Int. J. Nonlinear Anal. Appl.
Volume 12, Special Issue, Winter and Spring 2021, 2125-2136
ISSN: 2008-6822 (electronic)
http://dx.doi.org/11

MFCC based hybrid fingerprinting method for audio classification through LSTM

K. Banuroopa[a], D. Shanmuga Priyaa[a,*]

[a]*Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore, India*

(Communicated by Madjid Eshaghi Gordji)

Abstract

In this paper, a novel audio finger methodology for audio classification is proposed. The fingerprint of the audio signal is a unique digest to identify the signal. The proposed model uses the audio fingerprinting methodology to create a unique fingerprint of the audio files. The fingerprints are created by extracting an MFCC spectrum and then taking a mean of the spectra and converting the spectrum into a binary image. These images are then fed to the LSTM network to classify the environmental sounds stored in UrbanSound8K dataset and it produces an accuracy of 98.8% of accuracy across all 10 folds of the dataset.

Keywords: Audio fingerprinting, MFCC, Audio Classification, LSTM.

1. Introduction

The progression of internetwork technology over the decades has produced colossal quantities of data in multiple formats like text, image, audio and video, etc. Therefore, the requirement of potential software and applications to handle this data for identification, retrieval of such data has increased. One such task is classifying audio or sound produced as music, speech or other environmental sounds. Recently this is a leading territory in research and the numerous models have used varied features to produce valuable and precise outcome. The reasons behind all these researches are due to the various fields where the application of Environmental Sound Classification (ESC) is very important. Audio classification can be applied in varied applications like multimedia, bioacoustics surveillance, trespassers discovery in wildlife areas to audio forensics and monitoring environmental audio signals.

*Corresponding author
Email addresses: banuroopa@gmail.com (K. Banuroopa), shanmugapriyaait@kahedu.edu.in (D. Shanmuga Priyaa)

Received: November 2021 *Accepted:* December 2021

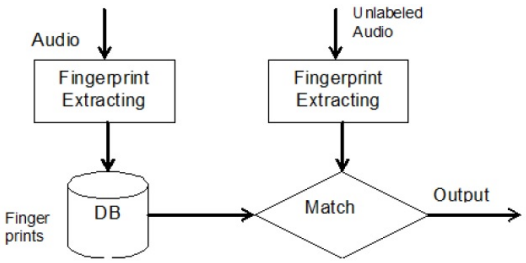

Figure 1: Content–based Audio Identification Framework.

Notwithstanding, this is a difficult issue because of the complex nature of ecological sounds as far as dimensionality, diverse instrument of sound creation mixture of various sources, and absence of significant level structures typically in discourse and in numerous sorts of melodic sounds. These complex natures of sound, when coupled with different natural noises as well as the various forms of sound production mechanism make it harder for classification of the sound to be done effectively.

The aim of the content-based audio analysis approaches is that an audio file can identified and described by the features extracted from its content rather than its meta data. Content –based information retrieval is an alternative approach to the traditional text-based searching process. Query by Humming (QbH) is a content-based information retrieval approach which uses a small piece of audio as a query instead using Meta data such as name, artist, album etc. QbH can be used when Meta data information is not available. The music information retrieval systems the piece of audio used and can be sample of original file, a hummed tune of the music or even whistling the tune. QbH uses complex algorithms to convert the hummed tune to query and search the music database for a match.

Another Content-based audio identification (CBID) system called Audio Fingerprinting is also used for audio file identification in a large database. These approaches calculate a small fingerprint or digest the each and every audio file in the database and use it as an index to identify the audio file searching. This method is used to automatically label audio data when presented with an unlabeled audio file. The audio fingerprint of the query audio file is calculated and matched against the fingerprints of audio files in the database instead of searching the entire file for a match.

2. Audio Fingerprinting

An audio finger print is a small set of features that uniquely identifies an audio file. An audio fingerprint can be used for broadcast monitoring, audience measurement and meta-data collection.

The audio fingerprinting system has two basic processes: (1) calculation of Fingerprint (2) matching fingerprints. [3]. In his paper Pedro Cano depicts a frame work for content-based audio identification using audio fingerprinting.

