Long Short-Term Memory Model for Singer Gender Identification and Singer Classification by Vocal Parts

K. Venkatakrishna[1], Resham Akhila[2], S. Sreenidhi [2], S. Rakesh Reddy[2], Ch. Sai Tejat[2], S. Shruthi[2]

[1]Assistant Professor, [2]UG Scholar, [1,2]Department of Computer Science Engineering
[1,2]Malla Reddy Engineering College and Management Sciences, Medchal, Hyderabad

ABSTRACT

In this paper we are using deep recurrent neural network algorithm called LSTM (long short-term memory) to predict gender by analyzing audio vocal part and to predict singer name. In proposed work author is building two LSTM model where one is used to predict singer gender or gender identification and other is used to predict/classify singer name. This work focused on implementation of LSTM model for signer classification (recognition of name) along with gender identification (i.e., either male or female). Initially, MIR-1k dataset is considered to implement this work, which contains both signer-name specific and gender specific speech files. Then, pre-processing operation is carried out on both datasets performed, which removed the noises from speech files. Then, MFCC features are extracted only from speech data, which contains the higher order spectral features. Further, LSTM model is trained with both speeches-based Mel-frequency cepstral coefficients (MFCC) features. Finally, test speech data is applied for prediction purposed and test features are compared with the pre-trained LSTM model features. Finally, the predicted singer-name and singer gender is obtained through this LSTM model. The simulation results shows that the proposed LSTM method shows superior performance over state of art techniques.

Keywords—Music Information Retrieval, LSTM Network, MLP network, Classification, Polyphonic Music signal

1. INTRODUCTION

The task of speaker identification is to determine the identity of a speaker by machine. To recognize voice, the voices must be familiar in case of human beings as well as machines. The second component of speaker identification is testing; namely the task of comparing an unidentified utterance to the training data and making the identification. The speaker of a test utterance is referred to as the target speaker. Recently, there has been some interest in alternative speech parameterizations based on using formant features. To develop speech spectrum formant frequencies are very essential.

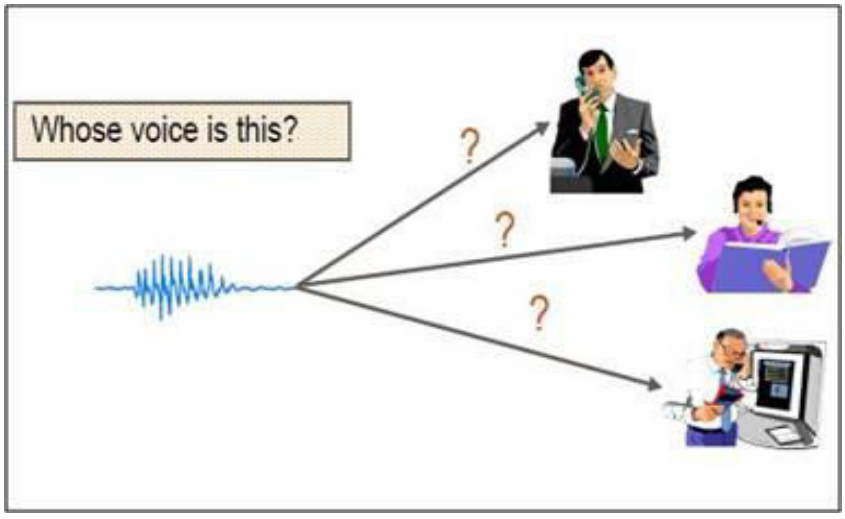

Fig.1. speaker Identification

But formants are very difficult to find from given speech signal and sometimes they may be not found clearly. That's why instead of estimating the resonant frequencies, formant-like features can be used. Depending upon the application the area of speaker recognition is divided into two parts. One is identification and other is verification. In speaker identification the aim is to match input voice sample with available voice samples. And in speaker verification, from available voice sample to determine the person who is claiming. Speaker identification and adaption have various applications than speaker verification. In speaker identification for example the speaker can be identified by his voice, where in case of speaker verification the speaker is verified using database. Speaker recognition is a biometric scheme applied to authenticate user's individuality using the specific characteristics elicited from their speech utterances. It is the automatic process of acknowledging the speaker depending on the speech signal's characteristic features. The Speaker recognition system uses the speaker's voice utterances to recognize their individuality and control access to services, such as voice dialing, voice mail, security control, etc.

