Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7778269

DESIGN AND IMPLEMENTATION OF DEEP LEARNING AND FEED FORWARD NEURAL NETWORK INTEGRATION MODEL FOR SINGER RECOGNITION AND CLASSIFICATION USING LSTM APPROACH

Arvind Kumar Sharma[1*], Sheng-Lung Peng[1], Pravin R. Kshirsagar[2], Prasun Chakrabarti[3]

[1]Department of Creative Technologies and Product Design, National Taipei University of Business, Taipei City, TAIWAN
[2]Department of Data Science, Tulsiramji Gaikwad Patil College of Engineering and Technology, Nagpur, INDIA
[3]Department of Computer Science and Engineering, ITM (SLS) Baroda University, Vadodara, INDIA

[*]Corresponding Author: drarvindkumarsharma@gmail.com

Abstract– This paper employs the use of a Deep Recurrent Neural Network approach namely Long Short-Term Memory (LSTM) to predict gender and singer name by analyzing audio vocal portions. The ultimate aim of this paper is to build two Long Short-Term Memory (LSTM) models, one for predicting singer gender or gender identification and the other for classifying singer name. The accuracy of different existing algorithms such as SVM, CNN and MLP is then compared to the LSTM algorithm. The MIR-1K dataset that contains audio recordings from singers is used to train all algorithms, including LSTM, SVM, CNN and MLP, with LSTM being the proposed algorithm and SVM, CNN and MLP being existing algorithms, to perform this integration model. The proposed LSTM approach for a deep recurrent neural network offers better performance than other existing ones. The obtained results show that the effectiveness of the proposed model is used together with a good enough feature vector which works well than the existing methods.

Keywords: *Music information retrieval (MIR), Speaker identification (SPID), Deep learning, MLP Network, LSTM, RNN, SVM, Classification*

1. INTRODUCTION

Now days, Music information retrieval (MIR) is an interdisciplinary area that analyzes musical contents by applying techniques of signal processing, deep learning and music theory. Speaker Identification is considered as a key area for research in the field of signal processing. Thus, singer identification as a subcategory of speaker identification (SPID) is a hot topic which has been grabbed the attention of the young researchers and refers to the identification of a singer in a music accompanied by other instruments [1]. Besides it, the growth of the music industry and while sophisticated recording techniques became more convenient, huge amount of songs are released and played through the internet, TV and radio channels every day. One of the most important features associated to a song is the singer. Many people use the voice of a singer to identify the songs quickly. The goal of

speaker identification is to identify a speaker using a computer. Both human and computer voices must be recognizable for one to be able to distinguish them. The job of comparing an unidentified speech to the training data and making the identification is the second part of speaker identification, or testing. The target speaker is the person who delivers a test speech. Alternative speech parameterizations based on formant characteristics have drawn considerable attention recently. Formant frequencies are crucial for the development of the speech spectrum. But it may be quite challenging to identify formants from a particular speech signal, and sometimes they may not be identified properly. For this reason, formant-like properties may be employed in place of predicting the resonant frequencies. Speaker recognition is classified into two parts: speaker identification and speaker verification. Matching the incoming voice sample with the existing voice samples is the goal of speaker identification [2]. Moreover speaker verification uses the available voice samples to identify the claimant. Speaker adaptation and identification have more uses than speaker verification. For example, in speaker identification, the speaker's voice may be used to identify him, but in speaker verification, the speaker is confirmed using a database. Speaker recognition is a biometric technique used to verify a user's identity by drawing certain traits from their voice expressions. It is an automated method that depends on the unique characteristics of the voice signal to acknowledge the speaker. In order to identify the speaker's uniqueness and manage access to services like voice dialing, voice mail, security control, etc., the speaker recognition system analyzes the speaker's vocal expressions. The basic architecture of an automatic speaker recognition system is shown in figure 1 below.

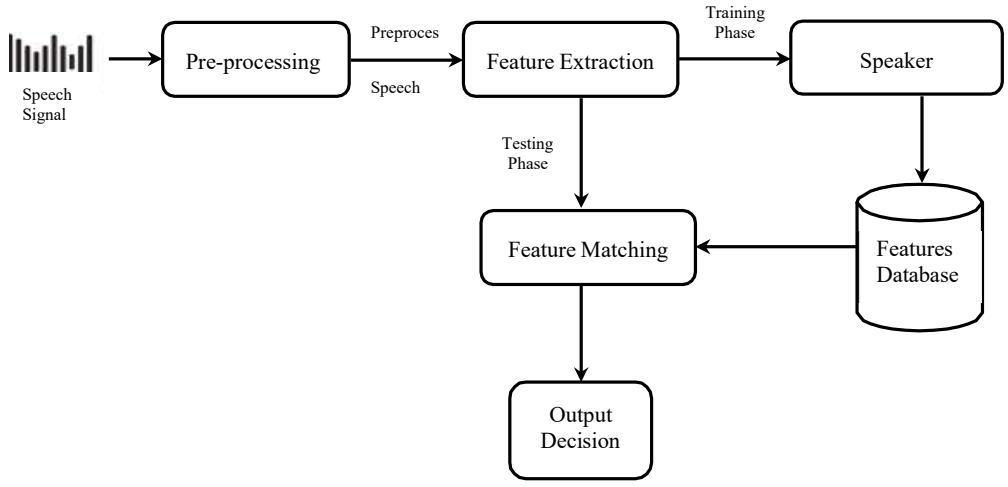

Figure 1: Architecture of an Automatic Speaker Recognition System

1.1 Pre-processing

The first step in an automated speech identification model is pre-processing. To create an efficient and dynamic ASR system, it is essential to execute this operation on the speech signal input. The voice signal is initially cleaned in this section of the speaker identification system. The signal is then cleaned up by removing the non-speech parts. Endpoint identification and pre-emphasis are the next basic tasks to be finished [3].