The first process of fingerprint extraction involves a frontend and fingerprint modeling for the entire database of the audio. The modeling process can use any one of the fingerprinting methods to compute the fingerprints of the audio signals by extracting its features. Then the audio fingerprints are stored in the database. The Fig. 2 gives the sequence of various stages of fingerprint modeling.

2.1. Preprocessing

The first step is to pre-process the audio file. In this step one or more of the following processes are done [10]:

1. Digitalize audio if necessary

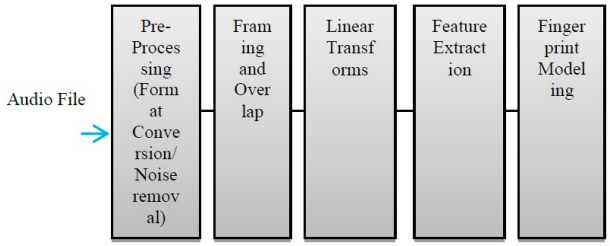

Figure 2: Audio Finger Print Extraction

2. Conversion to a general format
3. Averaging left and right channels
4. Bandpass filtering
5. GSM coder/decoder in a mobile phone
6. Pre-emphasis
7. Normalizing the loudness

2.2. Framing and Overlap

In this step the signal is divided into frames. The frames processed in a second are known as frame rate. A narrowed windowing is done to every frame to diminish the discontinuities at the commencement and conclusion. Frames overlapping are done which make sure the robustness of the frames in shifting.

2.3. Linear Transforms

In the next step of applying linear transforms, the transformation of the set of measurements to a new set of features is done. By choosing the correct transform the redundancy is reduced profoundly. The frequent transformation mostly used is the Fast Fourier Transform (FFT) and the Discrete Cosine Transform (DCT). DCT is said to exhibit shift invariance property [16].

2.4. Feature Extraction

Applications of other transformations are done to generate the final acoustic vectors once the audio file is represented in time-frequency domain. The objective of this is to diminish the dimensionality of the features as well as enhance the invariance to distortions and to extort additional significant parameters. The many methods used for feature extraction are as given below [3]:

- Mel-Frequency Cepstrum Coefficients (MFCC)
- Spectral Flatness
- High-level descriptors
- Pitch
- Bass
- Robust Hash
- Frequency Modulation

2.5. Post Processing

In the post processing step after features are extracted the following processes are done to make the signal a higher order derivative for better calculations [10]. Normalization is done to alleviate hardware implementations, trim down the memory necessities and for expediency in ensuing parts of the system. Decorrelation, differentiation Quantization is done to achieve robustness in opposition to distortions.

3. Review of Literature

The Shazam algorithm has been patented by Wang and Smith in the year 2002. In this method discrete-time Fourier transform is used to create a time-frequency spectrogram. Points with higher intensity than the neighbors called peaks are marked and the rest of the information is discarded. This is used to form a constellation. While matching the fingerprints, sample constellation is matched with database collection of constellations by using pattern matching algorithms. The disadvantage is large number of database entries should be searched [24].

The SHAZAM system has been created for identifying cover audios from other millions of songs stored in the database, with the purpose of using the harmonious assemble of the song. When applied to the Indian film songs database, the system is less effective due to delicate transformation in both rhythm and melody largely due to the semiclassical temperament of Indian film songs. The retrieval accuracy is found to be 85% [21].

The method of Kalker and Haitsma [9] follows the Kurth's [11] approach of a subfingerprint design based on the global energy of each interval and uses a decomposition of the spectrum of each frame into bands using a logarithmic spacing and Hamming distance between the sequences of sub fingerprint of two audio files. However, the corruption of subfingerprints by noise and alterations corrupts the Hamming distances and reduces the amount of information that an indexation algorithm may deduce from such distances.

Philips Robust Hashing (PRH) Algorithm uses Haitsma and Kalker's algorithm. In PRH the input, i.e. the audio signal is first divided into overlapping frames. FFT function is then applied to successfully obtain the power spectrum. The next step is the computation of the energies logarithmically spaced in sub bands. From each frame, subfingerprints, or hash strings are then calculated.

An another method called Waveprint [4] is proposed to be an innovative technique for audio identification. The combination of computer-vision and data stream processing algorithms are utilized by Waveprint to create a small acoustic fingerprint for matching process. The consequential method has admirable recognition potential for diminutive snippets of corrupted audio, as well as opposing noise, pitiable recording quality and cell-phone playback.