2. LITERATURE SURVEY

Fu,zhouyu, et al. [1] Music information retrieval (MIR) is an emerging research area that receives growing attention from both the research community and music industry. It addresses the problem of querying and retrieving certain types of music from large music data set. Classification is a fundamental problem in MIR. Many tasks in MIR can be naturally cast in a classification setting, such as genre classification, mood classification, artist recognition, instrument recognition, etc. Music annotation, a new research area in MIR that has attracted much attention in recent years, is also a classification problem in the general sense. Due to the importance of music classification in MIR research, rapid development of new methods, and lack of review papers on recent progress of the field, we provide a comprehensive review on audio-based classification in this paper and systematically summarize the state-of-the-art techniques for music classification. Specifically, we have stressed the difference in the features and the types of classifiers used for different classification tasks. This survey emphasizes on recent development of the techniques and discusses several open issues for future research.

Tsai, et al. [2] One major challenge of identifying singers in popular music recordings lies in how to reduce the interference of background accompaniment in trying to characterize the singer voice. Although a number of studies on automatic Singer Identification (SID) from acoustic features have been reported, most systems to date, however, do not explicitly deal with the background accompaniment. This study proposes a background accompaniment removal approach for SID by exploiting the underlying relationships between solo singing voices and their accompanied versions in Cepstrum. The relationships are characterized by a transformation estimated using a large set of accompanied singing generated by manually mixing solo singing with the accompaniments extracted from Karaoke VCDs. Such a transformation reflects the Cepstrum variations of a singing voice before and after it is added with accompaniments. When an unknown accompanied voice is presented to our system, the transformation is performed to convert the Cepstrum of the accompanied voice into a solo-voice-like one. Our experiments show that such a background removal approach improves the SID accuracy significantly; even when a test music recording involves sung language not covered in the data for estimating the transformation.

Pikrakis, et al. [3] This paper presents an unsupervised approach to vocal detection in music recordings based on dictionary learning. At a first stage, the recording to be segmented is treated as training data and the K-SVD algorithm is used to learn a dictionary which sparsely represents a short-term feature sequence that has been extracted from the recording. Subsequently, the vectors of the feature sequence are reconstructed based on the learned dictionary and the probability of appearance

of the dictionary atoms is estimated. The obtained probability serves to compute the value of a weight function for each frame of the recording. The histogram of this function is then used to estimate a binarization threshold that segments the recording into vocal and non-vocal segments. The performance of the proposed unsupervised method, when evaluated on two datasets of accompanied singing, presents comparable performance to supervised techniques.

Song, et al. [4] proposed a technique for the automatic vocal segment's detection in an acoustical polyphonic music signal. We use a combination of several characteristics specific to singing voice as the feature and employ a Gaussian Mixture Model (GMM) classifier for vocal and non-vocal classification. We have employed a pre-processing of spectral whitening and archived a performance of 81.3% over the RWC popular music dataset.

Tasi, et al. [5] Currently existing singer identification (SID) methods follow the framework of speaker identification (SPID), which requires that singing data be collected beforehand to establish each singer's voice characteristics. This framework, however, is unsuitable for many SID applications, because acquiring solo a cappella from each singer is usually not as feasible as collecting spoken data in SPID applications. Since a cappella data are difficult to acquire, many studies have tried to improve SID accuracies when only accompanied singing data are available for training; but, the improvements are not always satisfactory. Recognizing that spoken data are usually available easily, this work investigates the possibility of characterizing singers' voices using the spoken data instead of their singing data. Unfortunately, our experiment found it difficult to replace singing data fully by using spoken data in singer voice characterization, due to the significant difference between singing and speech voice for most people. Thus, we propose two alternative solutions based on the use of few singing data. The first solution aims at adapting a speech-derived model to cover singing voice characteristics. The second solution attempts to establish the relationships between speech and singing using a transformation, so that an unknown test singing clip can be converted into its speech counterpart and then identified using speech-derived models; or alternatively, training data can be converted from speech into singing to generate a singer model capable of matching test singing clips. Our experiments conducted using a 20-singer database validate the proposed solutions.

