Indonesian Journal of Electrical Engineering and Computer Science
Vol. 24, No. 3, December 2021, pp. 1481~1490
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v24.i3.pp1481-1490

Audio classification for music information retrieval of Hindustani vocal music

Amit Rege[1], Ravi Sindal[2]
[1]Department of Electronics Engineering, Medi-Caps University, Indore, India
[2] The Institute of Engineering and Technology, Devi Ahilya University, Indore, India

Article Info

Article history:

Received Mar 17, 2021
Revised Oct 13, 2021
Accepted Oct 18, 2021

Keywords:

Audio classification
Music information retrieval
Vocal music

ABSTRACT

An important task in music information retrieval of Indian art music is the recognition of the larger musicological frameworks, called ragas, on which the performances are based. Ragas are characterized by prominent musical notes, motifs, general sequences of notes used and embellishments improvised by the performers. In this work we propose a convolutional neural network-based model to work on the mel-spectrograms for classication of steady note regions and note transition regions in vocal melodies which can be used for finding prominent musical notes. It is demonstrated that, good classification accuracy is obtained using the proposed model.

Corresponding Author:

Amit Rege
Department of Electronics Engineering
Medi-Caps University
A B Road, Pigdamber, Rau, Indore-453331, India
Email: amit.rege@medicaps.ac.in

1. INTRODUCTION

The objective of this study is to find steady note regions from a vocal recital for the purpose of finding prominent notes used by the vocalists. The music that we take here for analysis contains audio files extracted from vocal Hindustani music without any accompaniment in the background. Music in the theoretical sense is believed to include all three of lyrics, singing/instrument-playing and dance, contained in any artistic performance, at least in Indian tradition [1]. However, the poetry and the dance are considered separate art forms. Generally, singing and instrument-playing are considered music as these require similar artistic capabilities. We use the word music, in this paper in the latter sense. Music information retrieval is a field of study which deals with finding out music related information like melody, chord, rhythm, and instrument(s) from a particular representation of music, say audio data, symbolic representation, and specific music file formats like musical instrument digital interface (MIDI) [2].

Musical notes: musical notes are basic entities which have a defined pitch frequency in a musical scale confined within an octave, (*saptak*). The vocal or instrumental audio of the notes have a quasi-periodic time domain description which amounts to an almost harmonic description in the frequency domain with little inharmonic components. These notes have been given different names in different traditions. In Indian tradition they are generally referred to as *sa, re, ga, and ma*. Figure 1 shows time domain description of the sound of a vocal musical note.

The melody (harmony) in music consists of a gradual (simultaneous) improvisation of musical notes by the musician or performer. It is obvious that harmony is not possible if there is only one singer present, because human voice cannot create two different musical notes simultaneously, being monophonic. Figure 2 shows the pitch frequency of vocal recitals as a function of time for the duration where the vocalist recites a note transition.