An improvement of the above said algorithm is proposed in [8]. This algorithm depends on the lifting wavelet packet rather than Waveprint and it also has an enhanced optimal-basis selection to calculate the coefficient of optimal wavelet packet. It proves that the computing and memory necessity is enhanced which make it more suitable for speech fingerprinting process.

The method proposed in [22] as a spectro-temporal landmarking approach to audio fingerprinting with the addition of a rank-ordering of local maxima that was submitted to MIREX 2015 for the Audio Fingerprinting task. This system is named STELLAR (Spectro-Temporal Landmarking with Rank ordering). This task involves database of ten thousand songs into a database of not more than 2GB and also taking not more than 24 hours. Having created a database, the system should identify approximately 6000 noisy queries within 24 hours.

In [7], the authors propose a new key-dependent audio fingerprinting method which uses discrete wavelet transform and quantization to prove that it performs well in terms of security and storage

concerns than [9]. In contrasting to a good number approaches described above, [1] changed song identification predicament into two dimensional image tasks. The proposed method has created a compact representation of the spectrogram into a 32 bit vectors and a traditional hash table is utilized to carry out the matching process. Based on [1], Baluja et al. extended the wavelet-based technique used in the near-duplicate image retrieval for the sound identification problem. They extracted the spectrogram and compute Haar wavelets for all the spectrogram images. The wavelets with the highest amplitude are taken into consideration as this will have degradation effects in the audio files. For computing the image features from spectrogram, Scale Invariant Feature Transform (SIFT) technique is used by Zhu et al. [25]. The effects of time scale modification and shifts in the pitch are reduced by SIFT features. Even when the audio files are lengthened from 65% to 100% of their original length this system shows potential by good identification results.

Mel frequencies Cepstral Coefficients (MFCC) are condensed version of the spectrogram which is classically used to automatically recognize speech signals and many other audio signal processing applications. In the 1980s, Davis and Mermelstein proposed MFCC and they are being used in much research as a main feature of audio signal. The proposed method uses the MFCC spectra as an input to create a unique fingerprint of the audio signal and uses it for audio classification purpose. The MFCC spectra is similar to a Mel-Spectrogram and applies the same triangular filters which mimic the human ear cochlea. When log is applied over FFT of a signal Mel-Spectrogram results, when a reverse DFT is performed using DCT MFCC is generated. These spectra will have intensity values of the original audio signals in the frequency range of human hearing, which makes it more suitable feature extraction for most audio signal processing applications.

The three stages of classification of audio signals are pre-processing, feature extraction and classification. Pre-processing the audio data partitions the raw audio data into frames. Feature extraction reduces the size of extracted information. Both of these steps are done in the audio fingerprinting system in the proposed method. MFCC is one of mostly used feature for classification of acoustic signals.

There are many AI or non-AI based models for audio classification task. The Environmental sound classification (ESC) task are carried out to automatically recognize the sound in the background of an audio, audio surveillance and audio forensics. Some of the recent classification models for ESC are discussed below.

A deep learning model was proposed by K.J Piczak [14]. It has two convolutional layers, max-pooling layers and two fully connected layers to be tested on unprocessed acoustic files. In order to be consistent with other state- of-the-art, the model exceeds simple implementations based on MFCC which can classify 5.6% more accurately than convolutional methods. Five folds cross validation for ESC-10 and ESC-50 and 10-fold for Urban- Sound8K was done. In the above datasets, the models which are based on a neural network (64.5%) performed better are compared to the respective implementations using manual engineered features (44%), particularly when classifying events in different categories of the ESC-50 datasets.

A deep CNN architecture was modeled by J. Salamon and J.P. Bello [18] for ESC. It has three convolutional layers in the midst of two max-pooling layers and followed by two fully connected layers at the end. They augumented the audio data with the techniques such as time stretching, adding noise, Dynamic Range Compression, pitch shifting to resolve the data scarcity problem and to explore the effects of various improvements on CNN architecture performance throughout with the UrbanSound8k dataset. Data augmentation is considered as a useful technique by enhancing the amount of training datasets without acquiring new data in case of building CNN model. The concept of data augmentation relies on duplicating the existing datasets with variation in order to give the CNN model to learn from more samples so that it can yield an improved accuracy rate. It

has also been shown that increment of class-conditional data can further improve performance. As conclusion of results of the estimated augmentation, the model considerably enhances the effectiveness with accuracy of 79%.

