Accepted Manuscript

International Journal of Pattern Recognition and Artificial Intelligence

Article Title:	Combined evidence of MFCC and CRP features using machine learning algorithms for singer identification
Author(s):	Sangeetha Rajesh, N. J. Nalini
DOI:	10.1142/S0218001421580015
Received:	06 June 2019
Accepted:	17 February 2020
To be cited as:	Sangeetha Rajesh, N. J. Nalini, Combined evidence of MFCC and CRP features using machine learning algorithms for singer identification, *International Journal of Pattern Recognition and Artificial Intelligence*, doi: 10.1142/S0218001421580015
Link to final version:	https://doi.org/10.1142/S0218001421580015

World Scientific
www.worldscientific.com

International Journal of Pattern Recognition
and Artificial Intelligence
Vol. XX, No. X (2006) 1–18
© World Scientific Publishing Company

Combined evidence of MFCC and CRP features using machine learning algorithms for singer identification

Sangeetha Rajesh*

K.J.Somaiya Institute of Management Studies and Research,
Mumbai, India
rajesh.sangeetha@gmail.com

N.J. Nalini

Annamalai University,
Annamalai Nagar, India
njncse78@gmail.com

Singer identification is a challenging task in music information retrieval because of the combined instrumental music with the singing voice. The previous approaches focus on identification of singers based on individual features extracted from the music clips. The objective of this work is to combine Mel Frequency Cepstral Coefficients (MFCC) and Chroma DCT-reduced Pitch (CRP) features for singer identification system (SID) using machine learning techniques. The proposed system has mainly two phases. In the feature extraction phase, MFCC, ΔMFCC, ΔΔMFCC and CRP features are extracted from the music clips. In the identification phase, extracted features are trained with Bidirectional Long Short-Term Memory (BLSTM) based Recurrent Neural Networks (RNN) and Convolution Neural Networks (CNN) and tested to identify different singer classes. The identification accuracy and Equal Error Rate (EER) are used as performance measures. Further, the experiments also demonstrate the effectiveness of score level fusion of MFCC and CRP feature in the singer identification system. Also, the experimental results are compared with the baseline system using support vector machines (SVM).

Keywords: Singer identification; Music information retrieval; Support vector machines; long short-term memory; recurrent neural networks; convolutional neural network; MFCC; CRP.

1. Introduction

The development of new technologies for storage and communication has led to the escalation of digital data volume, which also influenced the music industry. Music data stored on various digital repositories and cloud need to be organized and retrieved efficiently. The study on techniques for organizing and retrieving the substantial musical data is a fostering area of research in Music Information Retrieval (MIR). The indexing of music storage is based on various aspects such as genre, artist, instruments, and emotion. However, the singing voice is one of the essential music attributes which attracts the listener's attention. The music lovers are fascinated to listen to songs of their favourite singers. With the ingress of recommendation systems, streaming service providers recommend songs to the users based on their listening patterns. These services require a mechanism to identify and organize the songs to provide personalized recommendations. The proliferation of such systems demands the classification of music collection based on singers. Identifying the singers from the music clips is a challenging issue in MIR due to the following reasons.

- The instrumental music mixed with the singing voice acts as a noise in the process of identifying the singer.
- The type of instruments used to record affects the music data.
- The lack of standard data set available for singer identification of Indian language songs.

The singer identification process entails mainly two phases. i) Feature extraction and ii) identification. In the feature extraction phase, the acoustic features which accurately contribute to identifying the singers are extracted from the music clips. Most of the works in MIR focus on various audio features and its effectiveness in the information retrieval process. MFCC has been efficaciously utilized in most of the music and speech pattern recognition problems[23]. Speech recognition systems using MFCC features outperformed systems using other acoustic features due to its ability to represent the human auditory mechanism accurately [21]. However, in the music-related pattern recognition systems, the acoustic characteristics, such as harmony and pitch, also play an essential role. Chroma features represent the harmony aspect of the music signal[24]. In this work, the effectiveness of MFCC, its derivatives and CRP features for singer identification are analyzed.

In the identification phase, the extracted features are trained with deep learning algorithms. The selection of the machine learning algorithm that optimizes the performance of pattern recognition problems is also an exciting area of study. Traditional machine learning algorithms such as SVM, Decision Tree and K-Nearest Neighbors, have demonstrated their performance in MIR research[42]. However, SVM has revealed its superior performance over other machine learning algorithms[34]. Nonetheless, reinforcement learning has been widely used in various mobile applications; its invasion is less in music information processing[11, 12]. The acclimatization of deep learning techniques in various speech recognition and speaker verification systems has also influenced the optimization of MIR tasks. Nevertheless, the study of deep learning techniques to identify singers from the music clips needs to be reconnoitered.

The main objective of this work is to investigate the effectiveness of score level fusion of MFCC and its derivatives, ΔMFCC and ΔΔMFCC with CRP features in identifying singers from music clips using deep learning techniques, RNN and CNN. Since SVM has revealed its efficiency in MIR application, compared to Artificial Neural Networks, the results obtained using deep learning algorithms have been validated with the baseline machine learning technique SVM[8]. The significant contributions in this work are:

- The efficacy of MFCC and CRP features for singer identification is analyzed.
- The combined evidence of Mel Frequency Cepstral Coefficients (MFCC) and Chroma DCT-Reduced Pitch (CRP) features for singer identification is evaluated using deep learning algorithms, BLSTM-RNN and CNN.
- The results obtained are validated using baseline machine learning algorithm SVM.

2. Related Work

The research on various techniques for organizing and retrieving music data has gained more focus on rapid developments in the digital era. In this section, the various acoustic features and recognition techniques employed in research related to speaker verification and singer identification tasks are discussed. A suitable assortment of features plays a vital role in identification performance. The cepstral features of the speech signals are widely used in speaker verification systems[20]. A forensic speaker verification system is proposed by combining the MFCC and the discrete wavelet transform MFCC by Al-Ali et al. [.2] The work states, combining the features gives better verification performance even under noisy environments. The features of each frame of the music signal, which carries the essential information need to be captured for singer identification system. There are various techniques used for feature extraction from the speech or music signal such as MFCC, Linear Prediction Cepstral Coefficients, and Chroma features.

