification such as NLC (Natural Language Utterance Classification), MC (Music Classification), ADC (Acoustic Data Classification) and ESC (Environmental Sound Classification). In Machine learning, audio classification is the most widely explorable field. It provides the knowledge of predicting different sounds of the outside world. [3]Machine learning is an important technology for today's world and it is growing very fast day by day. Everyone in their daily life using machine learning even without realising it such as google assistant, Alexa, google maps etc. Machine learning algorithms have many applications such as traffic prediction, online fraud detection, medical diagnosis and many more. [4] The author understands the application of audio classification to the surrounding and its creatures. To go further in this study, it's important to understand that why audio classification is needed?

Investigation of vast field of audio classification research is done because there is much to learn. The authors studied that [5]in this world, around 285 million which is approximately equal to the 20 % of the Indian Population are impaired person. These people suffer from regular navigation problems especially when they are on their own. Such people depend on others for their daily needs. So, it is quite challenging task and much needed field to generate some equipment using trending technologies like machine learning. Thus, the authors try to classify audio of different types which helps in the development of such instruments that can detect objects by their sound and aware an impaired person from getting any harm.

2. Related Work

After survey of literature, it was found that maximum of the research work has used the

*School of Computing, DIT University, Dehradun, Uttarakhand, 248009

technology WER (word error rate) for determining the efficiency of their systems. [6] It is very interesting to know that many researchers have use the MFCC as a feature extraction for audio signals. MFCC's were heavily used in classical classification i.e. HMM and GMM. [22]75% of DNN models were standalone models where only 25% of the model uses the hybrid models. Researchers are more encourage to use hybrid model as research shows HMM or GMM inform of a DNN gives better results. Many researchers proposed audio classification system with the help of ANN. [25] Voice Activity Detection(VAD) was also for separation purpose ANN was used for modelling each individual utterance. MFCC is calculated to characterize audio content. Experimental results have shown that the proposed audio using ANN learning method was good and has 92% of speech recognized rate.

Support vector machines (SVM) and neural networks (NN) had performed better on audio dataset Parameter tuning has given 98.6% accuracy with SVM and 99.87 % with ANN. Hence many researchers have concluded ANN as best way for audio classification. It was also observed that Modified autocorrelation has the accuracy of 95 % and AMDF has accuracy of 85% and the proposed system is the combination of these two called combo classifier that has an accuracy of 99%. [24]Also the proposed algorithm has achieved accuracy rate of approximately 90% and LPC has achieved accuracy rate of 66.66%. So it can be concluded that proposed algorithm has improved accuracy as compared to LPC. After comparative studies it was concluded that Hidden Markov method (HMM) is best suitable as it is efficient, robust and reduces time and complexity.

[23]The TIMIT acoustic-phonetic continuous speech corpus dataset has been used to assess the word error rate performance of three voice recognition systems, namely GMM-HMM, DNN-HMM, and DBN. The DBN-based audio recognition system outperforms the other two audio recognition systems, according to the results. [8] For speech recognition systems based on GMM HMM, the word mistake rate is roughly 35%. [24] A speech recognition system using Deep Neural Network Hidden Markov Models (DNN-HMMs) with five hidden layers has a word error rate of roughly 32%. With three hidden layers and 2048 hidden units per layer, a DBN voice recognition system layer has a word error rate of roughly 24%. 98% of phonemes are accurately recognised by a probabilistic neural network and recurrent neural network combined, with HMM coming in second.

3. Methodology

Presently, the author studied and presented research on Machine Learning which is the prime technology used for audio classification. Audio classification has its wide application in various industries across different domains like voice lock features, music genre identification, Natural Language classification, Environment sound classification, and to capture and identify different types of sound. Thus, this study aims toward the audio classification of the 10 classes containing different audio files (i.e., dog bark, street sound, gunshot,) via advanced research methods using the UrbanSound8k dataset where the EDA and pre-processing is performed by MFCC architecture and model creation and its implementation is done by creating an ANN model with 3 dense layers using keras sequential API.

This work has been carried out in three distinct phases: -

1. Exploratory data analysis (EDA)
2. Data preparation
3. Development and application of models.

The workflow of the work is presented in Fig. 1.