J. Sharma, O. Granmo and M. Goodwin proposed a model consisting of a DCNN along with several feature channels provided for ESC task [20]. Compared to the other models, a deeper CNN (DCNN) made of 2D-separable time and domain convolutions has been employed in this model. There are max pooling layers in the model that separately test the duration and functionality of the domain. To further boost performance, different data augmentation techniques have been used in this model. The technique is pioneering as it uses a assortment of five feature channels. The accuracy of 95.75%, 90.48% and 97.35% are achieved for the datasets ESC-10, ESC-50 and UrbanSound8K respectively.

As LSTM are efficient in time dependency learning I. Lezhenin et.al proposed a model urban sound classification [12]. LSTMs are RNNs that is used to map the sequence of input to the output through using the background information over long intervals. The amplitude of the mel-spectrogram is derived from UrbanSound8k dataset is the feature with which this model is trained. The proposed model is tested through 5 cross validation in contrast to the standard CNN. In this paper, CNN and LSTM models present nearly analogous result with the accuracy of 81.67% and 84.25% respectively.

T. N. Sainath, et al, proposed a model that uses the CNN, LSTM and DNN are combined to form an unique architecture [17]. The complimentary modeling functions of CNNs, LSTMs and DNNs include CNN which are well suited for reducing variability in frequencies, LSTMs which are well known for its temporal modelers and DNNs which are suitable for mapping characteristics to a more separate space. The proposed model in this paper which is referred to as CLDNN has been explored to investigate a number of broad ranges of vocabulary activities from 200 to 2,000 hours. In comparison with LSTM which is the best out of three models, the CLDNN has been able to give a relatively 4-6% improvement in Word Error Rate (WER).

An approach on ESC classification system with sub-spectrum segmentation was presented in [15]. The effectiveness of features is derived from the ambient sounds is heavily dependent upon the ESC output. CRNN and score level fusion are used in combination model to enhance the precision of classification. Comprehensive truncation schemes are tested to determine the optimum number and corresponding sub-spectrogram band set. This proposed model has an accuracy of 81.9% on the ESC-50 data set offering an increase of 9.1% improvement over conventional baseline approaches.

A novel unsupervised generation system, ConvRBM [13] was trained to generate raw audio waves. The model proposes that the mid-frequency filters look like four bases while gamma tones resemble in the low-frequency range. With a CNN classifier in ConvRBM and score-level fusion with Mel filter bank energies, this model has been able to achieve 86.5% on the ESC 50 dataset.

An approach for ESC with Multi-temporal Convolutional Network in amalgamation with multi-level features is studied in [5]. To realize multi-temporal resolution functionality, raw audio files are used and assortments of autonomous CNNs are utilized in the layers with varied sizes and stages. The proposed model has been conducted on two data sets respectively: ESC-50 and DCASE 2017. ReLUs has been used to introduce nonlinear activation features in this proposed model. This proposed multi-temporal network with multi-level features achieved an enhancement of 3.0% on ESC-50 and 2.0% on DCASE 2017 in contrast to single-temporal models.

Another model called EnvNet proposed by Y. Tokozume and T. Harada also used raw audio signals and CNN for detection and classification tasks [23] has realized viable performance. EnvNetv2 which is the second version of the above model uses Before Class (BC) learning model. Two audio signals from various classes are randomly combined in the BC learning process. Then the CNN model provides feedback and is conditioned to generate the mixing ratio. The accurateness of classification