Hu and Liu proposed a singer identification system built on computational auditory scene analysis in which the missing features are analyzed and padded to improve the system[16]. The challenging problem of singer identification is the accompaniment of instrumental music. The authors first segregated the singing voice from the mixed signal. And then, Gammatone frequency cepstral coefficients (GMCC) are employed to achieve better performance. Fujihara et al.8 proposes a technique to solve the accompaniment impacts in the singer identification tasks. They employed Gaussian Mixture Models (GMM) to recognize the singers based on the extracted features. The work also demonstrates that higher accuracy can be achieved if the music data is of the same quality and with the same instruments. A distant speech recognition system using channel-wise convolution is proposed by Pawel et al. [.33]. A word error rate of 6.5% is achieved with AMI meeting corpus, which is comparatively better than the baseline deep neural network and Gaussian mixture models. The timbre invariance for instrumental music is also demonstrated with the use of CRP features for chord recognition[31].

The classification phase is the subsequent primary phase in the pattern recognition process. In this step, the extracted features are trained using the classifiers to generate the models. Various machine learning techniques have been proposed for music-related recognition tasks. The performance of the proposed systems varies based on the classification algorithm. Identifying the appropriate learning algorithm for various recognition applications is the main objective of research in this area. Fujimoto presented a factored deep convolution neural network for speech recognition[9]. The author has factored the feature enhancement, delta parameter tuning, and the HMM state classification with each factored method and the deep neural networks. The results indicate a substantial enhancement in noisy speech recognition. Ratanparna et al., anticipated a system to identify singers from the Indian video songs[37]. Naïve Bayes classifier and backpropagation neural networks are used to train the model with MFCC and LPC feature vectors. They demonstrated the importance of dimensionality reduction of the coefficients to improve the performance. MFCC is one of the prevalently used feature vectors for many speech and speaker recognition problems. It is also revealed its role in music-related recognition tasks. MFCC and Cepstral Mean Subtracted MFCCs are combined at score level, which has shown a substantial enhancement in the performance of the proposed system[32]. In another work, MFCC and residual phase features are used to recognize the emotion from the music. The authors employed SVM, AANN and RBFNN, among which SVM has shown the superior performance[29].

The findings and research gap apprehended from the literature review are listed below:

- The previous works on singer identification are based on the traditional machine learning techniques.
- Chroma features, which accurately epitomize the harmonic aspect of the music signal, are seldom used in the singer identification systems.
- The prior works mainly focus on individual features than the combination of features for capturing the essential information from the music signal for singer identification.

In this work, a singer identification system is proposed using MFCC, and CRP features with deep learning techniques, BLSTM-RNN and CNN. The effectiveness of derivatives of MFCC features, such as ΔMFCC and ΔΔMFCC, are also analyzed. The enhancement in the performance of the singer identification system is demonstrated by combining the MFCC and CRP features at the score level.

3. Proposed Singer Identification System

The process flow diagram of the proposed system is depicted in figure 1. Initially, the music data for 10 playback singers are collected from various online websites. The collected data is split into two sets of 70:30 for train and test sets respectively. Then, the acoustic features which accurately characterize the music data are extracted from the music clips. Various acoustic features such as MFCC, LPCC have been revealed their efficacy in speech recognition, speaker verification and speech emotion recognition systems[13,20, 35,38,41]. Nevertheless, most of the speech and music recognition tasks extensively employ the timbre features. MFCC is a common timbre feature which is analogous to human auditory mechanism[30]. CRP features enhance the degree of timbre invariance, which is a complement to the MFCC and are hardly used in recognition of speech and music.

In this work, MFCC and CRP features are engaged for the singer identification process. The derivatives of MFCC features (delta MFCC and delta-delta MFCC) are also extracted and analyzed to realize its efficacy in identifying the singers from the music signal. Then, the deep learning neural networks, BLSTM-RNN and CNN are trained with the extracted features from the training dataset. The trained models are evaluated using the extracted features from the test dataset. Finally, the proposed system is optimized by tuning the hyperparameters. The combined evidence of MFCC and CRP features at the score level illustrates a significant improvement in the performance of the singer identification system.

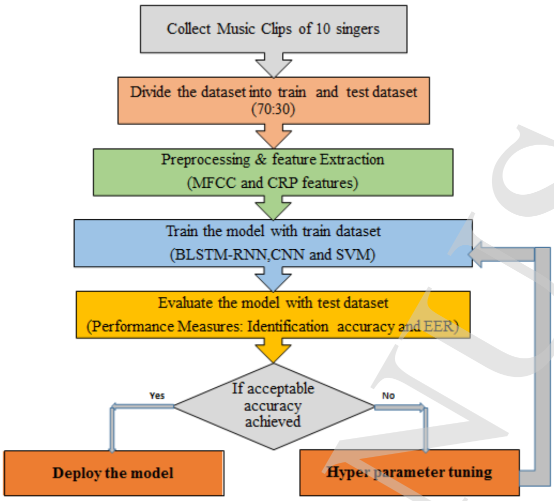

Collect Music Clips of 10 singers
Divide the dataset into train and test dataset (70:30)
Preprocessing & feature Extraction (MFCC and CRP features)
Train the model with train dataset (BLSTM-RNN,CNN and SVM)
Evaluate the model with test dataset (Performance Measures: Identification accuracy and EER)
If acceptable accuracy achieved
Yes
No
Deploy the model
Hyper parameter tuning

Fig.1. Process flow diagram of proposed singer identification system

3.1. *Data Organization*

The proposed work identifies ten different singer classes. Due to lack of standard dataset for identification of singers from Indian language music clips, music dataset used in this work is created using the songs collected from various websites. The dataset is a collection of songs by 10 playback singers. The collected songs are with 44.1 kHz sampling frequency and 16-bit wav format. Table 1 presents the details of singers. 200 music clips of 10 seconds duration are collected for each singer. A total of 2000 music clips which contains vocal of both male and female playback singers are used. From the collected data, 70% is engaged for training, and 30% is used for testing the model. The music signal is separated into frames of 20ms with 10ms shift.