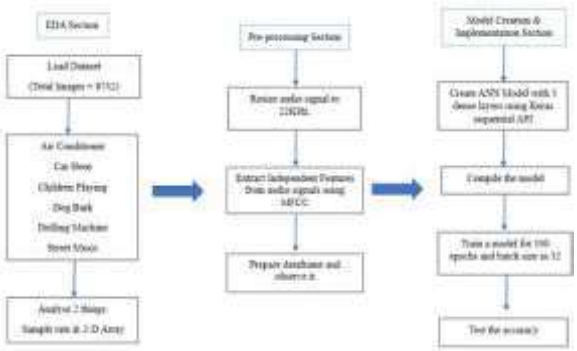

Figure 1. Work Flow Chart

3.1 Data Set Used

The dataset that author has used in this paper is referred to as 'UrbanSound8k' dataset. The dataset has 10 different classes in which there are 8732 sound files. All the classes are shown below. [9]The task is to extract various features from the files listed below and categorise them with their associated audio. Download the dataset from the official website www.freesound.org. [10]This dataset is also available on Kaggle. The different types of sounds available in the dataset are:

1. Car Hooter
2. Kids having fun
3. Drilling Equipment
4. Idling of Engine
5. Gunshot
6. Jackhammer
7. Siren
8. Urban Music
9. Dog Bark
10. Traffic Horn

3.2 Exploratory Data Analysis (EDA)

In the EDA part, load that audio data which has been taken from UrbanSound8K using Librosa library and give the audio file location directly to generate the waveforms. While loading the librosa audio file it gives the knowledge of two things first is the two-dimensional array and the other is the sample rate. Then the next part is to load the different audio files like dog bark, traffic horn, gunshot etc., and generate its waveforms. The generated sound waves are presented in Fig. 2 and Fig. 3.

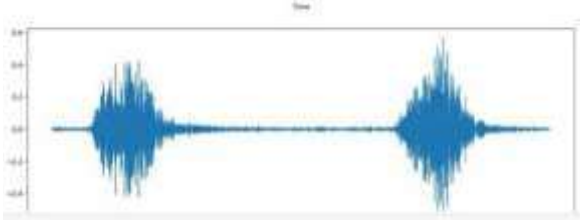

Figure 2. 1st Sound wave generated

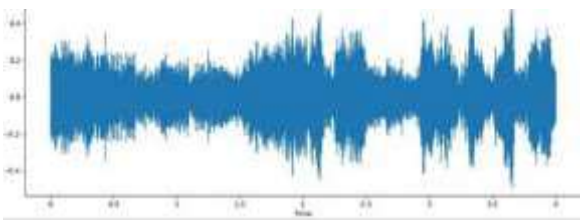

Figure 3. 2nd Sound wave generated

Wave Sample rate and audio data will be recorded as 22050 and 44100.

1-Sample rate – [12]It generally tells how many samles are recorded in a second. The default sample rate at which librosa library reads the file is 22050. It depends on the library selection. The sample rate depends on the library chosen.

2-2D Array – [12]The first axis represents the amplitude of the recorded samples while the second axis determine the number of channels. There are basically two types of channel first one is Monophonic that has only one channel in it. Second one is stereo in this audio there is two channels.

When the sample rate is printed, using scripy library then it would be different from librosa library. The difference between scripy and librosa library is that when the author tries to read the data using librosa then it is normalized but if it reads by scripy it cannot be normalized. Nowdays, librosa library is widely used because of following reasons.

1) It converges the signal in one signal.
2) It displays an audio signal in the normalize form that is -1 and +1 so that observe the regular pattern of audio signal can be observed.
3) By using librosa, the sample rate can be seen and converting sample rate to 22KHZ automatically while other libraries are observed, different values.
4) It is relatively easy to use and can be done by using variety of tools and packages.
5) It is also used to check the data quality issues.

3.3 Data Preprocessing

Some audio files are being recorded at different rates, such as 44 or 22 kHz. [11]That signal is brought up to 22 KHz using the Librosa package, and data can be viewed in a normalised pattern. The data is in the form of independent (extracted features from the audio signal) and dependent (class labels) characteristics. The main objective is to extract some key information from the data that we have loaded in EDA. Mel Frequency Cepstral Frequency (MFCC) is used to extract unique features from the audio sources.

Mel Frequency Cepstral Coefficients, or MFCC-[14]Cepstral was primarily created in the 1960s while researchers were looking at seismic signal echoes. The frequency distribution throughout the window size is summarised by the MFCC. Therefore, it is feasible to study the sound's frequency and time properties. Characteristics can be classified using this audio representation. [21]As a result, it will attempt to transform audio into features based on temporal and frequency characteristics that will aid in classification.