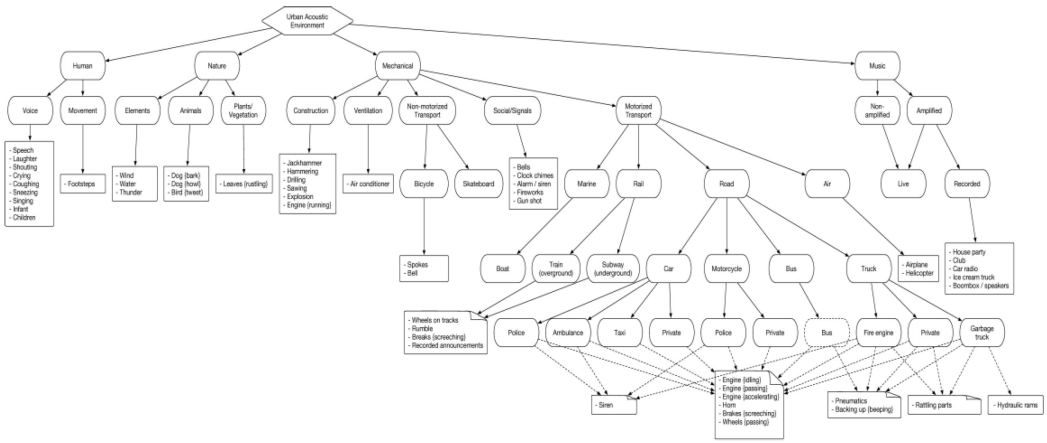

Figure 3: Urban Sound Taxonomy [19]

on the ESC-50 data set is 5.1% more than the static log mel-CNN model.

4. Proposed Method

The method proposed in this paper makes use of the feature that the spectrograms of similar audio will look similar. The changes in the audio file may distort the spectrogram images. To address this issue of mismatch due to noise a set of binary images of the spectrograms are generated by overlapping the frames of the audio. From these set of images at the least one image will make a match. MFCC-spectrograms are generated instead of standard spectrograms as Mel scale triangle filters when applied produce a waveform in time frequency domain where the frequency range matches with the human's audible frequency range.

MFCC-spectrograms are generated for each time frame of 2 second with an overlap of 1 second. And the generated MFCC-spectrograms for the audio file are then used to find out overall mean value and segmenting them into binary image by assigning 1 or 0 depending on the threshold value calculated. The segmented bitmap images are the fingerprints of the audio file. This vector is the audio fingerprint of the given audio file, which will be unique to the audio file. It will contain all the important features of the audio file and thus can be used to uniquely identify that particular audio file with great accuracy and more efficiency than other methods.

The calculated audio fingerprints are stored in a database for reference. If a query audio is presented, the audio fingerprint of the query audio is also generated by the same above said process. Then, it is used to compare the existing audio fingerprints in the database.

4.1. Fingerprints Generation

Fig. 4 illustrates the MFCC fingerprint making method. With the MFCC-spectrogram array of size $w \times h$, which denotes the size of the window from, the average magnitude this array is computed. After that, the magnitude values of this matrix are converted into binary values 0 or 1. When magnitude at given point is higher than the average then it is substituted by the value 1, if not it is a 0. This binary image is the audio fingerprint of the frame. Like this fingerprint are generated for all frames. Creating binary images for all the frames gives a better chance of finding a match between the query audio file and existing audio file with minimum possibility of mismatch.

This method gives robust fingerprints against a range of transformations. Irrelevant noise values in low intensity range in the MFCC-spectrogram can be discarded and only the essential information will be kept in this method because of the usage of the average of the intensity values as threshold

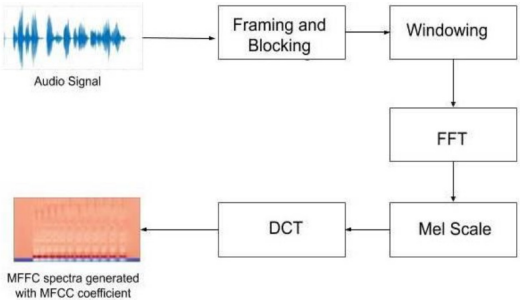

Figure 4: MFCC Spectrum generation

for segmentation. Additionally, the binary bit map creates a fingerprint which is tolerant to change in the relative intensity value.

In spite of changes in the intensity values in the time-frequency plotting in the original signal the binary bitmap has a value 1 denoted there. This property of the binary MFCC-Spectrogram matrix is very much comparable to the Shazam system where the magnitude value is not considered and only highest points are preserved.

4.2. Dataset

One of the main challenges of urban sound research persists in the categorization of audio data into specific classes as there is no urban taxonomy which would help in the labeling and creation of different datasets. In [19], an urban environmental audio categorization is defined which satisfies three elementary rules which are (1) fulfilling formerly defined taxonomies (2) specification of low-level sounds (3) contains of sounds which are known to produce noise pollution in urban cities.