Table 1. Details of Singer Dataset

Sl.No	Name of the singer	No. of Samples
1	Kumar Sanu	200
2	Atif Aslam	200
3	Hrohit Saboo	200
4	S.P. Balasubrahmaniam	200
5	K.J.Yesudas	200
6	Lata Mangeshkar	200
7	Sonu Nigam	200
8	Shreya Ghoshal	200
9	S.Janaki	200
10	K.S.Chithra	200

3.2. *Features employed in Singer Identification System*

3.2.1. *Mel Frequency Cepstral Coefficients (MFCC)*

MFCC is a powerful spectral feature in music and speech recognition applications. These features encode the timbre properties of the music signal and effectively represent the human auditory response[30]. Also, MFCC features are reliable for variations in speaker and recording conditions[10]. It is the representation of a music signal as a short-time power spectrum of a single frame based on a linear cosine transform of log power spectrum on a nonlinear Mel scale frequency[39]. Appending the information in the dynamics of the audio signal, known as delta coefficients (ΔMFCC) or differential coefficients improves the performance of the recognition systems. It is computed as given:

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \qquad (1)$$

where d_t is the delta coefficient, and c_t is static MFCC coefficient.

Delta-delta coefficients (ΔΔMFCC), also known as acceleration coefficients, are calculated using the same method using delta coefficients in place of static MFCC coefficients. In the proposed work, 13 MFCC(MFCC13) features, 13 ΔMFCC features and 13 ΔΔMFCC features are extracted using Python speech features package. ΔMFCC is appended with MFCC features which resulted in 26 MFCC feature values (MFCC26) and ΔΔMFCC is appended with MFCC26 to form 39 MFCC feature values (MFCC39). The extracted MFCC features from sample music clips of various singers are shown in figure 2.

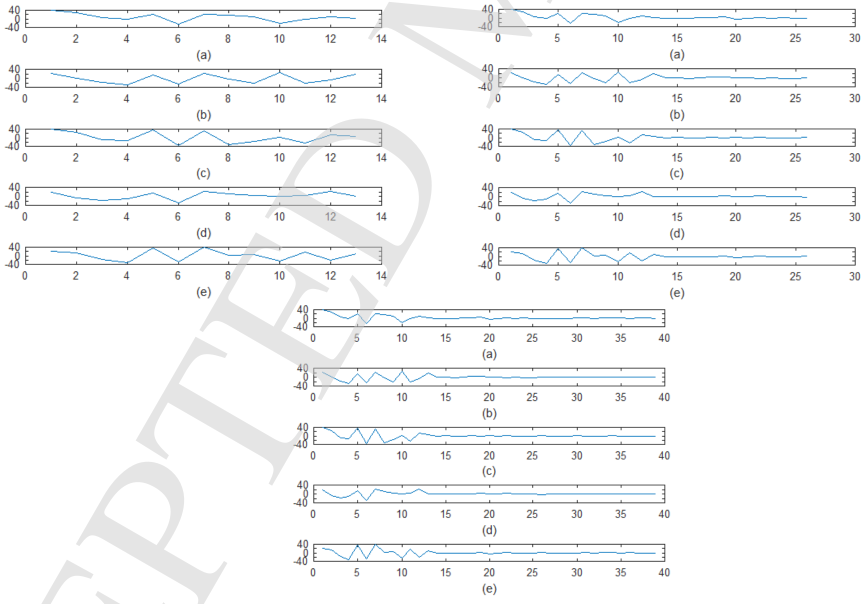


Fig. 2. Extracted MFCC13, MFCC26 and MFCC39 features from sample music clips of singers(a) - (d)

3.2.2. *Chroma DCT-reduced Pitch features*

Chroma features are introduced in order to extract timbre-invariance information from the music signal[26.] They are well suited for music data processing because of the high correlation to the harmony aspect of music. The overview of computing the CRP features is shown in figure 3[25-28]. It discards the timbre related information which is captured while extracting MFCC. CRP is computed by applying logarithmic compression to pitch features and transforming using DCT. From the resulting Pitch Frequency Cepstral Coefficients (PFCCs), discard the lower coefficients and apply inverse DCT to the upper coefficients. Finally, project the pitch vector to 12 bins which results in 12-dimensional chroma vectors. These vectors are normalized to compute CRP[31]. CRP features have few characteristics which make it different among all feature types. Primarily, they are intended to be invariant with timbre. Secondly, they integrate logarithmic compression in computing intensity. They also include smoothing technique. Figure 4 shows the CRP features extracted from a sample music clip.

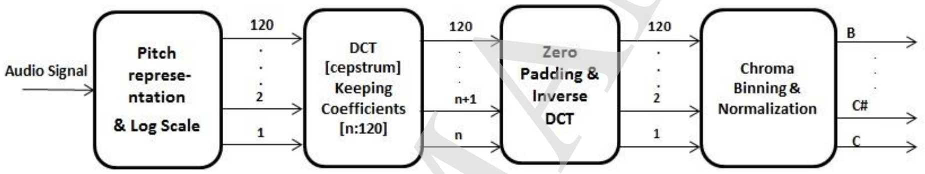

Audio Signal
Pitch representation & Log Scale
120
2
1
DCT [cepstrum] Keeping Coefficients [n:120]
120
n+1
n
Zero Padding & Inverse DCT
120
2
1
Chroma Binning & Normalization
B
C#
C

Fig.3. Chroma reduced pitch (CRP) features extraction process[26]

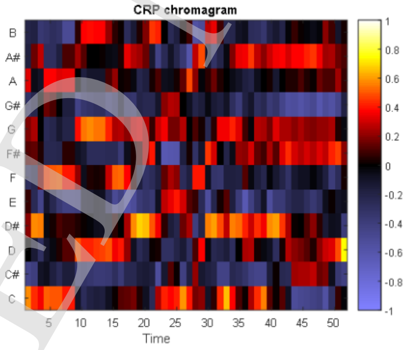


Fig.4. CRP features extracted from a sample singer music signal

3.3. *Techniques for Singer Identification*

3.3.1. *Bidirectional Long Short-Term Memory -Recurrent Neural Networks (BLSTM-RNN)*