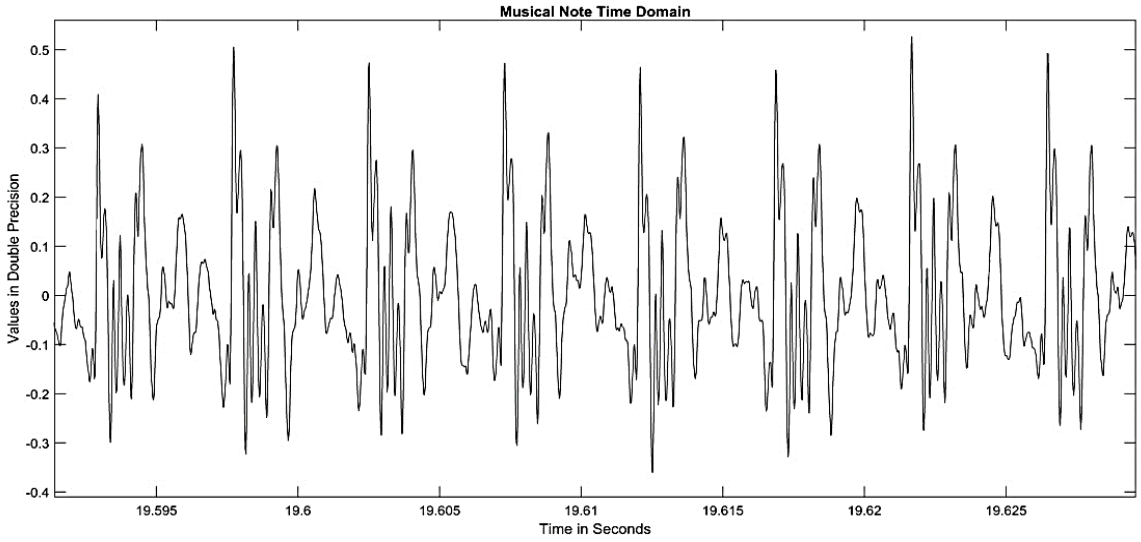

Figure 1. Time domain description of a typical musical note

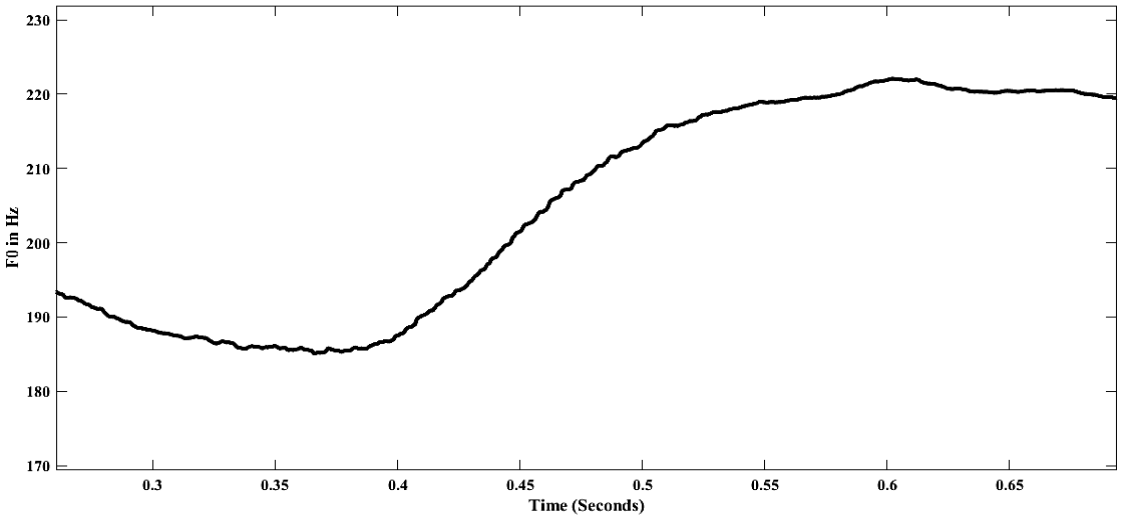

Figure 2. Change in pitch frequency of a vocal signal computed by Yin algorithm

Indian art music: Indian art music largely consists of monodies based on *ragas* which are well-defined musicological frameworks [3]. Ragas are characterized by the defines combination of notes, the prominence of certain notes in the octave, the general sequence of improvising the different notes and the peculiar embellishments which are used with certain notes and note transitions [4]. Vocal melodic improvisations and raga features**:** In both performance and identification of a certain raga some notes where the rendition halts temporarily, have an important role. These are called *nyas-swaras* in Indian musical terminology [5]. For example, in *raga Bhupali ga* is considered to be *vadi-swara* meaning it is used prominently and the rendition would often halt there [1]. On the other hand, certain notes are generally improvised with particular type of continuum or embellishment. For example, in *raga Bihag* while

descending back in the note sequence, there is always a continuum from *ni* to *pa* and also *pa* to *ga*. [1]. Finding out prominent notes regions and ornamentations used by the performer hence, becomes an important task in music information retrieval of Indian music for the larger objective of view of raga recognition.

Computational approaches of programming for finding out musical information have proved less effective in identification of such musical entities as there is considerable deviation from the expected ideal behavior. For example, a vibrato, i.e. a continuous oscillation of the underlying pitch frequency, makes the tone richer and is considered a steady note [2]. The identification of the raga is more important from perception point of view. Machine Learning models, in particular neural networks are believed to mimic human learning. This has been corroborated by our experimentation and study.

In this work we devise a model in which the first stage computes the mel-spectrograms of small audio clips extracted from a vocal recital. The second stage takes those clips to train a convolutional neural network for classification. The aim is finding out the regions in the audio where the singer has recited steady notes as against other regions where the singer has used note transitions.

2. RELATED WORK

The larger objective which is partially addressed by the present work is faithful and accurate transcription of the steady note regions and embellishment regions. Initial approaches towards this objective used feature-based methods. In particular, the instantaneous pitch frequency F0, is the main feature used for discerning the regions of steady notes and transitions and therefore different approaches towards the estimation of F0 were reviewed in [6] which re-affirmed the usefulness and accuracy of the YIN algorithm for finding out the instantaneous pitch frequency. A lot of work on feature-based music information retrieval (MIR) for different genres of world music has been done and documented. For example, an elegant way of classification of melodic motifs in raga music using time-series matching is given in [7]. Hidden Markov models (HMM) have also been used extensively for music transcription and musical expression detection tasks. For example, a remarkable method of transcribing frequency modulated musical expressions using HMM and spectrogram factorization approach is suggested in [8]. The limitation, however, of the feature-based computational methods is the need of hand engineering of the features and their accurate computation. Moreover, the need for discovery of rules for rule-based programming makes the process complex.