In addition, the UrbanSound8K database is created on the foundation of the above said taxonomy [19]. A total of 27 hours of audio files are in the dataset, which are contained in 8732 tagged audio files with different length and sampling rate. The 10 classes are sound of working of an air conditioner, horn of a car, sound of playing of children, bark of dog, sound of drilling machine, automobile engine idling sound, a gunshot sound, the noise of a jackhammer, sirens issued by police vehicles and music in the street. All the classes contain roughly 1000 audio clips except the car horn and gunshot classes. They have 429 and 374 audio files respectively. Consequently, a few of baseline ML algorithms performed on the dataset in [19] have a reduction of precision for car horn and gunshot classes. Nevertheless, the number of classes is fewer when compared to other datasets as wav files were labelled manually. The eight columns in the csv file of UrbanSound8K are namely; slice file name, fsID, start, end, salience, fold, classID and class names.

- Slice file name - Determines the name of the wav files in the folders.
- FsID - Each wav file is given a unique ID for identification
- Start and end - Labels the start and end time of every occurrence in the 27 hours of audio.
- Salience - Indicating whether the sound was in the background or foreground of the recordings.
- Fold - Specifies the folder to look into for each wav file. In each folder, attributes are selected based on correlation to circumvent overfitting of training data.
- ClassID and class name - Each class is given a name along with an ID from 1-10.

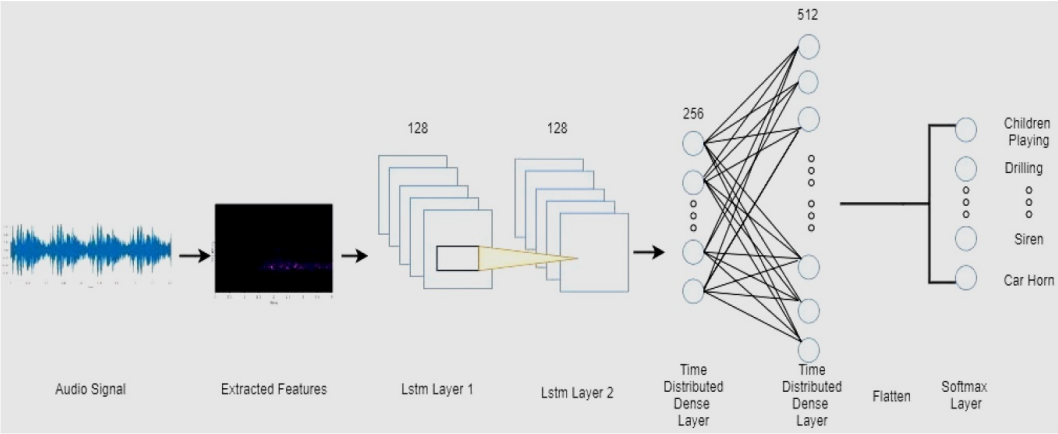

Figure 5: Model Architecture of the Proposed Method

In addition, the 10 folders are made with random assignment of the slices to avoid unnaturally high classification accuracies. In [6] a subsection of the larger dataset is created and named as UrbanSound8K subset which can be also be used for several researches on sound source determination. The subset also contains 10 folds with comparatively less number of wav files in them. Finally, UrbanSound8k was selected to run the model as the dataset was both informative and large.

4.3. Audio Classification Methodology

Classification of audio data through LSTM has developed into popular model lately and it has revealed a remarkable accurateness in the classification task [2]. This model has different layers that are involved in the process of training as well as testing the model. The proposed method for the classification of audio has two LSTM layers, followed by two Time Distributed Layers, a flattening layer and at last a dense layer as made known in the Fig. 5. Furthermore, different activation functions like Rectified Linear Unit (ReLU) and Softmax at different layers have been used. With regards to the model categorical cross entropy as a loss function, the metric accuracy and Adam optimizer have been used as well.

Although seen as a part of RNN, LSTM actually overcomes shortcomings of the RNN itself. LSTM model have blocks of memory which can be referred to as a set of subnets. A memory cell which is also known as cell state and three other gates are present in each block. These gates are: forget gate, input gate and output gate.