1.2 Feature Extraction

Feature extraction, often referred to as front end pre-processing, is used during the training and testing stages of speaker recognition systems. It uses feature vector or numerical descriptor sets to transform digital voice signals. The key elements of the speaker's speech are represented in these feature vectors [4].

1.3 Speaker Modeling

Speaker recognition algorithms are created using modeling techniques in order to match speaker speech features. Speaker models are characterized as processes that combine increased speaker-specific information with reduced volume [5]. State speaker models are created throughout training or enrollment by repeating the specific traits taken from the current speaker. For identification or verification tasks in the recognition stage, the speaker model is compared to the current speaker architecture. The development of several new services made possible by speaker recognition technology is anticipated to improve the convenience of our everyday lives.

1.4 Audio Features

A wide range of audio features has been used for Singer Identification (SID) purpose. The state of the art SID method selects a feature set, with which it can efficiently perform two tasks of detecting vocal segments and distinguishing among different singers.

Aim- *The purpose of this research work is to build and simulate a deep learning and feed forward neural network integration model for singer recognition and classification using an LSTM recurrent neural network approach. We propose a novel LSTM model to achieve the better accuracy in each sub-problem of music information retrieval research field. Moreover, experimental investigation presents that proposed method offers overall superior performance compared to other existing methods like SVM, CNN, MLP, and along with LSTM.*

Organization- *The rest of paper is organized as follows: Section 2 presents a brief review of related work in this field. All the deep learning methods used to identify singers are mentioned in Section 3. Section 4 covers the proposed methodology. Section 5 discusses experimental evaluation for the proposed model. Finally, Section 6 concludes the paper.*

2. RELATED WORK

Deep learning approaches have shown to be quite beneficial in the past few years as many researchers have worked on music information retrieval. In this section, we discuss relevant literature as well as contemporary developments.

In [6], Fu, Zhouyu *et al.* presented a music information retrieval (MIR), is an ascent study field that is attracting an increasing amount of interest from the scholarly community as well as the music business. It solves the challenge of searching and obtaining certain styles of music from a huge music data collection. The difficulty of classification is one of the most essential aspects of MIR. Numerous tasks in MIR are naturally suited to be cast in a framework of categorization, including but not limited to classification of genres, classification of moods, identification of artists and instruments, and so on. This review

places a strong emphasis on the most recent advancements in methodology and examines a number of unresolved problems that need more investigation.

In [7], Tsai, Wei-Ho *et al.* attempted to identify vocalists on recordings of popular music one of the most significant obstacles is finding a way to minimize the interference caused by the accompaniment in the background while still accurately describing the singer's voice. Despite the fact that a number of researches on automated singer identification (SID) based on acoustic aspects have been published, the majority of systems that have been developed to this day do not explicitly deal with the background accompaniment. Through the use of the fundamental connections that exist between solo singing voices and their accompanied counterparts in Cepstrum, this research work suggests a method for removing the background accompaniment that is present in singer identification (SID). It do the transition in order to turn the Cepstrum of an unknown accompanied voice into one that is similar to a solo voice whenever this system is provided with an unknown accompanied voice.

In[8], Pikrakis, Aggelos *et al.* presented a test music recording contains sung language that is not included in the data for calculating the transformation, findings suggest that using such a background reduction strategy enhances the singer identification (SID) accuracy substantially. When attempting to identify the vocalist on a recording of popular music, the primary emphasis of this research is on methods to lessen the interference caused by the background accompaniment. It suggests a strategy to the elimination of background accompaniment by making use of the underlying links that exist in cestrum between solo singing voices and the versions of those voices with accompaniment. A transformation that was approximated using a set of accompanied singing that was constructed by manually blending solo singing with accompaniments derived from Karaoke VCD was used to define the relationships. Even when a test music tape contains sung language that is not covered in the data for estimating the transformation, our findings demonstrate that such a background accompaniment elimination strategy results in a considerable improvement in singer identification (SID) accuracy.

In [9], Song, Liming, *et al.* proposed a method for the automated recognition of voice parts inside an acoustical polyphonic music signal. As the feature, make use of a mixture of many features that are unique to the singing voice, and they apply a Gaussian Mixture Model (GMM) classifier to differentiate between vocal and non-vocal sounds. They used a pre-processing technique called spectrum whitening, and as a result, they are able to get an overall performance of 81.3% on the RWC popular music dataset.

In [10], Tsai, Wei-Ho, and Hsin-Chieh Lee presented a framework of speaker identification (SPID) followed by the currently available techniques for singer identification (SID), which necessitates the collection of singing data in advance in order to determine the unique voice characteristics of each performer. This framework, on the other hand, is not appropriate for many SID applications due to the fact that collecting. Those applications focus on voice recognition. Many experiments have been conducted in an effort to enhance. However, the gains that have been made are not always adequate. This is because a cappella data are difficult to gather. This research studies the idea of defining singers' voices by utilizing the singers' spoken data rather than their singing data. The first approach proposes modifying a model that is developed from speech in order to account

for singing voice characteristics. The offered solutions have been validated by our studies, which were carried out utilizing a 20-singer database. The purpose of this research was to determine whether or whether it is possible to characterize singer's voices by utilizing their spoken data for SID.