Deep neural networks (DNN) have evidenced its superior performance over the conventional approaches for speech recognition[3]. The recurrent neural network is the only deep learning algorithm with internal memory which makes it appropriate to work on the sequential data such as weather, speech, and music[18]. RNN accepts the current input, and the learning from the previous input received. It uses the Backpropagation algorithm for training. Hochreiter et al. have introduced long short-term memory (LSTM) [15] to reduce the vanishing gradient problem of backpropagation neural networks (BPNN). The authors reveal LSTM as a suitable technique for pattern recognition from the sequential data. It also increases the recall capability of RNN. In LSTM, the information is flowing in a forward direction from time step t-1 to the time step t, which can handle only forward dependencies. Both forward and backward dependencies need to be captured in pattern recognition from sequential data. To fill this gap, Graves et al. introduced the Bidirectional Long Short-Term network that can handle bidirectional dependencies in the data[14]. An LSTM network figures a mapping from an input vector x= $(x_1,...,x_T)$ to an output vector y=$(y_1,...,y_T)$ by computing the network unit activations using the following equations iteratively from t= 1 to T:

$$i_t = \sigma(W_{ix}x_t + W_{im}m_{t-1} + W_{ic}c_{t-1} + b_i) \quad (2)$$

$$f_t = \sigma(W_{fx}x_t + W_{fm}m_{t-1} + W_{fc}c_{t-1} + b_f) \quad (3)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g(W_{cx}x_t + W_{cm}m_{t-1} + b_c) \quad (4)$$

$$o_t = \sigma(W_{ox}x_t + W_{om}m_{t-1} + W_{oc}c_{t-1} + b_o) \quad (5)$$

$$m_t = o_t \odot h(c_t) \quad (6)$$

$$y_t = \emptyset\left(W_{ym}m_t + b_m\right) \quad (7)$$

where

W: weight matrices,
b: bias vectors,
σ: logistic sigmoid function, and
i: input gate,
f: forget gate,
o: output gate,
c: cell activation vectors,
m: cell output activation vector,
⊙: element-wise product of the vectors,
g: cell input activation functions
h: cell output activation functions

3.3.2. *Convolutional Neural Networks (CNN)*

CNN is primarily applied for pattern recognition in image applications. It is motivated by the biological process; the neurons are connected in a way resembling the animal visual cortex pattern[43]. The network learns non-linear mappings from complex data which makes it a popular technique for pattern recognition. The major difference between CNN and multi-layer perceptron (MLP) is the structure of hidden layers. In CNN, hidden layers comprise of convolution, pooling and fully connected layers. In each CNN-layer, Conv1D forward propagation is articulated as follows:

$$x_k^l = b_k^l + \sum_{i=1}^{N_{l-1}} conv1D(w_{ik}^{l-1}, s_i^{l-1}) \quad (8)$$

x_k^l : input of the k^{th} neuron at level l
b_k^l : bias of the k^{th} neuron at level l

w_{ik}^{l-1}: kernel from the i^{th} neuron at layer $l{-}1$to the k^{th} neuron at layer l.
s_i^{l-1}: output of the i^{th} neuron at layer $l{-}1$

Another essential building block is the pooling layer or subsampling layer mainly introduced for parameter reduction[6]. Fully connected layers are intended for high-level reasoning. CNN has been recognized as a successful recognition technique in natural language processing, image and video recognition, recommender systems, and medical applications[17, 1]. However, its intervention in MIR applications is minimal.

3.3.3. *Support Vector Machines (SVM)*

SVM is a non-probabilistic classification algorithm[4]. It has been extensively used in various pattern recognition applications such as text classification[40], image classification[7], speech emotion recognition and speaker verification[5]. SVM performs the classification based on the concept of maximizing the margin or distance between two classes. Data is transformed into a high dimensional feature space to extend the application of SVM in nonlinear decision support, as shown in figure 5.

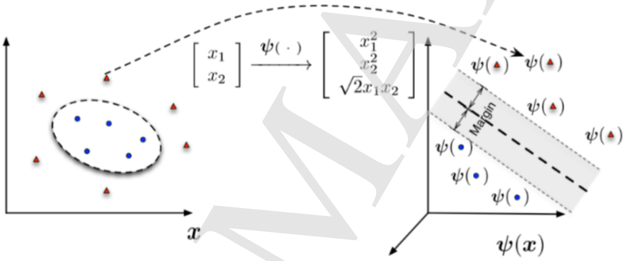

$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$
$\psi(\cdot)$
$\begin{bmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \end{bmatrix}$
Margin
$\boldsymbol{x}$
$\boldsymbol{\psi(x)}$

Fig.5. Feature mapping to high dimension feature space in SVM

SVM uses various kernels to compute the high dimensional feature space. The kernel function is given by:

$$k(O_i, O_j) = \Psi(O_i)^T \Psi(O_j) \tag{8}$$

where O_i is the supervised training samples.
The classification is represented as given:

$$y = sgn(\textstyle\sum_{i=1}^{R} \propto_i^{dual} y_i \; k(O, O_j) + b \tag{9}$$

where α is the weight vector and b is bias.

3.4. *Performance Metrics*

In this work, the performance of different approaches is evaluated by using mainly two performance measures, identification accuracy and Equal Error Rate (EER). Identification accuracy is the correctly predicted testing samples among the total number of the testing samples.

$$Identification\ Accuracy = \frac{number\ of\ \ correct\ predictions}{Total\ number\ of\ predictions} \tag{10}$$

EER is the result obtained by adjusting the system threshold such that the false acceptance rate (FAR) and false rejection rate (FRR) are equal.

$$FAR = \frac{False\ Positives}{False\ Positives + True\ Negatives} \quad (11)$$

$$FRR = \frac{False\ Negatives}{True\ Positives + False\ Negatives} \quad (12)$$

4. Experimental results and analysis

The experiments are conducted using Python 3.7 on a system with 16GB RAM and 4GB Graphics Card. The deep neural networks BLSTM-RNN and CNN are implemented using high-level neural network API Keras with Tensorflow as backend. The python library, Scikit-learn is used as a tool to perform the necessary engineering tasks.

From the collected data, 70% is randomly selected for training, and the remaining 30% is used to test the trained models. In this work, the performance of MFCC, CRP and the combined MFCC and CRP features for singer identification is evaluated using CNN and BLSTM-RNN. In addition to 13-dimensional MFCC features, the efficacy of delta MFCC (MFCC26) and delta-delta MFCC (MFCC39) features are also analyzed to identify singers from the music clips. The performance of the proposed singer identification system is evaluated using identification accuracy and EER. The results are compared with the baseline machine learning technique SVM and the achieved accuracy, and EER is consolidated in Table 6.