Regnier, et al. [6] This paper proposes a method to verify the singer identity of a given song. The query song is modeled as a GMM learned on the features extracted from sustained sung notes of the song. Each note is described by the shape its spectral envelope and by the temporal variations in frequency and amplitude of its fundamental frequency. The singer identity is verified with two approaches: the model of the query song is compared to a singer-based GMM or compared to the GMM of another song performed by the same singer. The comparison is done using a dissimilarity measurement given by the Kullback Leibler divergence. When the two types of features are combined, the proposed approach verifies the singer identity of a given a cappella song with an error rate lower than 8% when the whole song is considered and an error rate lower than 10% when a short excerpt of the song (i.e. 15 consecutive sustained notes) is considered. The second approach, song-level, verifies if two songs are performed by the same singer by measuring the similarity between two song-based GMM. For all experiments, the singer-based and the song-based GMM were computed using either TECC or INTO features. On our dataset, INTO features perform better than the TECC features. Since the song-to-song (and song-tosinger) similarity is computed by means of a distance, information conveyed by each type features can be simply combined by summing the distances obtained on each feature type separately. Using the combination of features, the identity of a singer is verified with an EER of 7.5% for the singer-level approach and 9% for the song level approach.

Zhu, et al. [7] Separating singing voice from music accompaniment can be of interest for many applications such as melody extraction, singer identification, lyrics alignment and recognition, and

content-based music retrieval. In this paper, a novel algorithm for singing voice separation in monaural mixtures is proposed. The algorithm consists of two stages, where non-negative matrix factorization (NMF) is applied to decompose the mixture spectrograms with long and short windows respectively. A spectral discontinuity thresholding method is devised for the long-window NMF to select out NMF components originating from pitched instrumental sounds, and a temporal discontinuity thresholding method is designed for the short-window NMF to pick out NMF components that are from percussive sounds.By eliminating the selected components, most pitched and percussive elements of the music accompaniment are filtered out from the input sound mixture, with little effect on the singing voice. Extensive testing on the MIR-1K public dataset of 1000 short audio clips and the Beach-Boys dataset of 14 full-track real-world songs showed that the proposed algorithm is both effective and efficient.

Hu, et al. [8] In order to improve the performance of singer identification, we propose a system to separate singing voice from music accompaniment for monaural recordings. Our system consists of two key stages. The first stage exploits the nonnegative matrix partial co-factorization (NMPCF), which is a joint matrix decomposition integrating prior knowledge of singing voice and pure accompaniment to separate the mixture signal into singing voice portion and accompaniment portion. In the second stage, based on the separated singing voice obtained by the first stage, the pitches of singing voice are first estimated and then the harmonic components of singing voice can be distinguished. For a frame, the distinguished harmonic components are regarded as reliable while other frequency components unreliable, thus the spectrum is incomplete. With those harmonic components, the complete spectrums of singing voice can be reconstructed by a missing feature method, spectrum reconstruction, obtaining a refined signal with more clean singing voice. Experimental results demonstrate that, from the point view of source separation, the singing voice refinement can further improve ΔSNR in contrast with the singing voice separation using NMPCF, while for the point view of singer identification, the singing voice separated by NMPCF is more appropriate than the refined singing voice.

Logan, et al. [9] examine in some detail Mel Frequency Cepstral Coefficients (MFCCs) - the dominant features used for speech recognition - and investigate their applicability to modeling music. In particular, we examine two of the main assumptions of the process of forming MFCCs: the use of the Mel frequency scale to model the spectra; and the use of the Discrete Cosine Transform (DCT) to decorrelate the Mel-spectral vectors.