Machine learning, in particular convolutional neural network-based models have been used for audio classification tasks for a variety of applications and we have developed our insight reading related publications of both audio music applications and non-music applications. Moreover, the experts working in the area of machine learning in general, and convolutional neural networks in particular, suggest one to go through already used architectures of learning models before trying to create one's own. Therefore, here we briefly summarize only some of the classification tasks reported in the literature that we have gone through, to put the discussion in perspective.

Teeravajanadet *et al.* [9] have proposed a method to recognize the reason of infant cry based on convolutional neural network. They use dataset which is available for research purpose and have used an architecture with 6 two-dimensional convolution layers and able to achieve 84% accuracy. Chandu *et al.* [10] propose an automated bird species identification system using convolutional neural network (CNN), after creation and curation of dataset. A similar strategy is used by the Hall *et al.* [11] for classifying musical instruments with around 73% accuracy. Similarly, Viswanath and Babu [12] propose a vehicle classification system based on CNN. Neelima and Santiprabha [13] propose and automatic speaker verification system, against replay attacks, speech synthesis attacks, voice conversion and impersonation attacks, using CNN based classification, applied over a dataset created by the authors themselves, for this purpose, from social media. A new architecture of CNN has been proposed by Dawodi *et al.* [14] for classification of words or the dari language with creation and curation of dataset and a subsequent training. Dewa [15] use CNN for similar type of application with an aim to classify the vowels of Javanese language, with a different architecture. A combination of mel-frequency cepstral coefficients (MFCC) and CNN is used for speaker identification from a set of 60 speakers in [16]. These all works provide evidence to believe that CNN are capable of finding relevant features from audio for classifications tasks. But audio features of relevance in music are quite different from speech and other type of data. For example, describing in a simple way, the identity of the musical notes is governed largely by the steady state oscillation pattern of the waveform while the identity of syllables. required for speech recognition are largely governed by the transient components present generally at the onsets. However, CNN are also capable of learning relevant features from music data as well as summarized next.

Music genre classification for different datasets has been attempted by many researchers, for example the Sugianto and Suyanto [17] classify music in 10 different genre classes using mel-spectrograms over an available dataset. Cong *et al.* [18] propose a way to transcribe piano music against an available

labelled dataset using constant-Q-transform spectrogram and CNN. Pons and Serra [19] establish the importance of architecture of network used for feature extraction by taking non-trained (randomly weighted) CNNs and producing the features, so extracted, to a classifier. Different structures of CNNs have been investigated in [20] with computation of MFCC, discrete Fourier transform (DFT) and raw pulse code modulated (PCM) samples for detection of singing voice. An elaborate method for joint detection and classification of vocal melody from polyphonic signals with a main network and an auxiliary network with shared features is suggested in [21]. Anand [22] has developed a novel method of *raga* recognition using classification of pitch contour images using CNNs.

In order to get well versed with the methods of dataset creation and curation and for the want of publicly available datasets for Hindustani vocal music research, we have created our own dataset. Then we have devised architecture for classification of audio as steady note and note transition region. We give the details of the dataset in section 4. In the following section we present conceptual background to well understand the proposed model.

3. CONCEPTUAL BACKGROUND FOR PROPOSED MODEL

Again, to put the discussion in perspective, here we build a simple model, which takes small audio clips of vocal improvisations as input, and convert those to images using a time frequency analysis method, for classification. After experimentation with different time frequency representations, we decide to use the mel-spectrograms for this task, because these are giving results better than the mel-frequency cepstral coefficients and vanilla spectrograms. The images so obtained are fed to a machine learning model based on the CNN, which perform very well as far as classification of images is concerned. In the following part we describe the mel-spectrograms and the convolutional neural networks in brief.

3.1. Mel-spectrograms

Figure 3 shows mel-spectrogram for an audio for both steady note and note transition case. Conventional frequency domain representation of signals using Fourier transform has the limitation that it gives a global representation of frequency components present in any signal. Both time and frequency localized information is more relevant for better recognition and localization of different attributes of interest. Moreover, we require a two-dimensional representation for feeding to the CNN model. Therefore, windowed Fourier transform or more commonly called short time Fourier transform (STFT) is a better time-frequency representation which is used in this work, in a modified form as illustrated ahead. The formula for calculation of STFT is given in (1).

$$X(m,\omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega n} \quad (1)$$

The (1) is used to calculate discrete time STFT as a function of the angular frequency and the time index. Then conversion from angular frequency units to cycles per seconds can be done easily.