In the proposed model stacking two LSTM layers is done. A LSTM layer can be defined by the dimension of the hidden states or cells along with the number of layers that are used. Each LSTM layer has hidden cells, as many as the number of time steps, which is 20 is this case. Each of the 20 hidden cells contain 128 hidden units, and each hidden unit contains some information from the immediate previous hidden cell. An increase in the number hidden units used in each cell affects the training data causing it to overfit. Furthermore, in both the LSTM layers return sequence is put as "true" in order to stack LSTM layers. The two LSTM layers give 3D array as output which will be given as input to the subsequent Time Distributed Dense layer. The input shape of the first LSTM layer is 20x5, in which 20 signifies the timesteps, the number of times LSTM layer repeats itself once it is applied to input. This means each of the input passes 20 times through each of the hidden cells and the 128 hidden units within it. The second layer of LSTM has a dropout of 0.3 in addition to the return sequence which is set to "true". A dropout of 0.3 reduces the effect of overfitting of the training data. The output of the two LSTM layers which is a 3D output and passed on to the time

Layer (type)	Output Shape	Param #
lstm_3 (LSTM)	(None, 20, 128)	68608
lstm_4 (LSTM)	(None, 20, 128)	131584
time_distributed_1 (TimeDist	(None, 20, 256)	33024
time_distributed_2 (TimeDist	(None, 20, 512)	131584
flatten_1 (Flatten)	(None, 10240)	0
dense_3 (Dense)	(None, 10)	102410

Total params: 467,210
Trainable params: 467,210
Non-trainable params: 0

Figure 6: Model summary of LSTM layers with Time Distributed Dense layer

Table 1: Comparison of the proposed model and other state of art models.

Model	**Classification Accuracy(%)**
PiczakCNN [20]	73.70
AlexNet [13]	92.00
GoogleNet [13]	93.00
ADCNN-5 []	97.52
MFCC AF with LSTM (Proposed)	98.80
Average	114.7

distributed dense layers for further processing. The Fig. 6 shows a model summary of LSTM layers with Time Distributed Dense layer.

5. Results

The proposed MFCC based hybrid features audio finger printing method for feature extraction for audio classification through LSTM networks classified the audio files of the urban sound 8K dataset with improved accuracy than the already existing models. The proposed model produced 98.8% accuracy in classifying the ten classes of the data set. The Table 1 shows the accuracy of the some of the state of art models.

The Chart shown in Fig. 7 shows the same graphically.

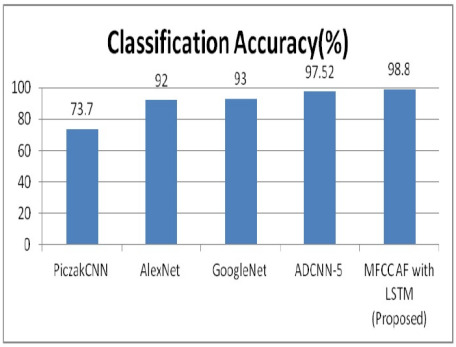

Figure 7: Comparison of Classification Accuracy

6. Conclusion

The proposed model uses the audio fingerprinting methodology to create a unique fingerprint of the audio files. The fingerprints are created by extracting a MFCC spectrum and then taking a mean of the spectra and converting the spectrum into a binary image. These images are then fed to the LSTM network to classify the environmental sounds stored in UrbanSound8K dataset and it produces an accuracy of 98.8% of accuracy across all 10 folds of the dataset.

In future, other features of audio signals like Chromagram can be used along with MFCC to improve the accuracy and to apply the model for speech classification.