In [11], Regnier *et al.* carried out two separate studies in order to confirm the identity of the song's performer. The second method, known as the song-level approach, examines the degree to which two song-based GMMs are comparable in order to determine whether or not they were sung by the same vocalist. TECC and INTO characteristics were used in the computation of the singer-based and song-based GMMs for each and every trial. On the basis of our dataset, the performance of the INTO features is superior than the TECC features. The information communicated by each kind of characteristics may be easily integrated by simply adding the distances that were acquired on each feature type individually. This is possible since the degree of similarity between songs, as well as between singers and songs, is determined by a distance. An EER of 7.5% is achieved while attempting to verify the identification of a vocalist by using a combination of characteristics, whereas an EER of 9% is achieved when attempting to verify the identity of a singer using a song.

In [12], Zhu, Bilei *et al.* presented a process of isolating the singing voice from the instrumental accompaniment may be useful for a variety of applications, including melody extraction, singer identification, lyrics alignment and recognition, and content-based music retrieval, among others. A brand new algorithm for separating the singing voice from monaural mixes is offered in this piece of research. The approach is divided into two steps, with the first stage using non-negative matrix factorization (NMF) to decompose the mixed spectrograms using large windows and the second stage using short windows.

In [13], Logan, Beth described Mel Frequency Cepstral Coefficients (MFCCs), which are the primary characteristics that are employed for speech identification, and analyze whether or not they can be applied to the modeling of music. In particular, they investigated the use of the Mel frequency scale to describe the spectra and the use of the Discrete Cosine Transform (DCT) in order to decorrelate the Mel-spectral vectors, which are two of the most important assumptions that are made throughout the process of creating MFCCs.

In [14], Hu, Ying, and Guizhong Liu discussed a technique that could isolate singing voice from music accompaniment for monaural recordings in order to enhance the performance of singer identification. This process is broken down into two main steps. In the first stage, the nonnegative matrix partial co-factorization (NMPCF), which is a joint matrix decomposition integrating prior knowledge of singing voice and pure accompaniment, is used to separate the mixture signal into singing voice portion and accompaniment portion. This is accomplished by separating the mixture signal into singing voice portion and accompaniment portion. In the second step, the separated singing voice that has acquired in the previous stage used as a basis for first estimating the pitches of the singing voice and then distinguishing the harmonic components of the singing voice.

In [15], Eronen *et al.* presented a method for recognizing musical instruments that is independent of pitch provided. The functionality of the features was tested by utilizing test data that comprised of 1498 samples encompassing the complete pitch ranges of 30 symphonic instruments from the string, brass, and woodwind families, each performed with

a distinct style. Accuracy of 94% has achieved in identifying the proper instrument family, while accuracy of 80% has achieved in identifying individual instruments.

3. NEURAL NETWORKS- DEEP LEARNING

This section describes the deep learning techniques used to address the problem specified earlier. Every action is controlled by neurons, i.e., our Nervous System. The actions depend on how the neurons are connected within the system and how strong is the connection between them. This is the main concept of Deep Learning. Several layers such as input, hidden and output are contained. Some basic terminology regarding neural networks includes Perceptron which can be defined as the simplest artificial neuron that only contains two layers- input and output. The first layer is processed by an activation function and the output is attained. The following figure 2 presents a simple perceptron.

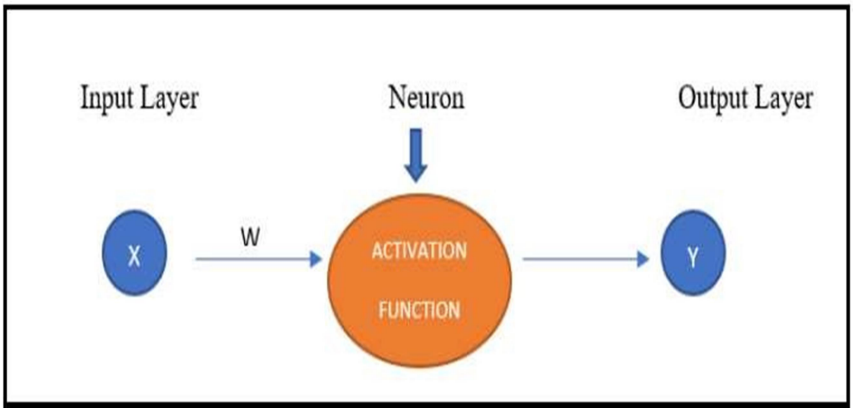

Figure 2: Representation of a Simple Perceptron

Activation function is responsible for converting the sum of weighted inputs into outputs. The result of one layer is passed over to the next layer as the input. Some activation functions are Sigmoid, Tanh, Relu, Softmax. A simple neural network consists of three layers. The first being the Input Layer, responsible for perceiving the input features of a problem. In certain cases, the number of nodes in this layer equals the number of input features. The data doesn't get changed when it passes through the input layer. The second one is the Hidden layer, which is responsible for transforming data. The processing in the hidden layer is done by weighted connections. The hidden layers increase the predictive power of a neural network. The final layer is the output layer which finally receives the connections from the hidden layer or the input layer directly and is responsible for predicting the output. In classification problems, we give the number of nodes in an output layer as one. The basic architecture of an artificial neural network is depicted in figure 3.