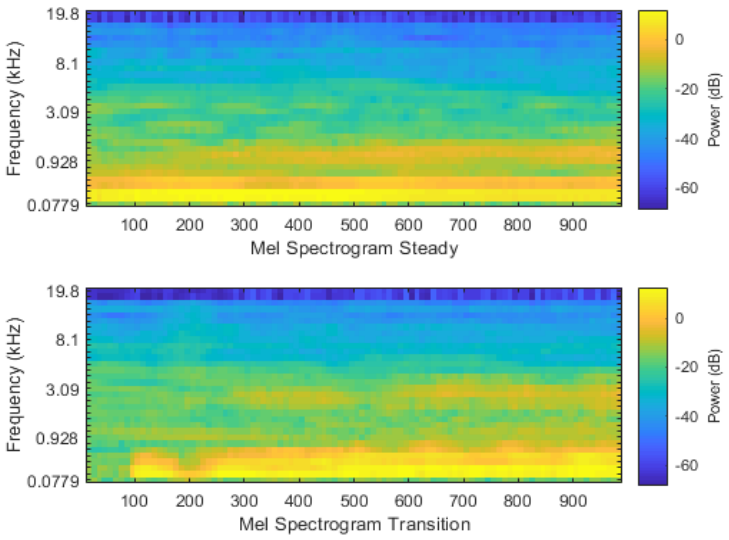

Figure 3. Mel-spctrogram represented as a heat map

Human auditory perception of both amplitude and frequencies is logarithmic. To accommodate this, mel scale of frequencies has been devised and is used frequently is speech analysis and recognition discipline. So, for calculation of mel-spectrogram the frequency axis is converted to mel-scale using the (2) and the squared-amplitude is converted to dB scale.

$$k_{mel} = 2595.log_{10}\left[1 + \frac{k}{700}\right] \quad (2)$$

The window length, type and other implementation details are given in a subsequent section.

3.2. Convolutional neural network (CNN)

CNN are a class of machine learning models, which are primarily used to operate on images, for different objectives like classification, and segmentation. Conventional image processing paradigm has had a legacy of hand engineered geometrical features, calculated and fed to neural networks, for different tasks. In the case of convolutional neural networks there is no need of those features. The feature extraction is also learnt by the model, via learning weights in convolutional kernels through back propagation. General architecture of a convolutional neural network consists of a stack of convolution layers and sub-sampling layers followed by some fully connected layers and finally the classification layers [20]. Convolution Kernels are two dimensional filters, the size of which are the decided at the time of design of model but the filter coefficients or the weights are learnt at the time of training. Sub-sampling layers can either be max pooling or mean pooling. Again, the pooling window size and the stride are decided at the time of design. There are no learnable parameters in the sub-sampling layers in general. A detailed explanation of convolutional neural network architecture can be found in [23]. Figure 4 shows a simplistic view of convolution and sub-sampling layer which is taken from [23].

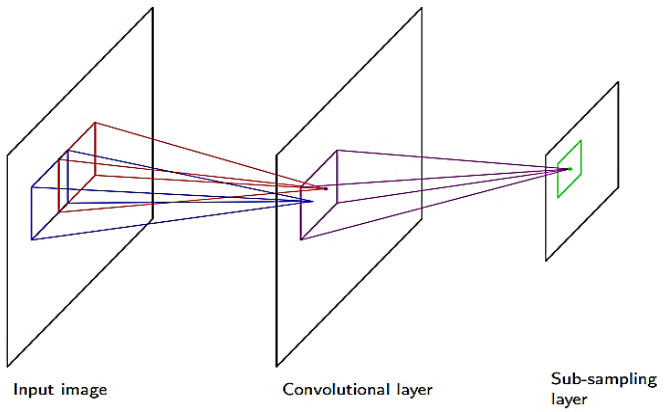

Figure 4. Convolutional neural network basic architectural unit from [23]

General formula which is used for calculation of the dimensions of layers is given in (3) which is used by us in designing the layers of the architecture that we have used for this task. We avoid a detailed explanation of CNN for limitation and space and the readers can go through [23] for details.

$$O = \frac{I-K+2P}{S} + 1 \quad (3)$$

The (3) gives the dimension of the activation in the next layer as calculated from the previous layer. This is applicable for both length and breadth. O stands for output size, I, input size, K is the convolution kernel size, P is padding and S stands for stride.

Different algorithms for training of neural network have been devised by researchers and available in the literature. After experimentation with our model, we used the stochastic gradient descent algorithm for training of our network because it worked best and is often used for training of basic CNN architectures. As the case with any classification task using neural networks, we use cross entropy loss for training our network. In the following section elaborate on our used dataset.

4. DATASET

In the creation of dataset for this work, we have tried to follow the principles of research corpora creation viz. purpose, coverage, completeness, quality and reusability as outlined in [24]. Moreover, we have been keen on the ethical dimensions of music information retrieval technology as explained in an elegant manner in [25]. Our field of interest for this work, according to our larger research objective, is vocal signals without any background accompaniment or percussion instrument. The audio data from research datasets, which are useful for this kind of research, are not available publicly due to copyright issues. Different researchers working on similar problems only provide pre-calculated features; the corresponding audio can be listened to online, but cannot be downloaded for experimentation. Moreover, it is hard to find recordings without background sound of instruments because in all performances, accompanying instruments are always there. Even in the initial unmetered improvisations of singers, called *alap*, the drone (*tanpura*) which is the reference tonic pitch is always there in background. However, after some experimentation we found out that the background of the drone does not affect the process of classification, because it is often low intensity, compared to the singing voice in the voiced part. The learning-based classifiers learn that it is always there in the background and differentiate the data based on the labels using the features which are different across differently labelled data.