4.1. *Singer identification using BLSTM-RNN*

During the training phase, BLSTM-RNN is trained for each singer. The MFCC and CRP features extracted from the music signal are fed to the recurrent neural network's input layer. The BLSTM-RNN is fabricated with 3 LSTM hidden layers. The first hidden layer used in the network is bidirectional LSTM with 45 units which can handle the forward and backward information flow in the network. The subsequent two layers are LSTM layers with 30 and 20 nodes and ReLU activation function. Rectified Linear Units (ReLU) activation function overcomes the vanishing gradient problem, which helps the network to learn faster and perform better[22]. Drop out of 20%, 40%, and 50% are added in the network to avoid the overfitting. It drops out the neurons from the network by passing zero weights to the neurons. The output layer is a dense layer with 10 units to represents 10 different singer classes. 'adam' optimizer is used to update the network weights iteratively[19, 36]. The categorical cross-entropy loss function used is as singer identification is multiclass pattern recognition task. 'softmax' activation is used in the output layer to output the probability with each singer class. The architecture of BLSTM-RNN designed for the proposed singer identification system is shown in figure 6. Based on various experiments conducted, the proposed architecture gives the optimum performance for singer identification. The SID performance with MFCC39 and CRP using BLSTM-RNN features is described in the following sections.

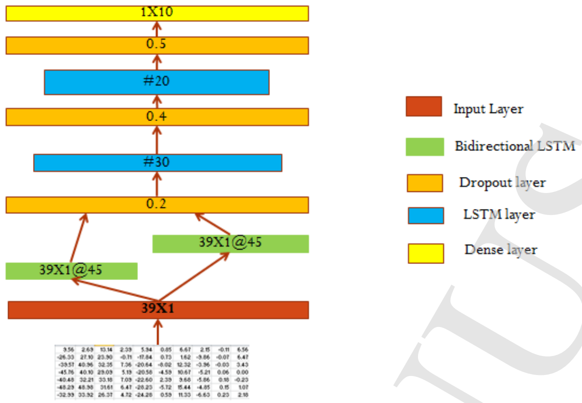


Fig. 6. Recurrent neural network architecture for the singer identification system

4.1.1. *SID using MFCC*

During the testing phase, MFCC features are extracted from the testing dataset and evaluated against BLSTM-RNN model. Also, the performance with variations in derivatives of MFCC coefficients is also analyzed. The identification rate of 92.0%, 96.0% and 96.0% and EER of 8.0%, 4.0% and 4.0% is obtained with MFCC13, MFCC26 and MFCC39 features respectively. Identification rate is computed using the confusion matrix. Table 2 shows the confusion matrix of singer identification using MFCC39 with BLSTM-RNN.

Table 2. SID performance using MFCC39 features with BLSTM-RNN

	Sn1	Sn2	Sn3	Sn4	Sn5	Sn6	Sn7	Sn8	Sn9	Sn10
Sn1	94.0	1.0	1.5	1.5	0.0	0.0	1.2	0.0	0.8	0.0
Sn2	2.5	93.1	0.0	1.0	1.4	0.0	2.0	0.0	0.0	0.0
Sn3	0.5	1.2	96.5	1.2	0.6	0.0	0.0	0.0	0.0	0.0
Sn4	2.3	1.7	1.5	92.0	1.0	0.0	1.5	0.0	0.0	0.0
Sn5	0.5	0.0	0.5	1.0	98.0	0.0	0.0	0.0	0.0	0.0
Sn6	0.0	0.0	0.0	0.0	0.0	97.5	0.0	1.0	0.0	1.5
Sn7	1.5	0.0	1.9	1.0	0.9	0.0	94.7	0.0	0.0	0.0
Sn8	0.0	0.0	0.0	0.0	0.0	0.7	0.0	98.2	0.0	1.1
Sn9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	99.0	1.0
Sn10	0.0	0.0	0.0	0.0	0.0	1.2	0.0	1.0	0.8	97.0
Overall performance = 96.0%										

4.1.2. *SID using CRP*

Singer identification system is also evaluated using 12-dimensional CRP features extracted from the testing music clips. Table 3 shows the confusion matrix of singer identification using CRP with BLSTM-RNN. The overall identification accuracy of 42.0% is obtained and EER of 30.0% is achieved from FAR and FRR curves.

Table 3. SID performance using CRP features with BLSTM-RNN

	Sn1	Sn2	Sn3	Sn4	Sn5	Sn6	Sn7	Sn8	Sn9	Sn10
Sn1	60.0	0.0	13.5	0.0	10.5	10.3	0.0	5.7	0.0	0.0
Sn2	30.0	40.0	6.5	0.0	0.0	0.0	15.0	5.0	0.0	3.5
Sn3	18.5	12.3	52.0	0.0	10.2	0.0	6.0	0.0	0.0	1.0

Sn4	22.8	13.3	5.0	30.0	0.0	6.4	22.5	0.0	0.0	0.0
Sn5	17.6	6.7	6.0	4.7	40.0	10.1	7.8	0.0	2.5	4.6
Sn6	3.5	0.0	4.1	2.0	2.0	40.0	0.0	18.2	6.7	23.5
Sn7	27.5	10.3	2.0	8.4	13.2	0.0	35.0	0.0	3.6	0.0
Sn8	3.6	0.0	1.6	0.0	20.8	4.2	0.0	51.0	13.2	5.6
Sn9	3.5	5.1	4.2	2.7	10.5	8.3	0.0	5.7	40.0	20.0
Sn10	10.5	14.2	6.5	8.6	12.5	12.5	0.0	2.2	0.0	32.0
Overall performance = 42.0%										

4.2. *Singer identification using CNN*

In this work, 1D CNN for singer identification system is designed which comprises of two convolution layers with kernel size 3 and 120 filters with ReLU(Rectified Linear Units) activation, one pooling layer with pool size 2 to perform the max pooling and a dense layer with 100 units. CNN is trained with the MFCC and CRP features extracted from the training music clips. The output layer is a dense layer with softmax activation, which results in the probability of each singer class. The hyperparameters are selected based on an empirical study performed with different values. Figure 7 shows the architecture of CNN for singer identification. The following sections describe the evaluation of singer identification using MFCC and CRP with CNN.