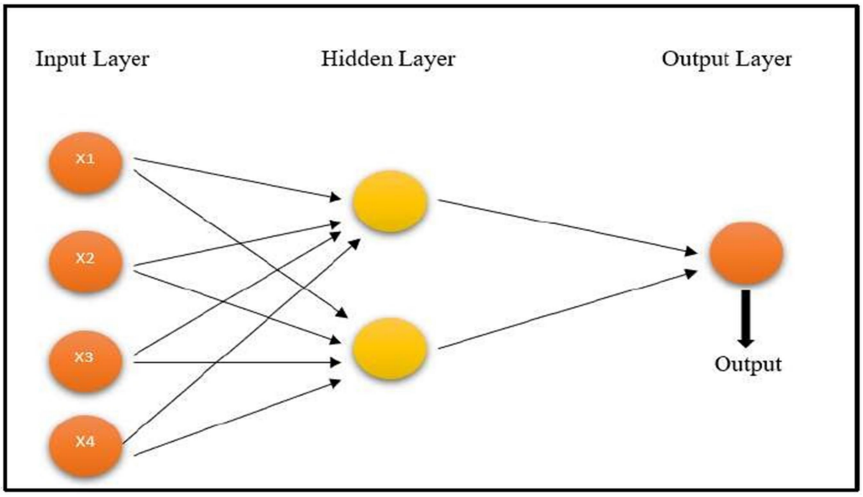

Figure 3: Layers of an Artificial Neural Network

The neural network learns to adjust the weights and also the threshold value (formula associated with the weights) to predict the correct outputs. Complex problems cannot be solved simply by using these three layers. Hence, more hidden layers are added to the neural network. Therefore, when we have more than three layers including the input and the output layer, the network is called a Deep Neural network. Also, training such complex networks is referred to as Deep Learning. The limitation that one comes across using neural networks is that neural networks need a large diversity of data in order to train themselves and give the correct predictions. Common applications of Neural networks are Voice recognition, character recognition, signature verification, and financial forecasting, etc.

3.1 LSTM Neural Network

One of the most well-known types of recurrent neural networks (RNN) is called a Long Short-Term Memory (LSTM) network. Long Short Term Memory Networks (LSTMs) are capable of learning long-term dependencies in data by addressing the vanishing gradient problem. Remembering information for long periods of time is practically their default behaviour, not something they struggle to learn [16]. LSTM models have the capacity to review past data and choices and draw conclusions from them. However, they must also use the same procedure to generate intelligent hypotheses and predictions about potential outcomes. Because of this, feeding sequential data to this model works well. It will look for tendencies and utilize those trends to forecast future developments. Moreover it contrasts LSTMs with GRUs.

3.2 Multi-Layer Perceptron (MLP)

An MLP is an artificial neural network that works like a feed-forward network. An MLP is made up of three layers: firstly Input Layer, secondly Hidden Layer, and thirdly Output Layer. Every layer besides the input layer utilizes a nonlinear activation function for training purposes. The distinction between an MLP and a single layer perceptron is that the MLP has multiple layers, allowing it to differentiate between data that are not linearly separable.

3.3 Recurrent Neural Network (RNN)

The RNN is a neural network that is designed specifically for processing a sequence of values $\left(x^{(1)}, \ldots\ldots, x^{(\tau)}\right)$ which is based on an early idea proposed in machine learning and statistical models: sharing parameters across different parts of a model [17]. Similar to convolutional neural networks that share weights across pixels, RNNs share weights across time steps. Due to their capacity to memorize, RNNs are very efficient in making predictions as they are able to remember and hence expect what is about to come. Due to this reason precisely, they are the preferred choice of algorithm for data which is sequential like series of time, text, speech, data on finance, video, audio, climate and many more as they can develop an understanding of the deeper kind along with its context when compared to other algorithms.

3.4 Support Vector Machine (SVM)

Support Vector Machine (also referred to as Support Vector Networks) could be a supervised learning model that's accustomed to analyze each classification and regression. When compared to different algorithms, Support Vector Machines proves to be a lot of sturdy prediction methodology. It is a non-probabilistic binary linear classifier as a result of which it assigns examples to a minimum of one class or the alternative. SVM maps coaching examples to points in the house thus maximizing the breadth of the gap between the two classes. New examples square measure a unit then mapped into that very same house and predicted to belong to a category supported that facet of the gap they fall. Support Vector Machines may classify non-linear problems by mapping the inputs to higher-dimensional space. Following figure 4 shows SVM, which finds a hyperplane or a bunch of hyperplanes in degree infinite-dimensional space.

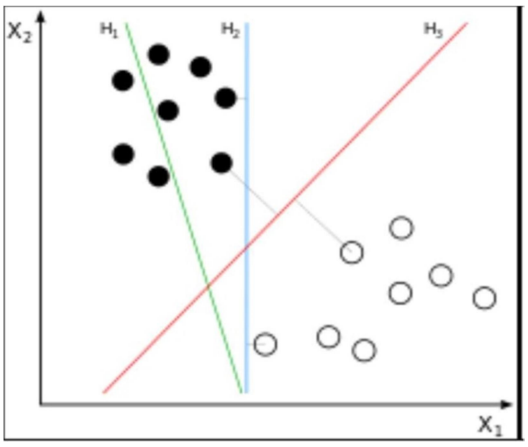

Figure 4: Hyperplane Selection in SVM

The equation of a hyperplane is given as: $w.x + b = 0$ (1)

To calculate the distance of a point from a line in a plane, the following formula is used:

$$d=\frac{|ax+by+c|}{\sqrt{a^2+b^2}} \quad (2)$$

Now, the distance of a hyperplane equation for a given vector point Φ(x) as follows:

$$d(\emptyset(x)) = \frac{wx+b}{\sqrt{w1^2+w2^2+\cdots.}} \quad (3)$$

4. PROPOSED METHODOLOGY

Our proposed methodology describes how actionable development work should be conducted out. It is recognized as a tool that is constantly employed to discover some area for which data is collected and analyzed. The proposed framework is intended to simulate in the development of deep recurrent neural networks model using various deep learning techniques. This proposed work presents a singer recognition system by incorporating LSTM (Long Short Term Memory) Deep learning and Feed-forward neural networks, which has not already been used for this purpose to the best of our knowledge.