A system for musical instrument recognition was presented that uses a wide set of features to model the temporal and spectral characteristics of sounds. Signal processing algorithms were designed to measure these features in acoustic signals. Using this input data, a classifier was constructed and the usefulness of the features was verified. Furthermore, experiments were carried out to investigate the potential advantage of a hierarchically structured classifier. The achieved performance and comparison to earlier results demonstrates that combining the different types of features succeeded in capturing some extra knowledge about the instrument properties. Hierarchical structure could not bring further benefits, but its full potential should be reconsidered when a wider data set including more instruments, as well as different examples from a particular instrument class is available. Future work will concentrate on these areas, and on integrating the recognizer into a system that is able to process more complex sound mixtures.

3. PROPOSED METHOD

This work introduces LSTM model of singer identification. To the best of our knowledge, it is the first SID system which involves deep neural networks for this purpose, to date. This system consists of multi-stage processing blocks. Figure 2 shows the proposed block diagram of signer classification

(recognition of name) along with gender identification (i.e., either male or female). Initially, MIR-1k dataset is considered to implement this work, which contains both signer-name specific and gender specific speech files. Then, pre-processing operation is carried out on both datasets performed, which removed the noises from speech files. Then, MFCC features are extracted only from speech data, which contains the higher order spectral features. Further, LSTM model is trained with the both speeches based MFCC features. Finally, test speech data is applied for prediction purposed and test features are compared with the pre-trained LSTM model features. Finally, the predicted singer-name and singer gender is obtained through this LSTM model.

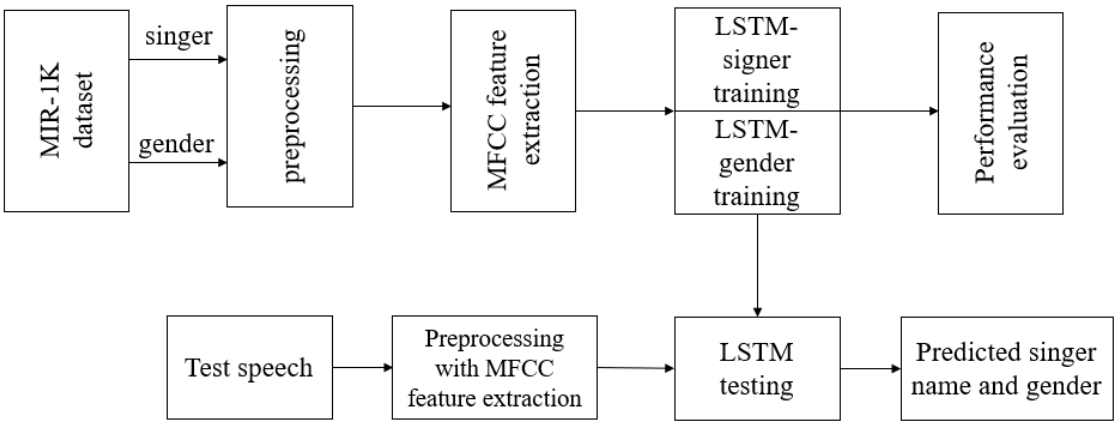

Fig.2. Proposed block diagram

3.1 MIR-1K dataset

The MIR-1K dataset contains the mainly two sub-parts, those are singer specific and gender specific. Here, singer-name specific dataset is gathered from 10-speakers, with maximum of five samples from each speaker. Similarly, gender specific dataset is formed by separating male and female singers voice clips of same singer-name specific dataset. Finally, the 108 samples presented in gender dataset and singer dataset.