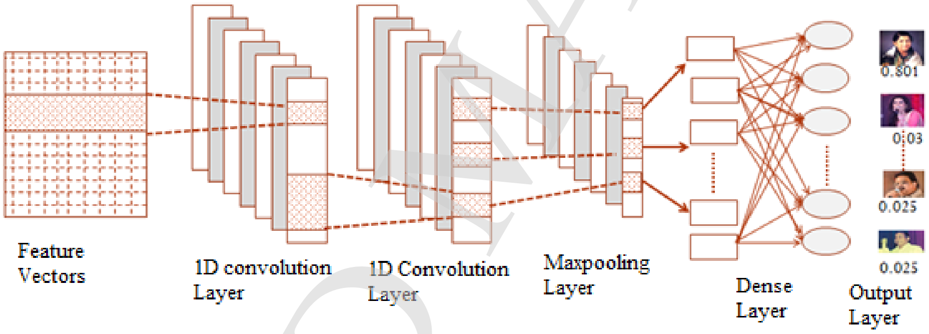


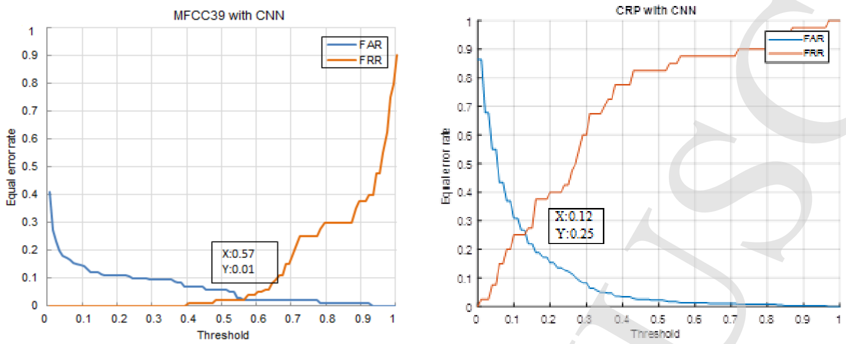
Fig.7. The architecture of 1D CNN designed for singer identification

4.2.1. *SID using MFCC*

The extracted MFCC features from the test music samples are fed as input to the trained CNN model. An identification rate of 90.0%, 92.0% and 98.0% is reported when tested against the CNN model using MFCC13, MFCC26 and MFCC39 features respectively. By evaluating the FAR and FRR curves, an EER of 8.0%, 8.0%, and 2.0% is obtained at a threshold of 0.4, 0.8, 0.5 respectively. FAR and FRR curves using MFCC39 features with CNN is shown in figure 8(a).

4.2.2. *SID using CRP*

The performance of singer identification using CNN is also evaluated using CRP features. The features extracted from test music clips of 10 singers are fed to trained CNN model. Overall identification accuracy of 48.0% and an EER of 25.0% are obtained using CRP features with CNN. EER using CRP features with CNN is shown in figure 8(b).



(a) (b)

Fig. 8. EER with CNN using (a) MFCC39 features (b) CRP features

4.3. *Singer identification using SVM*

SVM is trained with the extracted MFCC and CRP features from the training music clips. It distinguishes the features of one singer with a positive class label and all other features in the training set as a negative class label. The extracted MFCC and CRP features are given as input to the SVM. For each singer, a single SVM model is created. Totally 10 SVM models are created. 'RBF' kernel has been used in the proposed SVM model in order to adapt the non-linearity in the singer identification system.

4.3.1. *SID using MFCC*

During the testing phase, the SVM model is evaluated with the extracted MFCC features from the test singer music clips. An identification accuracy of 86.0%, 90.0% and 92.0% and EER of 14.0%, 10.0% and 8.0% is achieved with MFCC13, MFCC26 and MFCC39 respectively.

4.3.2. *SID using CRP*

The CRP features extracted from the test music clips are used to evaluate the SVM model. Identification accuracy of 44.0% is conveyed, and by varying the threshold, an EER of 30.0% is obtained.

4.4. *The combined evidence of MFCC and CRP features*

MFCC features provide the timbre information and CRP features captures the timbre invariance in the music signal. This complementary nature of the spectral feature MFCC is combined with Chroma feature CRP at the score level using the following equation.

$$c = ws1 + (1 - w)s2 \qquad (4)$$

where s1 and s2 are the scores of MFCC and CRP and w is the weight assigned to score in the range 0 to 1.

4.4.1. *Combined MFCC39 and CRP features with BLSTM-RNN*

By combining MFCC features and CRP features with BLSTM-RNN, the identification rate is improved when compared to the individual features. An accuracy of 94.0%, 97.0%, and 98.0% and EER of 6.0%, 3.0% and 2.0% is attained with MFCC13+CRP, MFCC26+CRP and MFCC39+CRP respectively. The experimental results show a significant increase in performance.

4.4.2. *Combined MFCC39 and CRP features with CNN*

By evaluating the FAR and FRR with varying threshold values, EER of 4.0%, 6.0%, and 1.0% is obtained with MFCC13+CRP, MFCC26+CRP and MFCC39+CRP respectively. Identification accuracy is computed using the confusion matrix. Table 4 shows the confusion matrix of singer identification using MFCC39+CRP with CNN. The overall identification rate of 96.0%, 94.0%, and 99.0% is achieved by the combination of MFCC13, MFCC26 and MFCC39 with CRP feature at score level.

4.4.3. *Combined MFCC39 and CRP features with SVM*

By evaluating SVM model, it is observed that EER of 12.0%,8.0% and 6.0% is obtained for the weight 0.6, and identification rate of 88.0%,92.0% and 94.0% is obtained using MFCC13+CRP, MFCC26+CRP and MFCC39+CRP respectively.

Table 4. SID performance using MFCC39+CRP features with CNN

	Sn1	Sn2	Sn3	Sn4	Sn5	Sn6	Sn7	Sn8	Sn9	Sn10
Sn1	99.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
Sn2	0.0	99.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
Sn3	1.0	0.0	98.5	0.0	0.0	0.5	0.0	0.0	0.0	0.0
Sn4	1.0	1.5	0.0	97.5	0.0	0.0	0.0	0.0	0.0	0.0
Sn5	0.0	0.0	0.0	0.0	99.5	0.5	0.0	0.0	0.0	0.0
Sn6	0.0	0.0	0.0	0.0	0.0	99.3	0.0	0.7	0.0	0.0
Sn7	0.0	0.2	0.0	0.0	0.0	0.0	99.2	0.0	0.6	0.0
Sn8	0.0	0.0	0.0	0.0	0.0	0.0	0.5	99.5	0.0	0.0
Sn9	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.0	98.9	0.7
Sn10	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0	99.6
Overall performance = 99.0%										

The experimental results exhibit the optimum performance of singer identification using MFCC39 with CRP features using CNN with overall accuracy of 99.0% and EER of 1.0%. The CRP features represent the harmony aspect which combined with 39-dimensional MFCC features contains the desired information for singer identification. Also, the complementary nature of MFCC and CRP features resulted in the escalation of the performance of the system when compared to the SID system using individual features. A comparison of SID performance with CNN, BLSTM-RNN, and SVM is shown in figure 9. The experiment results also demonstrate the efficiency of deep learning techniques over traditional machine learning algorithms when the data availability is high. Due to the lack of availability of standard music dataset of Indian playback singers, the proposed system is validated only on the basis of machine learning algorithm.

