Oktoechos Classification in Liturgical Music Using Musical Texture Features

Rajeev Rajan *, Amlu Anna Joshy, and Varsha Shiburaj

College of Engineering, Trivandrum, Thiruvananthapuram
APJ Abdul Kalam Technological University, Kerala, India
⋆rajeev@cet.ac.in

Abstract. A distinguishing feature of the music repertoire of the Syrian tradition is the system of classifying melodies into eight tunes, called 'oktoēchos'. It inspired many traditions, such as Greek and Indian liturgical music. In oktoēchos tradition, liturgical hymns are sung in eight modes or eight colours (known as eight 'niram', regionally). In this paper, the automatic oktoēchos genre classification is addressed using musical texture features (MTF), i-vectors and Mel-spectrograms through deep learning strategies. The performance of the proposed approaches is evaluated using a newly created corpus of liturgical music in Malayalam. Long-short term memory (LSTM)-based experiment reports the average classification accuracy of 83.76%, with a significant margin over other frameworks. The experiments demonstrate the potential of LSTM in learning temporal information through MTF in recognizing eight modes in oktoēchos system.

Keywords: liturgy, colour, timbral, deep learning.

1 Introduction

Oktoēchos classification in liturgical music (music using in worship) is addressed using deep learning frameworks in the paper. Music plays a vital role in liturgy because music itself is a language that goes beyond even cultures and races. The vast diversity of forms, styles, and functions in the music used for worship makes it challenging to categorize liturgical music. Musical roles have been distributed in different ways in different rites. Indian orthodox church has imbibed this music system into its liturgy through its relationship with the orthodox church in Syria (Antiochian liturgy). A distinguishing feature of the music repertoire of the Syrian tradition is the system of classifying melodies into eight tunes [15]. This musical tradition is transferred to Indian orthodox liturgical music through centuries with hymns in the Malayalam[1]. Most of the hymns used for various feasts and occasions are musically composed under eight tunes. The system of singing the same text in eight different melodies in an eight-week cycle is referred to as the'oktoēchos' [15].

[1] https://en.wikipedia.org/wiki/Malayalam

1.1 Oktoēchos

Western Syriac music is based on the classical tradition prescribed in 'Bethgazzo'[2]. In oktoēchos tradition, liturgical hymns are sung in eight modes, similar to the Greek liturgy. They are a group of eight adaptable melody types, known as eight 'colours' or 'niram' [27]. None of the Syriac melodies may cover eight notes in an octave. It may often cover three or four or five notes. There is a similarity in Syrian/Indian liturgical (Malankara) hymnal music and rāga[3] of Indian art music. But they cannot be taken in equal level because the rāga classification of Indian art music is incomparable in its scientific systematisation. Each rāga has a particular mode and temperament. Oktoēchos can be compared to rāga in the sense that they are also creating passion or rasa during singing [27]. In Indian art music, a hymn in a rāga can be sung or played in another rāga. The same principle is applied in the oktoēchos system that most of the liturgical hymns can be sung in all the eight tunes.

Oktoēchos is considered as a cyclic system because it is performed in a cycle of eight weeks with two colours in a week. Each colour begins with evening prayer of Sunday. If the first colour is used in the evening, the same is continued for the rest of the day. From Monday evening onwards the fifth colour is used. On Tuesday, it is again switched on to the first colour and so on. The next Sunday begins with the second colour. It is continued in the order 1-5; 2-6; 3-7; 4-8; till to the fourth Sunday and on the fifth Sunday onwards the order becomes 5-1; 6-2; 7-3; 8-4.

1.2 Related Work

Although there has been significant work in music genre classification, the proposed task of liturgical music genre classification is first of its kind. Melodic features [23] and local features [28] have been employed well for genre classification task. Researchers used both generative and discriminative models [12, 24] for music classification. Musical texture features are recently used in meter classification works [22, 21, 19]. Music genre classification is addressed using feature fusion in [20]. A model capable of learning distinctive rhythmic structures of different music genres using unsupervised learning is proposed in [16]. In contrast with the standard approaches, model-based distances between time series can take into account the structure of the songs by modelling the dynamics of the parameter sequence [7]. More recent deep learning approaches process spectrograms for the task of music genre classification [17, 3]. Regarding multimodal approaches found in the literature, most of them combine audio and song lyrics [11] through a fusion framework. The proposed task is similar to music genre classification, but shares the textual content across modes is one of the specific traits of the oktoēchos genre system. The aim of the work is to explore the ability of LSTM to capture the long range dependency in learning temporal patterns.

The rest of the paper is organized as follows; Section 2 describes the proposed system followed by the performance evaluation in Section 3. The analysis of results is given in Section 4. Finally, the paper is concluded in Section 5.

[2] Bethgazzo is a Syriac liturgical book that contains a collection of Syriac chants and melodies.