With this background we take publicly available recordings of renowned male and female singers, in different ragas. After ripping the audio, we cut the initial *alap* renditions manually using the Audacity software which is an open-source digital audio workstation (DAW). The numerical details of created dataset are shown in Figure 5. Care is taken to discard the initial silence before beginning of the rendition and the latter portion which includes percussion instruments like *tabla* or *pakhawaj*, for providing the rhythm. Thereafter, the recordings are cut to small audio clips of 1 second each, using a program. Then, these all-audio clips which are now large in number are manually labelled in three categories. In (1) steady note (2) note transition (3) partially or fully unvoiced.

- Steady note: These are audio clips where a constant note is sung by the performed in the complete duration of one second and there are no silence regions.
- Note transition: These are audio clips which are also fully voiced but the note sung goes on changing and the change in the note is quite perceptible.
- Partially or fully unvoiced: If the sound clip has region, where voice is not there then it is labelled in this category.

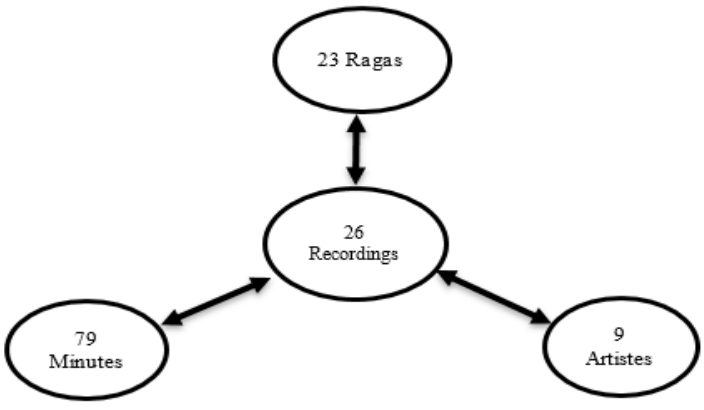

Figure 5. Numerical details of the dataset used

This manual labelling is done by one of us who has been receiving training in music for long, and then cross checked by a trained musician, to avoid any potential erroneous labelling. For the final training of the network, we only take the steady note and the note transition clips. Because unvoiced ones are very less in number, and classification of those is beyond the scope of objective of this work. Moreover, elegant computational algorithms are available to decide on the voiced or unvoiced portions from the discipline of speech recognition. After all the preprocessing and labelling that we have discussed so far, we have a total of 953 audio clips containing steady notes and 1028 audio clips containing note transitions.

Conversion to images: The audio clips are converted to portable graphics format (PNG) images in the following manner:

- We compute mel-spectrogram of the sound data and get matrix.

- We normalize each entry of the matrix by dividing it by maximum value is that matrix and multiply by 255 for full 8-bit representation.
- We store the matrix as a grayscale image in portable graphics format (PNG) format.

We show sample PNG image after reversal (black for larger values) for better depiction here in Figure 6. In the following section we present the overall model that we propose in this work.

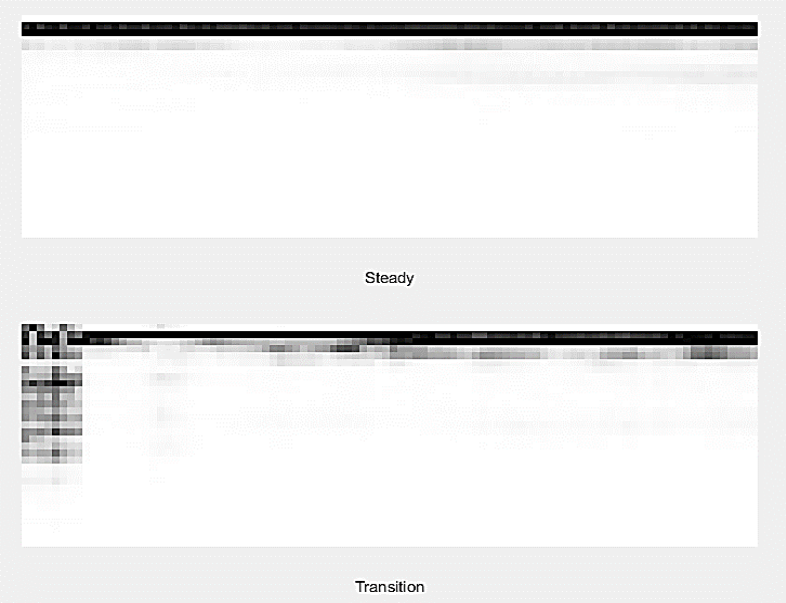

Figure 6. Reversed PNG images drawn from mel-spectrograms

5. PROPOSED MODEL

The proposed model is given in Figure 7. For training of the CNN, we create image data store folders with images of the two classes in separate sub-folders. The class labels are same as the sub-folder names. As discussed, earlier mel-spectrogram is a specific time frequency representation. The Implementation details of mel-spectrogram are given in Table 1.