First, in order to show how actually MFCC is used, apply just one audio file to it. Then take features from each audio file in order to create the dataframe. [13]Consequently, developing a function that accepts the filename (file where it is found on the journey). Now, loading the file using the librosa library and obtain 2 pieces of information from it. To determine scaled features, first determine the MFCC for the audio data and then determine the mean of the transpose of an array.

To extract all the features for each audio file, loop through each row of the dataframe. [15]To track the development, additionally use the TQDM Python package. The method to extract MFCC features will be called inside the loop, and then establish a unique file path for each file before calling it to append features and accompanying labels to a freshly created dataframe. Due to its iterative nature, the loop will take some time to complete after which you may view the dataframe of extracted characteristics, which has about 8000 rows.

First, divide the dependent and independent characteristics for the train test. [20]Then, utilise label encoding (Integer label encoding) from numbers 1 to 10 and then transform it into categories to convert all 10 classes. The data was then divided in half, 80/20, into train and test sets. The audio signals represented in one-channel and two-channel format are shown in Fig. 4 and Fig. 5.

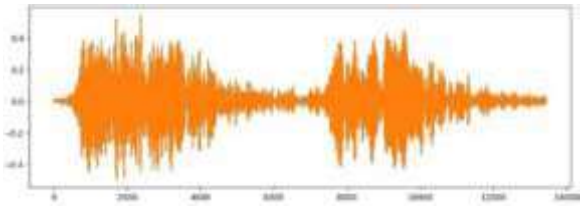

Figure 4. One Channel audio signal

3.4 Model Creation

To create a model, TensorFlow and ANN model are used by taking care of the version of TensorFlow to be greater than 2.0. To create an ANN model the libraries like sequential, Dense, Dropout, Activation, Flatten have been used.

The number of classes should be known for ANN model creation. Create ANN with three dense layers. The first dense layer will have 100 neurons and the input shape will be 40. [16]Take the input shape as 40 because there are 40 features in the training dataset so give the input same with respect to the number of features. An activation function called Relu is applied with dropout at a rate of 0.5. Using dropout library so as to avoid any overfitting in the model.

The second layer of the model, this model is going to have 200 neurons with rest of the features being same as used in the first layer. Our third layer will contain 100 neurons with rest of the features being similar to the first layer. Finally comes output layer which has 'softmax' as the activation function because it is a multiclass classification problem. The proposed model architecture is shown in Fig. 6.

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 100)	4100
activation (Activation)	(None, 100)	0
dropout (Dropout)	(None, 100)	0
dense_1 (Dense)	(None, 200)	20200
activation_1 (Activation)	(None, 200)	0
dropout_1 (Dropout)	(None, 200)	0
dense_2 (Dense)	(None, 100)	20100
activation_2 (Activation)	(None, 100)	0
dropout_2 (Dropout)	(None, 100)	0
dense_3 (Dense)	(None, 10)	1010
activation_3 (Activation)	(None, 10)	0

Figure 6. Model Summary

4. Result

After optimizing the proposed model, the final training accuracy achieved was 83.5 % and testing accuracy of 82.0 % was reached. The performance is moderate and it can be used to predict the different audio data into different predefined categories of classes.

For checking the test accuracy choose a sound file from the dataset 'UrbanSound8k'. The first information that needs to be executed is the extract features. Pre-process the new audio data wherein the extraction of the features is required from that specific data itself again with the help of MFCC. [18]Prediction of the class labels is done with the help of model that has been created previously. Inverse transform the label in order to get the class name.

Select any random audio file from dataset in order to test the model. The three steps are performed.

1. Loading the audio file with the help of librosa library and extracting the MFCC features.
2. Predicting the labels to which audio belongs.
3. To get the class name to which the audio file belongs, an inverse transformation of predicted label is done

[17]There are basically two changes in the information that is performed here. The first change is that initially the author was trying to use get dummies to do the label encoding with respect to the output feature but here one point needs to be understood that whenever one wants to work with the test data, inverse transform is needed from that label into the class name. This is the reason why Label encoder is used here instead of get dummies function. [19]Label Encoder basically converts the data into categorical form of numerals.

At last, the author successfully completed the audio classification of different sounds using machine learning with a good training accuracy and test accuracy.

5. Conclusion

In the present work, the authors experimented with the audio data classification and found that ANN classifier is very useful for this task. Using MFCC for extracting characteristics of the audio files and this provide learning and experimenting more on pre-processing and model creation. In order to categorise the audio in a different class, author built a straightforward ANN model on top of it.