The preprocessing task involves, studying of large sets of audio features in order to extract the most efficient set for the recognition stage. Implementation process of our proposed work is carried out into various stages. First of all the vocal frames of all music clips are detected using an LSTM network which can perform well for the time series data, such as audio signals. Then, an MLP network is incorporated and compared with an SVM classifier in order to classify the gender of singers. Finally, another LSTM network is simulated to detect each singer's ID and compared with an MLP network in the same way. In all stages, different classifiers are trained and tested here.

4.1 Key Objectives of Proposed Methodology

The key objectives of our proposed methodology are as follows:

- To propose a deep recurrent neural network approach named LSTM (Long Short-term Memory) to predict singer's gender and singer's name by analyzing audio vocal parts.
- To build two LSTM models where one is used to predict singer's gender or gender identification and other one is used to predict or classify singer's name.
- To use the MIR-1K dataset which contains signer's audio files and train various algorithms such as SVM, MLP, CNN and LSTM.
- To compare the accuracy of the existing algorithms such as SVM, CNN, and MLP along with LSTM.

4.2 Dataset Collection

To conduct experiments, a dataset is prepared. For this purpose, samples are collected from the MIR-1K dataset [18][19][20] designed for research on the separation of vocal voices. MIR-1K (Multimedia Information Retrieval lab) consists of 1000 songs with vocal activation and pitch contours annotations. MIR-1K has coined by Chao-Ling Hsu *et al.* in [21]. For each song, the segments are annotated as "voice" and "no voice". The sampling rate is 16 kHz, and for all the 1000 clips, the clip duration is 4-13 sec. The MIR-1K dataset is provided as input to the LSTM model, from which MFCC features are extracted and trained to our model. The assessment procedure then begins, in which new data or test data are provided and actual results are compared to the predicted results to find accuracy. All dataset audio files are saved on the local storage of our machine in the Dataset directory.

4.3 Dataset Upload

It begins with the collection of datasets and the identification of singing voice segments. Additionally, various timbre and temporal characteristics that are suitable for singer identification are extracted. They are supplied to two distinct classifiers for testing and training purposes. In this phase the dataset is initially loaded from MIR-1K Dataset and pre-processed to remove irrelevant information. The user interface of proposed model is shown in figure 5.

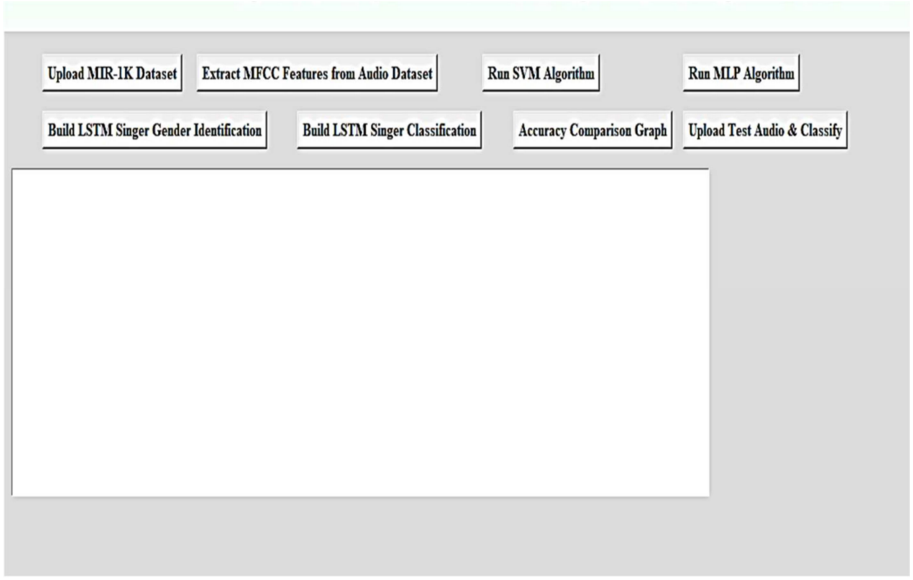

Figure 5: User Interface of Proposed Model

5. EXPERIMENTS AND RESULTS

The experimental setup is conducted on a machine with Intel Core i5-7500 CPU @3.4GHz, running Window 64-bit operating system with 16 GB RAM, GPU GeForce GTX 1080 Ti, and 1TB of local storage. Once the dataset has been preprocessed, this implementation of the proposed model should be carried out here. The model is designed and developed by using Python technologies.

5.1 Vocal and Non-vocal Segmentation

In this section we use deep recurrent neural network algorithm called LSTM (Long Short Term Memory) to predict singer's gender by analysing audio vocal parts and to predict singer's name. In this work we build two LSTM models where one is used to predict singer's gender and other one is used to predict/classify singer's name. Moreover, we compare accuracy of existing algorithms such as SVM, MLP, CNN, and with LSTM.