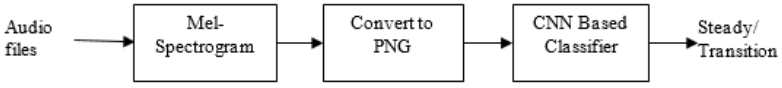

Figure 7. Overall model for proposed scheme of audio classification

Table 1. Implementation parameters of mel-spectrogram

Parameter	Value	Details
Window Length	1323	Default to 0.03×Fs
Overlap Length	882	Default to 0.02×Fs
FFT Length	1323	Default to Window Length
Number of bands	32	Default Value
Frequency Range	0 to 22050 Hz	Up to Nyquist frequency
Spectrum Type	Power Spectrum	Default Value

The conversion of matrices to PNG is done using the steps illustrated earlier. The implementation details of the CNN are listed in Table 2. Training: we use the stochastic gradient descent algorithm with momentum for training of the network, with maximum epoch set at 20. The initial learning rate is 0.0001. The overall division of the dataset for training purpose is done such that 70% of the data is for training, 20% for validation and 10% for testing. The randomization for training and validation is done by the training method itself. The loss function is the cross-entropy loss function.

The values of parameters obtained during and after completion of training are in Table 3. The figure showing the training progress is shown next in Figure 8. We can see from the Figure 8 that the training process is converging. In the next section we elaborate on the results obtained here.

Table 2. Architecture of CNN implementation

Layer Name	Details	Type	Activations	Learnables
Imageinput	32×98×1 Images (Gray-Scale)	Image Input Layer	32×98×1	-
Conv	5×17×1 Convolutions with stride [1, 3] and padding same (20)	Convolution Layer	32×33×20	Weights: 5×17×1×20 Bias: 1×1×20
Relu	Rectified Linear Activation Unit	Relu	32×33×20	-
Maxpool	2×2 max pooling with stride [2,2] and padding same	Max Pooling Layer	16×17×20	-
Fc	fully connected layer with 2 neurons	Fully Connected	1×1×2	Weights: 2×5440 Bias: 2×1
Softmax	Softmax	Softmax Layer	1×1×2	-

Table 3. CNN training details

Parameter	Value
Validation accuracy	71.46%
Elapsed time	21 sec
Epoch	20
Iterations per epoch	10
Initial Learning rate	0.0001

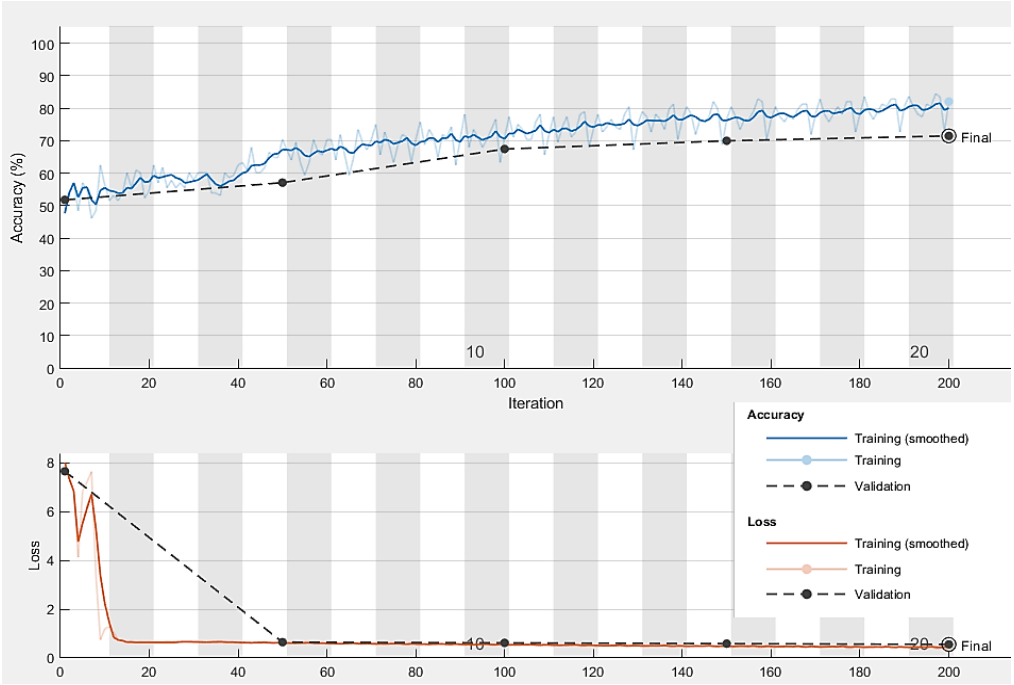

Figure 8. Training process: accuracy (above) and loss (below) as epochs and iterations progress

6. RESULTS AND DISCUSSION

We have been able to design a system which learns the relevant features from the mel-spectrogram of the audio clips for the classification task, as can be seen from the profile of change accuracy and loss function values over time. As can be seen from Figure 8, the network converges well. As can be seen from Table 3, we are able to achieve a validation accuracy of about 71% and testing accuracy of about 70% in the run that we are referring to. This means that the good generalization is possible in this training. On the other hand, we can see from the loss profile that there is little change in the value after around 50 iterations. This is due to the problem of vanishing gradients. Going by the literature studied by us, this is first work of its kind, and therefore accuracy of this range is considerable. Table 4 gives the confusion matrix.