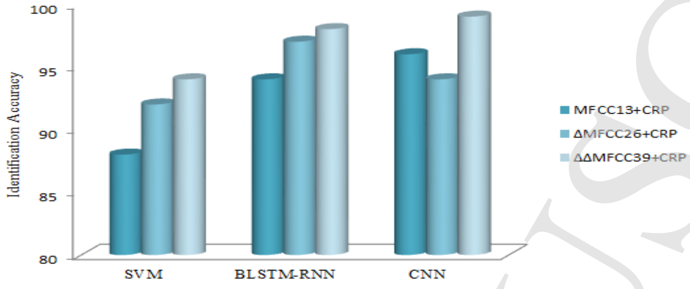


Fig.9. Comparison of performance of singer identification models

The analysis of the performance of singer identification for each singer class using various models is shown in Table 5. It clearly depicts that the combined evidence of MFCC39 and CRP features with CNN gives the highest accuracy for each singer class when compared to the baseline system.

Table 5. Singer identification performance of each singer class using SVM, BLSTM-RNN, and CNN

	SVM			BLSTM-RNN			CNN		
Singer Id	MFCC13 +CRP	MFCC26 +CRP	MFCC39 +CRP	MFCC13 +CRP	MFCC26 +CRP	MFCC39 +CRP	MFCC13 +CRP	MFCC26 +CRP	MFCC39 +CRP
S1	70.0	72.7	82.6	98.1	99.1	99.0	98.3	76.2	**99.0**
S2	73.2	80.0	83.2	80.0	99.2	99.2	84.2	90.2	**99.0**
S3	94.7	97.0	98.5	98.5	98.5	99.1	99.1	98.1	**98.5**
S4	65.9	79.5	82.4	70.0	80.4	92.0	84.8	81.0	**97.5**
S5	95.3	98.2	98.7	98.4	98.3	99.0	99.3	99.3	**99.5**
S6	95.2	98.5	99.1	98.7	99.4	98.2	98.7	98.4	**99.3**
S7	96.1	98.4	99.0	99.3	98.4	99.0	98.0	98.7	**99.2**
S8	96.3	97.8	98.0	99.3	98.1	98.5	99.5	99.1	**99.5**
S9	96.1	98.6	99.0	98.6	99.2	98.0	99.1	99.4	**98.9**
S10	97.2	99.3	99.5	99.1	99.4	98.0	99.0	99.6	**99.6**

Table 6. Singer Identification performance (%) using SVM, BLSTM-RNN and CNN

Acoustic Features	Performance measure	SVM	RNN	CNN
MFCC13	Accuracy	86.0	92.0	90.0
	EER	14.0	8.0	8.0
ΔMFCC26	Accuracy	90.0	96.0	92.0
	EER	10.0	4.0	8.0
ΔΔMFCC39	Accuracy	92.0	96.0	98.0
	EER	8.0	4.0	2.0
CRP	Accuracy	44.0	42.0	48.0
	EER	30.0	30.0	25.0
MFCC13+CRP	Accuracy	88.0	94.0	96.0
	EER	12.0	6.0	4.0
ΔMFCC26+CRP	Accuracy	92.0	97.0	94.0
	EER	8.0	3.0	6.0
ΔΔMFCC39+CRP	Accuracy	**94.0**	**98.0**	**99.0**
	EER	6.0	2.0	1.0

5. Conclusion

In this work, supervised deep learning techniques for singer identification from music signals is proposed. The experimental results demonstrate that BLSTM-RNN and CNN work well for singer identification task. The proposed work also reveals that appending delta and delta-delta features to static MFCC features improve the performance of the system. Further, combining MFCC and CRP features at score level increased the singer identification accuracy, due to the complementary timbre invariance information present in the CRP feature. Using MFCC39 and CRP features with BLSTM-RNN resulted in an accuracy of 98.0%, which demonstrate that BLSTM-RNN is capable of handling sequential data and works well for music information retrieval. However, the CNN model for singer identification obtained an accuracy of 99.0%. The proficiency of CNN in capturing the high dimensional information from the music data is also depicted in this work. In future work, unsupervised deep learning techniques which focus on the similarity information of different classes from the unknown music data need to be explored.

References

1. Acharya U. R., Oh S. L., Hagiwara Y., Tan J. H. and Adeli H., Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals, *Computers in Biology and Medicine*, **100** (2018) 270–278.
2. A. K. H. Al-Ali, David D., Bouchra S., Vinod C. and Ganesh R. N., Enhanced forensic speaker verification using a combination of DWT and MFCC feature wrapping in the presence of noise and reverberation conditions, *IEEE Access* **5** (2017) 15400-15413.
3. Alex G., Jaitly N. and Mohamed A. R., Hybrid speech recognition with deep bidirectional LSTM, in *Automatic Speech Recognition and Understanding* (2013) 273–278.
4. Asa B., David H., Hava S. and Vapnik V. N., Support vector clustering, *Journal of Machine Learning Research*, **2** (2001) 125-137.
5. Campbell W. M., Campbell J. P., Gleason T. P., Reynolds D. A. and Shen W., Speaker Verification Using Support Vector Machines and High-Level Features, *IEEE Transactions on Audio, Speech and Language Processing*, **15(7)** (2007) 2085–2094.
6. Chen, X., Kopsaftopoulos, F., Wu, Q., Ren, H., & Chang, F.-K., A Self-Adaptive 1D Convolutional Neural Network for Flight-State Identification, *Sensors*, **19(2)** (2019) 275.
7. Foody G. M. and Mathur A., Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification, *Remote Sensing of Environment* **93(1-2)** (2004) 107–117.
8. Fujihara H., Goto M. and Kitahara T., A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity based music information retrieval, *IEEE transactions on audio, speech and language processing,* **18** (2010) 638-648.
9. Fujimoto M., Factored deep convolution networks for noise robust speech recognition, in *Interspeech* (2017).
10. Furui, Speaker-independent isolated word recognition based on emphasized spectral dynamics, in *IEEE International Conference on Acoustics, Speech and Signal Processing* (1986).
11. Gai, K., & Qiu, M., Reinforcement Learning-based Content-Centric Services in Mobile Sensing, *IEEE Network*, **32(4)** (2018) 34–39.
12. Gai, K., & Qiu, M., Optimal resource allocation using reinforcement learning for IoT content-centric services. *Applied Soft Computing,* **70** (2018) 12–21.