5.2 Simulation Details

The simulation work of our proposed model is conducted in this section. We are used 75% of MIR-1K dataset containing signer's audio files for training the classifiers and remaining 25% dataset is used for testing process of the model. First of all, we import the dataset into the data model of the proposed framework followed by the training phase. After training the LSTM neural network model, test the system with different audio files which are not utilized to train the system. The experimental results are discussed below one by one.

Step-I: To upload the needed MIR-1K dataset from the proposed model, click on the 'Upload MIR-1K Dataset' button on the screen shown below in figure 6.

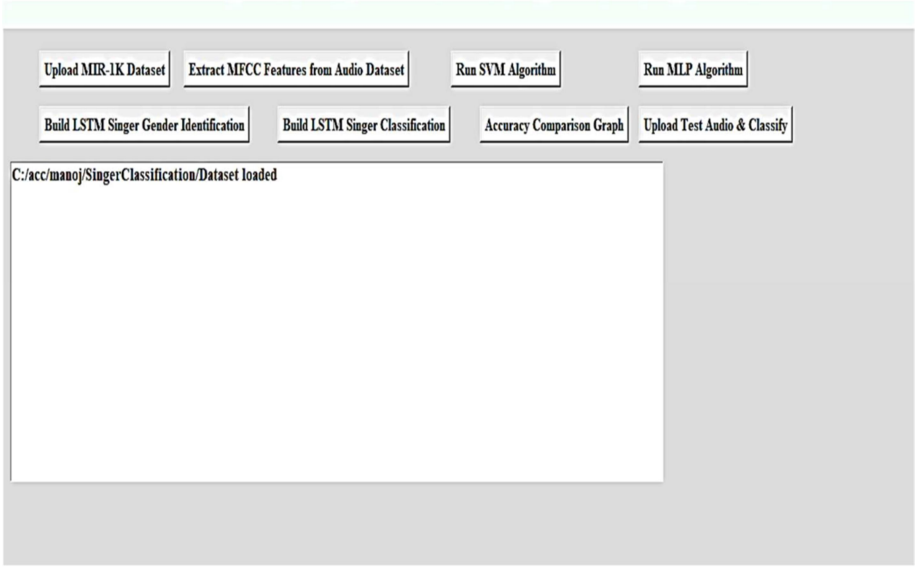

Figure 6: Extract MFCC Features from Audio Dataset Module

Step-II: In this phase, dataset contains total 108 audio files belong to Female and Male and total singers found in 108 files are 10. In above module 140 are the multidimensional array which contains audio features or data. We use these 108 audio files features to train algorithms. Now data is ready and click on 'Run SVM Algorithm' button to train SVM on above features and it calculates predicted accuracy.

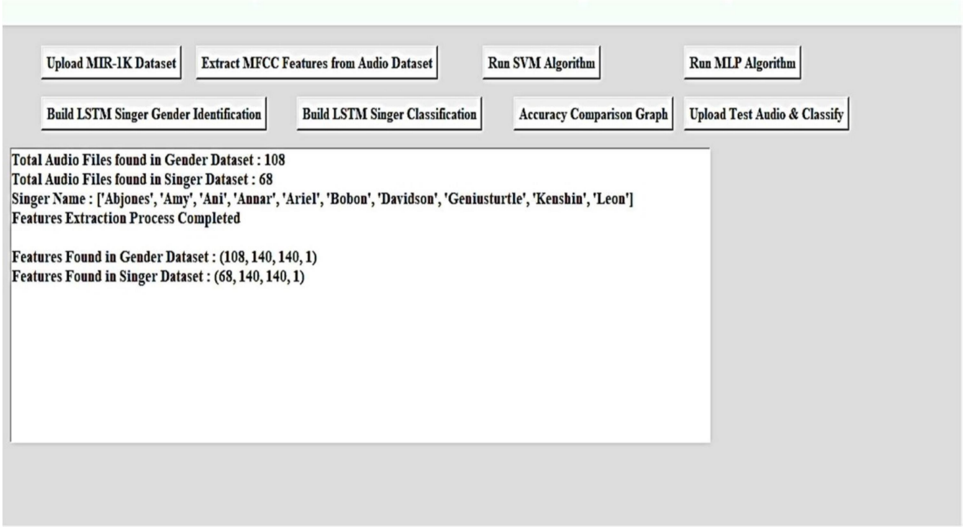

Figure 7: Results of Existing SVM Algorithm

5.3 Accuracy Analysis for Gender Prediction

After a comparative analysis has done on various Deep Learning classifiers by using MIR-1K Dataset, the following results are obtained as shown in table-1 below.

Table-1: Accuracy Analysis for Gender Prediction

Methods	Accuracy (in %)
SVM [22]	80.10
MLP [22]	87.60
CNN [23]	89.57
LSTM (Proposed)	93.12

The accuracy analysis is shown in the following figure 8 in a more visual manner.