Classification performance: In the calculation of the precision, recall and the F1-score value, we have considered the steady note as the positive class and transition as the negative class. These classification performance metrics are given in Table 5. Analysing in Table 5 parameters we can say that we have achieved good accuracy of classification. In the following section we conclude the discussion.

Table 4. Confusion matrix of classification

		Predicted Class	
		Steady	Transition
True Class	Steady	83	12
	Transition	29	74

Table 5. Classification performance matrix

Parameter	Value
Precision	0.74
Recall	0.87
F1-Score	0.80

7. CONCLUSION

It can be seen that the network is able to capture relevant features and is converging. The accuracy for the test set is good enough taking in to account the simplicity of the network. Some errors might also have occurred due to inaccuracy of perception in creation of dataset. Trained musician with better hold on notes and embellishments would be able to curate the dataset with a larger accuracy. Another future improvement may be in the depth of the network, in that it is generally considered that deeper the network better are the chances of conversion and getting good accuracy in approximation of any function. The contribution of this work is finding out a simple network which works for audio classification towards finding steady note regions and note continuum regions. The trained network can be used as a part of a larger system which would give a complete transcription of melodies, giving out details of the notes and embellishments. This work also gives a direction for design and development of specialized network architectures for audio classification tasks like the ones already available for image classification tasks and training of those, so that later on transfer learning can be used for the audio classification tasks at hand.

ACKNOWLEDGEMENTS

We acknowledge the support provided by the team of CIDI and SIF, SGSITS Indore. The musicians at Pancham Nishad Sangeet Sansthan, Indore are acknowledged for the support provided for labelling. The routines for dataset creation and curation were written and the neural network was implemented in Matlab language. The manual cutting of audio files was done in Audacity software.

REFERENCES

[1] V. N. Bhatkhande, “Hindustani Sangeet Paddhati: Kramik Pustak Malika Vol. I-VI*,” Sangeet Karyalaya*, 1990.
[2] A. Klapuri and M. Davy, “Signal Processing Methods for Music Transcription,” *Springer,* 2006.
[3] A. Datta, S. Solanki, R. Sengupta, S. Chakraborty, K. Mahto, and A. Patranabis, “Signal Analysis of Hindustani Classical Music,” (1st. ed.) *Springer Publishing Company, Incorporated,* 2017.
[4] S. Gulati, K. K. Ganguli, S. Gupta, A. Srinivasamurthy, and X. Serra, “Ragawise: A Lightweight Real-time Raga Recognition System for Indian Art Music,” *Late breaking demo at the 16th International Society for Music Information Retrieval Conference (ISMIR),* October 2015, Malaga, Spain, 2015.
[5] S. Gulati, “Computational Approaches for Melodic Description in Indian Art Music Corpora,” Ph D Thesis, UPF, Barcelona, Spain, 2016.
[6] A. Rege and R. Sindal, “Review of F0 Estimation in the Context of Indian Classical Music Expression Detection,” *Social Networking and Computational Intelligence. Lecture Notes in Networks and Systems*, vol. 100. Springer, Singapore, 2020, doi: 10.1007/978-981-15-2071-6_21.
[7] P. Rao *et al.,* “Classification of Melodic Motifs in Raga Music with Time-series Matching,” *Journal of New Music Research*, vol. 43, no. 1, pp. 115-131, 2020, doi: 10.1080/09298215.2013.873470.
[8] D. Sung and K. Lee, "Transcribing Frequency Modulated Musical Expressions from Polyphonic Music Using HMM Constrained Shift Invariant PLCA," *2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, Kitakyushu, Japan, 2014, pp. 562-565, doi: 10.1109/IIH-MSP.2014.145.
[9] K. Teeravajanadet, N. Siwilai, K. Thanaselanggul, N. Ponsiricharoenphan, S. Tungjitkusolmun, and P. Phasukkit, "An Infant Cry Recognition based on Convolutional Neural Network Method," *2019 12th Biomedical Engineering International Conference (BMEiCON),* UbonRatchathani, Thailand, 2019, pp. 1-4, doi: 10.1109/BMEiCON47515.2019.8990191.
[10] B. Chandu, A. Munikoti, K. S. Murthy, G. Murthy V., and C. Nagaraj, "Automated Bird Species Identification using Audio Signal Processing and Neural Networks," *2020 International Conference on Artificial Intelligence and Signal Processing (AISP)*, Amaravati, India, 2020, pp. 1-5, doi: 10.1109/AISP48273.2020.9073584.
[11] J. Hall, W. O’Quinn, and R. J. Haddad, "An Efficient Visual-Based Method for Classifying Instrumental Audio using Deep Learning," *2019 SoutheastCon*, Huntsville, AL, USA, 2019, pp. 1-4, doi: 10.1109/SoutheastCon42311.2019.9020571.