References

[1] S. Baluja and M. Covell, *Audio fingerprinting: combining computer vision data stream processing*, IEEE Int. Conf. on Acoustics, Speech, and Signal Process. 2 (2007).
[2] V. Boddapati, A. Petef, J. Rasmusson and L. Lundberg, *Classifying environmental sounds using image recognition networks*, Procedia Comput. Sci. 112 (2017) 2048–2056.
[3] P. Cano and E. Batlle, *A review of audio fingerprinting*, J. VLSI Signal Process. 41 (2005) 271–284.
[4] M. Covell and S. Baluja, *Waveprint: Efficient wavelet-based audio fingerprinting*, Pattern Recognit. 41(11) (2008) 3467–3480.
[5] J.K. Das, A. Ghosh, A.K. Pal, S. Dutta and A. Chakrabarty, *Urban sound classification using convolutional neural network and long short term memory based on multiple features*, Fourth Int. Conf. Intell. Comput. Data Sci. (2020) 1–9.
[6] T. Elliot, J. Howard, R. Lisam, K. Fehling, A.d. Luca and D. Haba, *Dense vs convolutional vs fully connected layers*, https://forums.fast.ai/t/dense-vs-convolutional-vs-fully-connected-layers/191, (2016).
[7] D. Ellis, *Robust landmark-based audio fingerprinting*, Online Serial, 2009.
[8] Y. Jiang, C. Wu, K. Deng and Y. Wu, *An audio fingerprinting extraction algorithm based on lifting wavelet packet and improved optimal-basis selection*, Multimed. Tools Appl. 78 (2019) 30011–30025.
[9] T. Kalker and J. Haitsma, *A highly robust audio fingerprinting system*, Proc. ISMIR'2002, 2002 (2002) 144–148.
[10] H.B. Kekre, N. Bhandari and N. Nair, *A review of audio fingerprinting and comparison of algorithms*, Int. J. Comput. Appl. 70(13) (2013).
[11] F. Kurth, *A ranking technique for fast audio identification*, Proc. Int. Workshop Multimedia Signal Process. (2002) 186–189.
[12] I. Lezhenin, N. Bogach and E. Pyshkin, *Urban sound classification using long short-term memory neural network*, Federated Conf. Comput. Sci. Inf. Syst. (2019) 57–60.
[13] V. Nair and G.E. Hinton, *Rectified linear units improve restricted boltzmann machines*, Proc. 27th Int. Conf. Mach. Learn. (ICML-10) (2010) 807–814.
[14] K.J. Piczak, *Environmental sound classification with convolutional neural networks*, IEEE 25th Int. Workshop on Machine Learn. Signal Process. (2015) 1–6.
[15] T. Qiao, S. Zhang, Z. Zhang, S. Cao and S. Xu, *Sub-spectrogram segmentation for environmental sound classification via convolutional recurrent neural network and score level fusion*, arXiv preprint arXiv:1908.05863, (2019).
[16] G. Richly, L. Varga, F. Kovacs and G. Hosszu, *Short-term sound stream characterization for reliable, real-time occurrence monitoring of given sound-prints*, Proc. 10th Mediter. Electrotech. Conf. MeleCon 2 (2000) 526–528.
[17] T.N. Sainath, O. Vinyals, A. Senior and H. Sak, *Convolutional, long short- term memory, fully connected deep neural networks*, IEEE Int. Conf. Acoustics, Speech and Signal Process. (2015) 4580–4584.
[18] J. Salamon and J.P. Bello, *Deep convolutional neural networks and data augmentation for environmental sound classification*, IEEE Signal Process. Lett. 24(3) (2017) 279–283.
[19] J. Salamon, C. Jacoby and J.P. Bello, *A dataset and taxonomy for urban sound research*, Proc. 22nd ACM Int. Conf. Multimedia (2014) 1041–1044.
[20] J. Sharma, O.-C. Granmo and M. Goodwin, *Environment sound classification using multiple feature channels and attention based deep convolutional neural network*, Proc. Interspeech 2020 (2020) 1186–1190.
[21] S. Sri Ranjani, V. Abdulkareem, K. Karthik and P.K. Bora, *Application of SHAZAM-based audio fingerprinting for multilingual Indian song retrieval*, Adv. Commun. Comput. 347 (2015) 81–92.
[22] T. Stokes, *Spectro-temporal landmarking with rank-ordered local maxima for audio fingerprinting*, 16th Int. Soc. Music Inf. Retr. Conf. (2015).

[23] Y. Tokozume and T. Harada, *Learning environmental sounds with end-to-end convolutional neural network*, IEEE Int. Conf. Acoustics, Speech and Signal Process. (2017) 2721–2725.
[24] A. Wang, *The shazam music recognition service*, Comm. ACM, 49(8) (2006).
[25] X. Zhang, B. Zhu, L. Li, W. Li, X. Li, W. Wang, P. Lu and W. Zhang, *SIFT-based local spectrogram image descriptor: a novel feature for robust music identification*, EURASIP J. Audio, Speech, and Music Process. 2015(1) (2015) 1–15.