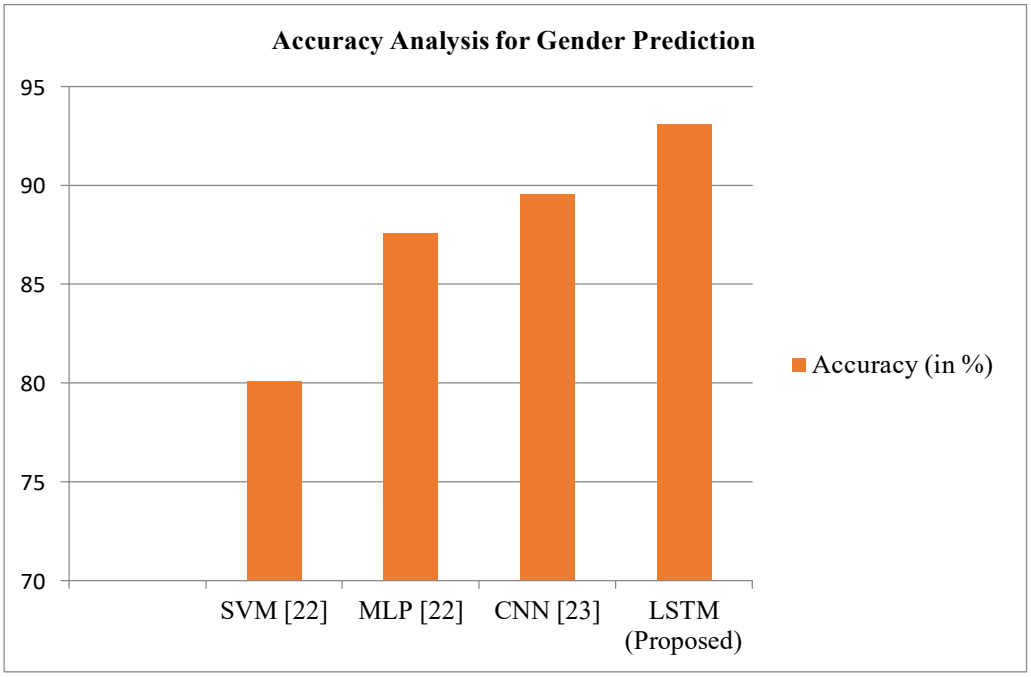

Figure 8: Performance Analysis of various Models for Gender Prediction

5.4 Accuracy Analysis for Singer Name Prediction

After training and evaluating the different Deep Learning models for singer name prediction,the following results are obtained as shown in table-2 below.

Table-2: Accuracy Analysis for Singer Name Prediction

Methods	Accuracy (in %)
SVM [22]	86.93
MLP [22]	90.01
CNN [23]	88.90
LSTM (Proposed)	94.24

The accuracy analysis is depicted in a more visual form in the figure 9 below.

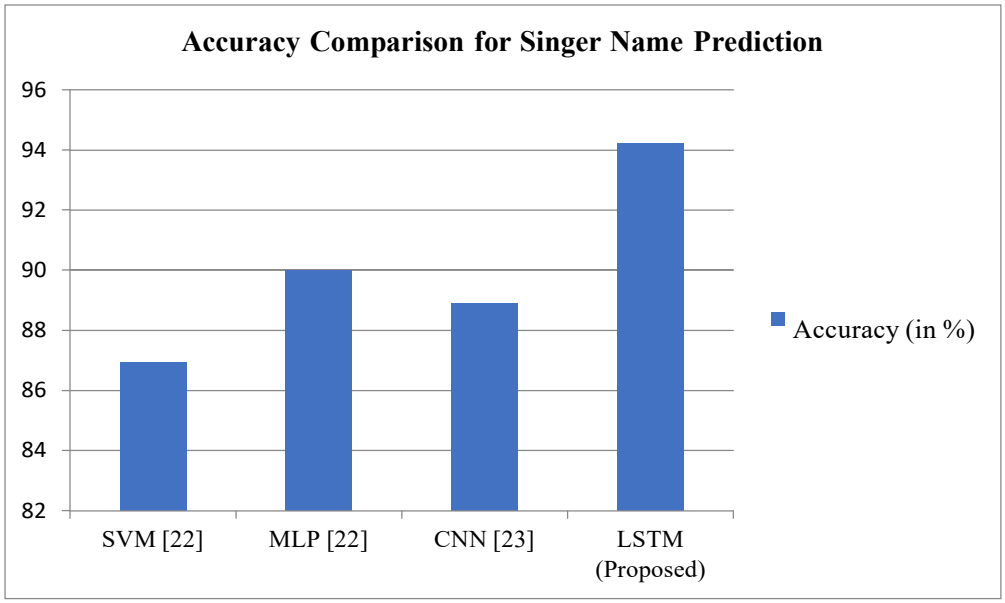

Figure 9: Performance Analysis of various Models for Singer Name Prediction

The accuracy of the classifier for the Support Vector Machine is 86.93%, that of the MLP classifier is 90.01%, and that of the CNN classifier is 88.90%. Better accuracy is provided by our proposed LSTM model, which is 94.24%.Thus, it follows that our proposed model performs better than the existing ones. The LSTM Network stands out as having the highest accuracy when the accuracy of the aforementioned deep learning models is evaluated. As a result, the LSTM Network is used to predict the name of singers.

6. CONCLUSION

In this paper a novel approach for singer recognition based on deep learning and feed-forward neural networks is presented. The methodology employed here is to first identify the vocal parts and classify the vocalist's gender before attempting to identify the singer. As a result, a model is suggested here to identify a singer's gender and singer's name from his or her voice recordings. After training various deep learning classifiers for gender and name, it is found that LSTM neural network model produces the most accurate results. Moreover the singer's gender and name are predicted as male or female and as a cohort, respectively. The experimental results show that the proposed method is more accurate than the existing methods for the MIR-1K dataset. The usage of real-time data will be recommended for a future work. Finally, the scope of our research is wide enough to cover a variety of scenarios and it may be simply modified to address new concerns as they arise in the future.