13. Ge Z., Iyer A. N., Cheluvarja S., Sundaram R., Ganapathiraju A., Neural network classification and verification systems with enhanced features, in *Intelligent systems conference* (2017).
14. Graves A. and Mohamed A. R. and Hinton G., Speech recognition with deep recurrent neural networks, *Acoustics, speech and signal processing* (2013) 6645 – 6649.
15. Hochreiter S. and Schmidhuber J., Long short-term memory, *Neural Computation* **9(8)** (1997) 1735-1780.
16. Hu Y. and Liu G., Singer identification based on computational auditory scene analysis and missing feature methods, *Journal of intelligent information systems*, **42** (2014) 333-352.
17. Ji S., Xu W., Yang M. and Yu K., 3D Convolutional Neural Networks for Human Action Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35(1)** (2013) 221–231.
18. Jozefowicz R., Zaremba W. and Sutskever I., An empirical exploration of recurrent network architectures, in *International Conference on Machine Learning*, (2015) 2342 – 2350.
19. Kingma D. P. and Ba J. A., A Method for Stochastic Optimization, arXiv: 1412.6980, December 2014.
20. Kinnunen T. and H. Li, An overview of text-independent speaker verification: From features to supervectors, *Speech communication* **52** (2010) 2-40.
21. Kumar P. P, Vardhan K. S. N., Krishna K. S. R. Performance evaluation of MLP for speech recognition in noisy environments using MFCC & wavelets, *International Journal of Computer Science & Communication* **1(2)** (2010) 41-45
22. Machine Learning Mastery https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/. Accessed 21 May 2019.
23. Mokhsin M. B., Rosli N. B., Zambri S., Ahmad N. D., and Rahah S., Automatic music emotion classification using artificial neural network based on vocal and instrumental sound timbres, *Journal of Computer Science* **10(12)** (2014) 2584–2592.
24. M ̈uller M., Information Retrieval for Music and Motion. *Springer Verlag*, 2007
25. Müller M., Kurth F. and Clausen M., Audio matching via chroma-based statistical features, in *International Conference for Music Information Retrieval Conference*, (2005).
26. M ̈uller M. and Ewert S., Chroma Toolbox: MATLAB implementations for extracting variants of chroma based audio features, in *International society for music information retrieval* (2011).
27. Müller M. and Ewert S., Towards timbre-invariant audio features for harmony-based music, *IEEE Transactions on Audio, Speech, and Language Processing*, **18(3)** (2010) 649–662.
28. Müller M., Ewert S. and Kreuzer S., Making chroma features more robust to timbre changes. In *IEEE International Conference on Acoustics, Speech, and Signal Processing* (2009) 1869-1872.
29. Nalini N. J., Palanivel S., Music emotion recognition: combined evidence of MFCC and residual phase, *Egyptian informatics journal*, **17** (2016) 1-10.
30. Nalini N. J., Palanivel S., Balasubramanian M., Speech Emotion Recognition Using Residual Phase and MFCC Feature, *International Journal of Engineering and Technology,* **5(6)** (2014) 4515-4527.
31. O'Hanlon K, Ewert S, Pauwels J. and Sandler M. B., Improved template based chord recognition using the CRP feature, in *IEEE International Conference on Acoustics, Speech and Signal Processing* (2017).
32. . Patil H. A. and Radadia P. G., Combining evidences from Mel Cepstral features and Cepstral Mean Subtracted features for singer identification. in *International conference on Asian language processing*, (2012).
33. Pawel S., Arnab G. and Renals S., Convolutional neural networks for distant speech recognition, *IEEE signal processing letters* **21** (2014) 1120-1124.
34. Pouyanfar, S., & Sameti, H., Music emotion recognition using two level classification, In *Iranian Conference on Intelligent Systems*, (2014).

35. Prem S., ZaJar R., Cover song identification with 2D Fourier transform sequences, in *IEEE International Conference on Acoustics, Speech and Signal Processing*, (2017) .
36. Rajesh K. A., Improving Hindi Speech Recognition Using Filter Bank Optimization and Acoustic Model Refinement, Ph.D thesis, National Institute of Technology, Kurukshetra, December, 2012
37. Ratanparna T. and Patel N., Singer Identification using MFCC and LPC coefficients from Indian video songs, *Advances in intelligent systems and computing*, (2015) 275-282.
38. Regnier L., Peters G., Singer verification: Singer Model Vs. Song Model, in *IEEE International conference on acoustics, speech and signal processing*, (2012) 437-440.
39. Sangeetha R. and Nalini N. J., Singer recognition using MFCC and CRP features with support vector machines. *Computational Intelligence in Pattern Recognition, Advances in Intelligent Systems and Computing*, (2019) 295-306.
40. Sun A., Lim E. P. and Liu Y., On strategies for imbalanced text classification using SVM: A comparative study, *Decision Support Systems*, **48(1)** (2009) 191-201.
41. Tsai W. H., Lee H. C., Singer Identification based on spoken data in voice characterization. *IEEE Transactions on Acoustics Speech and Signal Processing*, **20(80)** (2012) 2291-2300.
42. Valverde-Rebaza, J., Soriano, A., Berton, L., Oliveira, M. C. F. de, & Lopes, A. de A., Music Genre Classification Using Traditional and Relational Approaches, *Brazilian Conference on Intelligent Systems* (2014).
43. Wikipedia https://en.wikipedia.org/wiki/Convolutional_neural_network. Accessed 26 November 2018.
44. Youngmoo E. K., Erik M. S., Music Emotion Recognition: A State of The Art Review, In *International Society for Music Information Retrieval Conference*, (2010) 255-266.