3.2 Pre-processing

Digital speech processing is the use of computer algorithms to perform speech processing on digital speeches. As a subfield of digital signal processing, digital speech processing has many advantages over analogue speech processing. It allows a much wider range of algorithms to be applied to the input data — the aim of digital speech processing is to improve the speech data (features) by suppressing unwanted distortions and/or enhancement of some important speech features so that our AI-Computer Vision models can benefit from this improved data to work on. To train a network and make predictions on new data, our speeches must match the input size of the network. If we need to adjust the size of speeches to match the network, then we can rescale or singer-name & gender data to the required size.

we can effectively increase the amount of training data by applying randomized augmentation to data. Augmentation also enables to train networks to be invariant to distortions in speech data. For example, we can add randomized rotations to input speeches so that a network is invariant to the presence of rotation in input speeches. An augmented Speech Datastore provides a convenient way to apply a limited set of augmentations to 2-D speeches for classification problems.

we can store speech data as a numeric array, an Speech Datastore object, or a table. An Speech Datastore enables to import data in batches from speech collections that are too large to fit in memory.

we can use an augmented speech datastore or a resized 4-D array for training, prediction, and classification. We can use a resized 3-D array for prediction and classification only.

There are two ways to resize speech data to match the input size of a network. Rescaling multiplies the height and width of the speech by a scaling factor. If the scaling factor is not identical in the vertical and horizontal directions, then rescaling changes the spatial extents of the pixels and the aspect ratio.

crop extracts a subregion of the speech and preserves the spatial extent of each pixel. We can crop speeches from the centre or from random positions in the speech. A speech is nothing more than a two-dimensional array of numbers (or samples). It is defined by the mathematical function f(x,y) where x and y are the two co-ordinates horizontally and vertically.

Resize speech: In this step-in order to visualize the change, we are going to create two functions to display the speeches the first being a one to display one speech and the second for two speeches. After that, we then create a function called processing that just receives the speeches as a parameter.

Need of resize speech during the pre-processing phase, some speeches captured by a camera and fed to our AI algorithm vary in size, therefore, we should establish a base size for all speeches fed into our AI algorithms.

3.3. MFCC Feature extraction

Pre-emphasis is the initial stage of extraction. It is the process of boosting the energy in high frequency. It is done because the spectrum for voice segments has more energy at lower frequencies than higher frequencies. This is called spectral tilt which is caused by the nature of the glottal pulse. Boosting high-frequency energy gives more info to Acoustic Model which improves phone recognition performance. MFCC can be extracted by following method.

1) The given speech signal is divided into frames (~20 ms). The length of time between successive frames is typically 5-10ms.
2) Hamming window is used to multiply the above frames to maintain the continuity of the signal. Application of hamming window avoids Gibbs phenomenon. Hamming window is multiplied to every frame of the signal to maintain the continuity in the start and stop point of frame and to avoid hasty changes at end point. Further, hamming window is applied to each frame to collect the closest frequency component together.
3) Mel spectrum is obtained by applying Mel-scale filter bank on DFT power spectrum. Mel-filter concentrates more on the significant part of the spectrum to get data values. Mel-filter bank is a series of triangular band pass filters similar to the human auditory system. The filter bank consists of overlapping filters. Each filter output is the sum of the energy of certain frequency bands. Higher sensitivity of the human ear to lower frequencies is modeled with this procedure. The energy within the frame is also an important feature to be obtained. Compute the logarithm of the square magnitude of the output of Mel-filter bank. Human response to signal level is logarithm. Humans are less sensitive to small changes in energy at high energy than small changes at low energy. Logarithm compresses dynamic range of values.
4) Mel-scaling and smoothing (pull to right). Mel scale is approximately linear below 1 kHz and logarithmic above 1 kHz.
5) Compute the logarithm of the square magnitude of the output of Mel filter bank.
6) DCT is further stage in MFCC which converts the frequency domain signal into time domain and minimizes the redundancy in data which may neglect the smaller temporal variations in the signal. Mel-cestrum is obtained by applying DCT on the logarithm of the

mel-spectrum. DCT is used to reduce the number of feature dimensions. It reduces spectral correlation between filter bank coefficients. Low dimensionality and 17 uncorrelated features are desirable for any statistical classifier. The cepstral coefficients do not capture the energy. So, it is necessary to add energy feature. Thus twelve (12) Mel Frequency Cepstral Coefficients plus one (1) energy coefficient are extracted. These thirteen (13) features are generally known as base features.

7) Obtain MFCC features.