Acknowledgements

I would like to express my deep gratitude to Prof. (Dr.) Sheng-Lung Peng, Prof. (Dr.) Pravin R. Kshirsagar and Prof. (Dr.) Prasun Chakrabarti of my Postdoctoral research for their valuable guidance, enthusiastic encouragement, and constructive suggestions during the planning and development of this research work.

References:

[1] Fu, Zhouyu, Guojun Lu, *et al.,* "A survey of Audio-Based Music Classification and Annotation", IEEE Transactions on Multimedia 13, No. 2 (2011): 303-319.

[2] Tsai, Wei-Ho, and Hao-Ping Lin, "Background Music Removal Based on Cepstrum Transformation for Popular Singer Identification", IEEE Transactions on Audio, Speech, and Language Processing 19, No. 5 (2011): 1196-1205.

[3] Pikrakis, Aggelos, *et al., "*Unsupervised Singing Voice Detection using Dictionary Learning", In Signal Processing Conference (EUSIPCO), 2016 24th European, pp. 1212-1216. IEEE, 2016.

[4] Song, Liming, Ming Li, and Yonghong Yan, "Automatic Vocal Segments Detection in Popular Music", In 2013 Ninth International Conference on Computational Intelligence and Security, pp. 349-352. IEEE, 2013.

[5] Tsai, Wei-Ho, and Hsin-Chieh Lee, "Singer identification based on spoken data in voice characterization", IEEE Transactions on Audio, Speech, and Language Processing 20, No. 8 (2012): 2291-2300.

[6] Tang, Z., Li, L., Wang, D., "Multi-Task Recurrent Model for Speech and Speaker Recognition", In: 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), South-Korea, 2016, pp.1-4.

[7] Sharmila Biswas, Sandeep Singh Solanki, "Speaker recognition: an Enhanced Approach to Identify Singer Voice using Neural Network", International Journal of Speech Technology, Vol. 24, No.1, pp.9, 2021.

[8] Pikrakis, Aggelos, *et al.,"*Unsupervised Singing Voice Detection using Dictionary Learning", In Signal Processing Conference (EUSIPCO), 24th European, pp. 1212-1216. IEEE, 2016.

[9] Song, Liming, Ming Li, and Yonghong Yan, "Automatic Vocal Segments Detection in Popular Music", In the Ninth International Conference on Computational Intelligence and Security, pp. 349-352. IEEE, 2013.

[10] Graves Alex, *et al., "*Hybrid Speech Recognition with Deep Bidirectional LSTM", Automatic Speech Recognition and Understanding (ASRU) 2013 IEEE Workshop on, pp. 273-278, 2013.

[11] Regnier, Lise, and Geoffroy Peeters, "Singer Verification: Singer Model vs. Song Model", In Acoustics, Speech and Signal Processing (ICASSP), In the IEEE International Conference on, pp. 437-440. IEEE, 2012.

[12] Zhu, Bilei, Wei Li, Ruijiang Li, and Xiangyang Xue, "Multi-stage Non-negative Matrix Factorization for Monaural Singing Voice Separation", IEEE Transactions on Audio, Speech, and Language Processing 21, No. 10 (2013): 2096-2107.

[13] Logan, Beth, "Mel Frequency Cepstral Coefficients for Music Modeling", In ISMIR, Vol. 270, pp. 1-11. 2000.

[14] Hu, Ying, and Guizhong Liu, "Separation of Singing Voice by use of Nonnegative Matrixpartial Co-factorization for the Purpose of Singer Identification", IEEE Transactions on Audio, Speech, and Language Processing: (2015) 643-653.

[15] Eronen, Antti, and Anssi Klapuri, "Musical instrument recognition using Cepstral coefficients and temporal features", In Acoustics, Speech, and Signal Processing (ICASSP), IEEE International Conference on, Vol. 2, pp. II753-II756. IEEE, 2000.

[16] Yashu Kedi, Nagadevi S, "Speaker Identification from Voice", Journal of Current Research in Engineering & Science, Vol. 5, Issue-1, January, 2022.

[17] Good fellow, Y. Bengio, and A. Courville, Deep Learning, Cambridge, MA, USA: MIT Press, 2016.

[18] J.S. R. J. Chao-Ling Hsu, "http://www.mirlab.org" [Online]: http://mirlab.org/ dataset/public/MIR-1K.rar.

[19] C.L. Hsu *et al.* "Dataset for Singing Voice Separation" [Online]: http://mirlab.org/ dataset/public/MIR-1K.rar.

[20] S. Chen, S. Paris *et al.*, "Singing-Voice Separation from Monaural Recordings using Robust Principal Component Analysis," IEEE ICASSP, Vol. 987-1-4673-0046-9, No.12, pp. 57-60, 2012.

[21] Hsu, C.L.; Jang, J.S.R, "On the Improvement of Singing Voice Separation for Monaural Recordings using the MIR-1K dataset", IEEE Transactions Audio Speech Lang. Process, 2009, 18, 310-319.

[22] Seyed Kooshan Hashemi Fard, Rahil Mahdian Toroghi, "Singer Identification by Vocal Parts Detection and Singer Classification using LSTM Neural Networks," 4th International Conference on Pattern Recognition and Image Analysis (IPRIA), 6-7 March, Tehran, Iran.

[23] Gui, W.; Li, Y.; Zang, X.; Zhang, J., "Exploring Channel Properties to Improve Singing Voice Detection with Convolutional Neural Networks", Applied Science, 2021, 11, 11838.