[12] V. Viswanath and B. P. Babu, "Vehicle Classification with Audio and Video Modalities Using CNN and Decision-Level Fusion," *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)*, Bhimtal, India, 2020, pp. 482-486, doi: 10.1109/CICN49253.2020.9242556.
[13] M. Neelima and I. Santiprabha, "Mimicry Voice Detection using Convolutional Neural Networks," *2020 International Conference on Smart Electronics and Communication (ICOSEC),* Trichy, India, 2020, pp. 314-318, doi: 10.1109/ICOSEC49089.2020.9215407.
[14] M. Dawodi, J. A. Baktash, T. Wada, N. Alam, and M. Z. Joya, "Dari Speech Classification Using Deep Convolutional Neural Network," *2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS),* Vancouver, BC, Canada, 2020, pp. 1-4, doi: 10.1109/IEMTRONICS51293.2020.9216370.
[15] C. K. Dewa, "Javanese vowels sound classification with convolutional neural network," *2016 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, Lombok, Indonesia, 2016, pp. 123-128, doi: 10.1109/ISITIA.2016.7828645.
[16] A. Ashar, M. S. Bhatti, and U. Mushtaq, "Speaker Identification Using a Hybrid CNN-MFCC Approach," *2020 International Conference on Emerging Trends in Smart Technologies (ICETST)*, Karachi, Pakistan, 2020, pp. 1-4, doi: 10.1109/ICETST49965.2020.9080730.
[17] S. Sugianto and S. Suyanto, "Voting-Based Music Genre Classification Using Melspectogram and Convolutional Neural Network," *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI),* Yogyakarta, Indonesia, 2019, pp. 330-333, doi: 10.1109/ISRITI48646.2019.9034644.
[18] F. Cong, S. Liu, L. Guo, and G. A. Wiggins, "A Parallel Fusion Approach to Piano Music Transcription Based on Convolutional Neural Network," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Calgary, AB, Canada, 2018, pp. 391-395, doi: 10.1109/ICASSP.2018.8461794.
[19] J. Pons and X. Serra, "Randomly Weighted CNNs for (Music) Audio Classification," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Brighton, UK, 2019, pp. 336-340, doi: 10.1109/ICASSP.2019.8682912.
[20] H. Huang, W. Chen, C. Liu, and S. D. You, "Singing voice detection based on convolutional neural networks," *2018 7th International Symposium on Next Generation Electronics (ISNE),* Taipei, Taiwan, 2018, pp. 1-4, doi: 10.1109/ISNE.2018.8394727.
[21] S. Kumand and J. Nam, "Joint Detection and Classification of Singing Voice Melody Using Convolutional Recurrent Neural Networks," *Appl. Sci.,* vol. 9, no. 7, 2019, doi: 10.3390/app9071324.
[22] A. Anand, "Raga Identification Using Convolutional Neural Network," *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*, Gangtok, India, 2019, pp. 1-6, doi: 10.1109/ICACCP.2019.8882942.
[23] C. M. Bishop, "Pattern recognition and machine learning," *New York: Springer*, 2006.
[24] X. Serra, "Creating Research Corpora for the Computational Study of Music: the case of the CompMusic Project," *AES 53rd International Conference on Semantic Audio*, 2014.
[25] A. Holzapfel, B. L. Sturm, and M. Coeckelbergh, "Ethical Dimensions of Music Information Retrieval Technology," *Transactions of the International Society for Music Information Retrieval*, vol. 1, no. 1, pp. 44-55, 2018, doi: 10.5334/tismir.13.

BIOGRAPHIES OF AUTHORS

Amit Rege received the degree of B.E. in Electronics and Instrumentation Engineering and M. Tech in Digital Instrumentation both from IET-DAVV Indore. He started his career in academics as a lecturer at Medicaps Institute of Technology and Management, Indore in 2007. He has 12 years of teaching experience. His research area includes music information retrieval. He has received training in vocal and instrumental music. His current research is focused on singing transcription and musical expression detection.

Ravi Sindal received Ph D. degree in Electronics and Telecommunication from Devi Ahilya University in 2011, respectively. Since 2013, he has been a Professor of Electronics and Telecommunication at Institute of Engineering and Technology, Devi Ahilya University, Indore. His research areas include Radio Resource Management, Modeling and Simulation, Digital Design, Wireless Networks and Signal Processing.