The MFCC i.e. frequency transformed to the cepstral coefficients and the cepstral coefficients transformed to the MFCC by using the equation.

$$mel(f) = 2595 \times \log 10 \left(1 + \frac{f}{700}\right)$$

Where f denotes the frequency in Hz. The Step followed to compute MFCC. The MFCC features are estimated by using the following equation.

$$C_n = \sum_{n=1}^{K} (logS_k) \left[n\left(K - \frac{1}{2}\right)\frac{\pi}{K}\right] wheren = 1,2,\ldots..K$$

Here, K represents the number of Mel cepstral coefficient, C0 is left out of the DCT because it represents the mean value of the input speech signal which contains no significant speech related information. For each of the frames (approx. 20 ms) of speech that has overlapped, an acoustic vector consisting of MFCC is computed. This set of coefficients represents as well as recognize the characteristics of the speech.

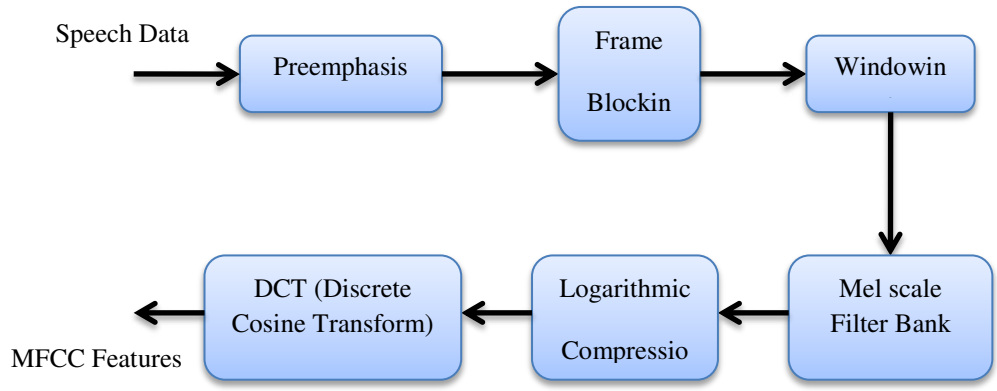

Fig.3. MFCC operation diagram

3.4 LSTM Multi-stage Classification

First, the vocal parts are detected. Then, a gender classifier determines the singer's gender. Finally, a singer classifier is used to identify the singer. Here, we use an LSTM recurrent neural network for vocal and singer classification stages. Figure 4 shows the proposed LSTM model for Signer-name specific training. Figure 5 shows the proposed LSTM model for signer-gender specific training. Figure 6 shows the overall testing process testing for singer-name and gender predication from test MFCC features.