By the end of this study a successful model is created and tested, where overall test accuracy of 82.0 % has been achieved. Any loss in the given model is observed to be the losses obtained due to an error in the data available. The current error rate is a bit high for practical use. To improve the output, there is the need of some better features that can explain the time dependencies in signals in a better or clear way.

REFERENCES

Soontorn Oraintara, Ying-Jui Chen Et.al. IEEE transactions on signal processing, IFFT, Vol. 50, No. 3, March 2002

Kelly Wong, Journal of Undergraduate Research, Florida, Vol 2, Issue 11 - August 2011

M.A. Anusuya and S.K. Katti. Speech recognition by machine: A review: International Journal of Computer Science and Information Security, Vol. 6, No. 3, 2009.

Sadaoki Furui, 50 years of progress in speech and Speaker Recognition Research, ECTI transactions on Computer and Information Technology, Vol.1. No.2 November 2005.

D.R. Reddy, an approach to computer speech recognition by direct analysis of the Speech Wave, Tech. Report No.C549, Computer ScienceDept., Stanford Univ., September 1966.

Simon Kinga and Joe Frankel, recognition, speech production knowledge in automatic speech recognition, Journal of Acoustic Society of America, 2006.

Nathaniel Morgan and Herve Bourlard. Continuous speech recognition using multilayer perceptrons with Hidden Markov models. In IEEE International Conference on acoustics, speech and signal processing, 1990.

Abdel-Rahman Mohamed, George E. Dahl, and Geoffrey E. Hinton. Deep belief networks for phone recognition. In neural information processing systems: Workshop on deep learning for speech recognition and related applications, 2009.

Jen-Tzung Chien, linear regression base bayesia predictive classification for speech recognition, IEEE transactions on speech and audio processing, Vol. 11 No. 1, January 2003.

Mohamed Afify and Olivier Siohan, sequential estimation with optimal forgetting for robust speech recognition, IEEE transactions on speech and audio processing, Vol. 12, No. 1, January 2004.

Shinji Watanabe, variational bayesian estimation and clustering for speech recognition, IEEE transactions on speech and audio processing, Vol. 12, No. 4, July 2004.

Mohamed Afify, Feng Liu, Hui Jiang, A new verification-based fast-match for large vocabulary continuous speech recognition, IEEE transactions on speech and audio processing, Vol. 13, No. 4, July 2005.

Giuseppe Riccardi, active learning: Theory and applications to automatic speech recognition, IEEE transactions on speech and audio processing, Vol. 13, No. 4, July 2005.

Frank Wessel and Hermann Ney , unsupervised training of acoustic models for large vocabulary continuous speech recognition, IEEE transactions on speech and audio processing, Vol. 13, No. 1, January 2005 .

Mathias De-Wachter et.al., Template based continuous speech recognition, IEEE transactions on Audio, speech and Language processing, Vol.15, No.4, May 2007.

J.W.Forgie and C.D.Forgie, results obtained from a vowel recognition computer program, J.A.S.A., 31(11),pp.1480-1489.1959.

V.M. Velichko and N.G. Zagoruyko, automatic recognition of 200 words, Int.J.Man-Machine Studies, 2:223, June 1970.

Michael Unser, Thierry Blu, IEEE Transactions on Signal Processing, Wavelet Theory Demystified, Vol. 51, No. 2, Feb'13

Soundararajan et.al., Transforming Binary uncertainties for Robust Speech Recognition, IEEE Transactions on Audio, Speech and Language processing, Vol.15,No.6, July 2007.

Ram Singh, proceedings of the NCC, spectral subtraction speech enhancement with RASTA filtering IIT-B 2012.

Nitin Sawhney, situational awareness from environmental sounds, SIG, MIT Media Lab, June 13, 2013.

Chadha, N., Gangwar, R.C., Bedi, R. 2015. current challenges and applications of speech recognition process using natural language processing: A survey, International Journal of Computer Applications (0975-8887), 131(11), 28-31.

Petkar, H. 2016. A review of challenges in "Automatic Speech Recognition", International Journal of Computer Applications (0975-8887), 151(3), 23-29.

Ramírez, J., Górriz, J.M., Segura, J.C. 2007. Voice activity detection, fundamentals and speech recognition systems robustness, University of Granada, Spain.

Hirsch, H.G., Pearce, D. 2000. The aurora experimental framework for the performance evaluation of speech recognition system under noisy conditions, ASR-2000 Automatic speech recognition: Challenges for the new millennium Paris, France, 181-188.