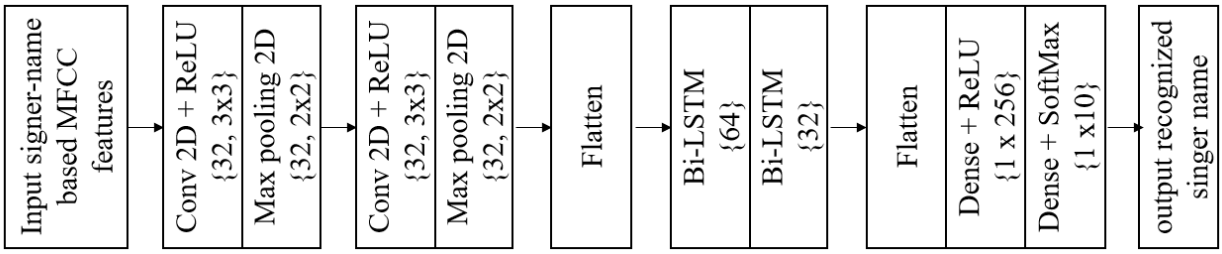

Fig.4. Signer-name specific training

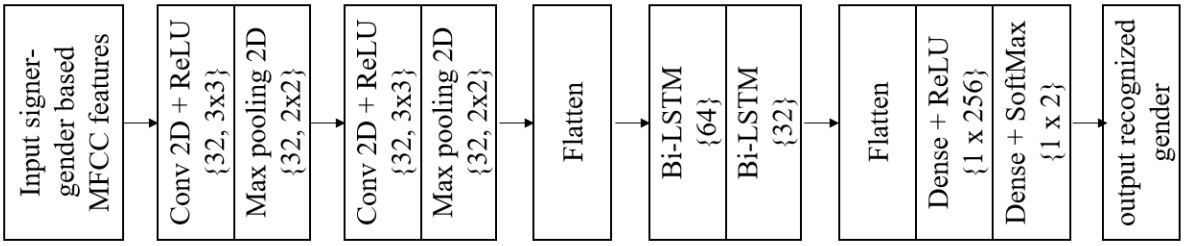

Fig.5. Singer-gender based training

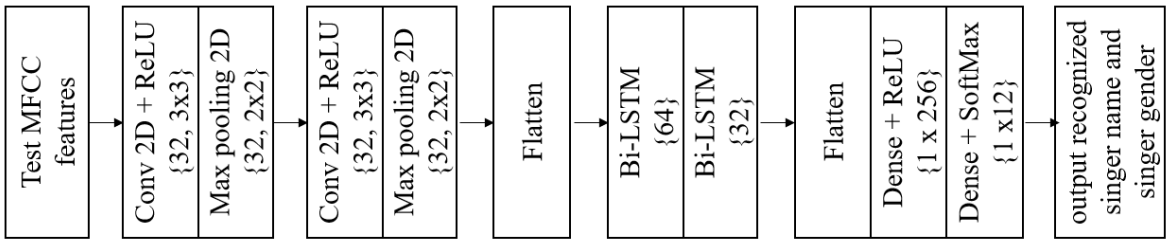

Fig.6. Overall testing for singer-name and gender predication

4. RESULT

Singer Identification by Vocal Parts Detection and Singer Classification Using LSTM Neural Networks In this paper author is using deep recurrent neural network algorithm called LSTM to predict gender by analyzing audio vocal part and to predict singer name. In propose work author is building two LSTM model where one is used to predict singer gender or gender identification and other is used to predict/classify singer name. To implement this project author is using MIR-1K dataset which contains signer's audio files and by using this dataset we will train all algorithms such as LSTM, SVM and MLP where LSTM is the propose work and SVM and MLP is the existing algorithms. Figure 7 compares performance of proposed LSTM comparison with two existing algorithms accuracy such as SVM and MLP. Table 1 shows the performance measures and Figure 8 shows the predicted outcome.

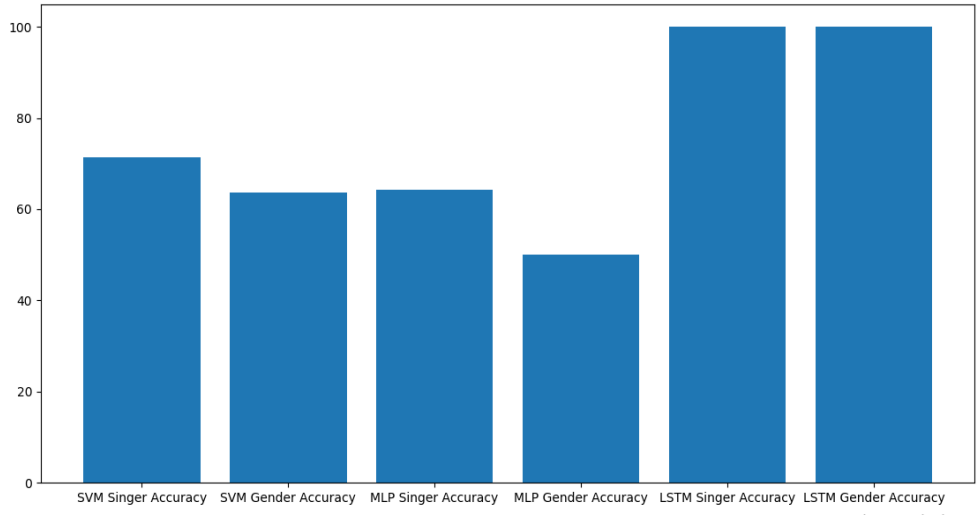

Fig.7. Accuracy comparison graph.

Table.1. Accuracy comparison.

Algorithms	singer	Gender
MLP	64.2	50.0
SVM	71.4	63.6
LSTM	100.0	100.0

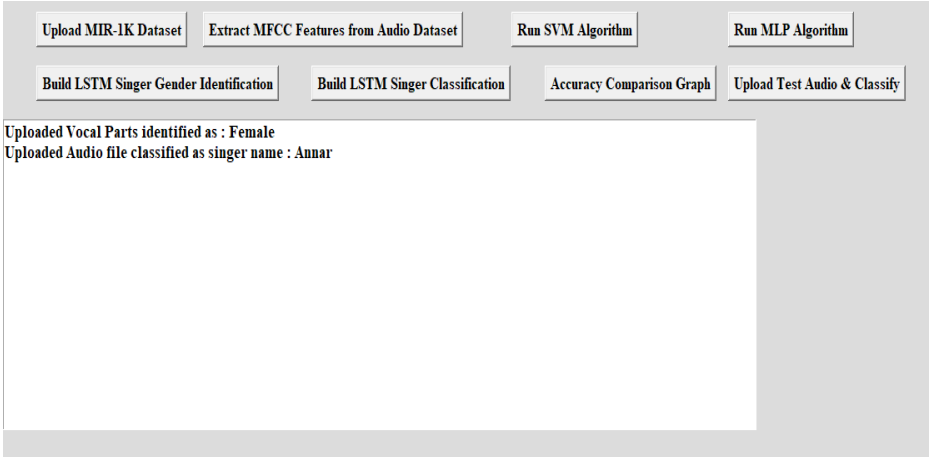

Fig.8. predicted outcome.

5. CONCLUSION

A supervised singer identification system has been proposed in this paper, which combines multi-stages of processing blocks, in which deep recurrent neural networks-based LSTM cells are mainly involved. The strategy adopted here, is to first detect the vocal parts and classify the gender of the singer, and then to apply for identifying the singer. Along with an appropriate feature vector, the achieved results indicate the efficacy of our strategy and the proposed system with respect to the state-of- the-art. For our future works, it would be interesting to determine the singer vocal types and incorporate the optimal feature vectors using a deep Auto-Encoder. Further, a measure of similarity

could be defined in order to detect similar styles among singers. Using further processing tuning steps, such as drop out and batch normalization will also increase the system overall accuracy, for sure.

REFERENCE

[1] Fu, Zhouyu, Guojun Lu, Kai Ming Ting, and Dengsheng Zhang. "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, Yannis Kopsinis, Nadine Kroher, and José- Miguel Díaz-Báñez. "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] Regnier, Lise, and Geoffroy Peeters. "Singer verification: singer model. vs. song model." In Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pp. 437-440. IEEE, 2012.

[7] 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.

[8] Hu, Ying, and Guizhong Liu. "Separation of singing voice using nonnegative matrix partial co-factorization for singer identification." IEEE Transactions on Audio, Speech, and Language Processing 23, no. 4 (2015): 643-653.

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

[10] Eronen, Antti, and Anssi Klapuri. "Musical instrument recognition using cepstral coefficients and temporal features." In Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on, vol. 2, pp. II753-II756. IEEE, 2000.

[11] Dixon, Simon. "Onset detection revisited." In Proceedings of the 9th International Conference on Digital Audio Effects, vol. 120, pp. 133-137. 2006.

[12] F. Gers, N. Schraudolph, and J. Schmidhuber, “Learning Precise Timing with LSTM Recurrent Networks,” Journal of Machine Learning Research, vol. 3, pp. 115–143, 2002.

[13] Graves, Alex, Navdeep Jaitly, and Abdel-rahman Mohamed. "Hybrid speech recognition with deep bidirectional LSTM." In Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on, pp. 273-278. IEEE, 2013.