[3] rāga is the fundamental melodic framework for both Carnatic and Hindusthani traditions

2 System Description

2.1 Feature Extraction

It has already been proven that timbral and rhythmic features are useful in genre classification task [1]. In our experiment, we extracted timbral and rhythmic features as musical texture features. Timbral features, namely Mel-frequency cepstral features (MFCC) and low-level timbral feature-set (T_{LF}), are computed in the front-end. Spectral centroid, spectral roll-off, spectral flux, and spectral entropy [13] are extracted as low-level timbral feature set. Besides, features namely tempo, pulse clarity, event density [10] are computed as rhythmic cues (R_F). Event density represents the number of events per unit time in the music piece. It is a measure that captures how easily "listeners can perceive the underlying rhythmic or metrical pulsation of music" [10]. This feature plays an important role in musical genre recognition, in particular, allowing a finer discrimination between genres that present similar average tempo, but that differ in the degree of emergence of the main pulsation over the rhythmic texture [10]. The distribution of pulse clarity for the corpus is shown in Fig. 1. It can be seen that the pulse clarity distribution for niram 1, niram 2 and niram 3 is different from the rest. Low-level timbral features and rhythmic features are computed using MIRToolbox [4].

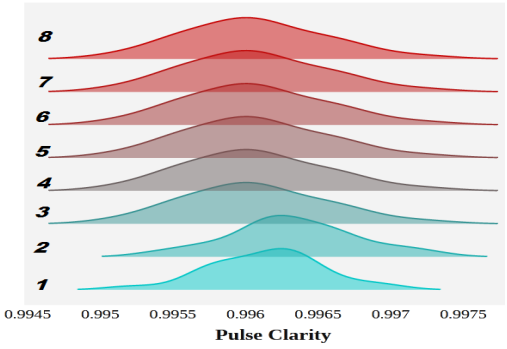

Fig. 1: Distribution of pulse clarity for the colours

Given the success of using i-vectors for speaker and music processing tasks [29, 6], we use the i-vector framework in the proposed task for performance comparison. The i-vector-based statistical feature has been employed well in the task of music genre classification [4]. In i-vector system [5], the high dimensional GMM super vector space (generated from concatenating the mean values of GMM) is mapped to a low dimensional space called total variability space. The target utterance GMM is adapted from a universal background model (UBM) using eigenvoice adaption. The target GMM super vector can be viewed as a shifted version of UBM. Formally, a target GMM super vector M can be written as:

$$M = m + Tw \tag{1}$$

[4] https://www.jyu.fi/hytk/fi/laitokset/mutku/en/research/materials/

where m represents the UBM super vector, T is a low dimensional rectangular total variability (TV) matrix, and w is termed as i-vector. Using training data, the UBM and TV matrix is modeled by expectation maximization. 100 dimensional i-vectors (i_{MFCC}) are computed for each song from MFCC using Alize tool kit [2].

In the final phase, visual representation of audio files, spectrograms are utilized for the proposed task. Since Mel-spectrogram has already been utilized well for music genre classification tasks [25, 8], we also experimented with mel-spectrogram-CNN framework for the proposed task. Mel-spectrogram can be seen as the spectrogram smoothed, with high precision in the low frequencies and low precision in the high frequencies. Mel-spectrogram is computed with frame size of 40 ms and hop size of 10 ms using 128 bins.

2.2 Classification Scheme

We experimented with four classifiers, namely, SVM, DNN, CNN and LSTM. DNN is based on six hidden layered network, which uses 64, 128, 256, 512, 1024, 2048 nodes in successive layers with a dropout of 0.25. The network is trained with the batch size is 32 for 150 epochs by AdaMax optimization algorithm. Relu and softmax have been chosen for hidden and output layers, respectively.

Table 1: LSTM architecture used for the experiment

Sl no.	**Output Size**	**Description**
1	(45,64)	LSTM, 64 hidden units
2	(46, 64)	Dropout (0.25)
3	(1024)	LSTM, 1024 hidden units
4	(1024)	Dropout (0.25)
5	(8)	Dense (8 hidden units)

The proposed CNN has six convolution layers, followed by max-pooling. We use filters with a very small 3×3 receptive fields, with a fixed stride of one and increase the number of filters for the layer by a factor of 2 after every layer. Global max-pooling is adopted in the final max-pooling layer, which is then fed to a fully connected layer. The training is done with 100 epochs by optimizing the categorical cross-entropy between predictions and targets using Adam optimizer, with a learning rate of 0.001.

LSTM architecture shown in Table 1 effectively utilized to track the temporal pattern embedded in the modes of the music. LSTM-RNNs can capture long-range temporal dependencies by overcoming the vanishing gradient problem in conventional RNNs [26]. RNN tap inherent temporal pattern embedded within the frame-wise computed MTF. Deep learning schemes and SVM are implemented using and Keras-TensorFlow and LibSVM, respectively.

Table 2: Overall classification accuracy for the experiments

Sl.No	Feature	Method	Accr.(%)
1	MFCC+ T_{LF} + R_F	SVM	42.65
2	MFCC + T_{LF}+ R_F	DNN	48.70
3	i_{MFCC} + T_{LF} + R_F	DNN	50.00
4	Mel-spectrogram	CNN	52.60
5	**MFCC +** T_{LF}**+** R_F	**LSTM**	**83.76**

3 Performance Evaluation

3.1 Database

A database is created in a studio environment and it consists of eight nirams (colours), with 384 audio tracks with duration 25 to 45 sec per file. No accompaniments were there in the audio files. A total of 15 professional singers in the age group 12 to 50, were participated in the data recording and the whole session was recorded at 44.1kHz. All the singers were very much familiar with the singing modes in 'oktoēchos'. Malayalam hymns were collected from the liturgical book of Indian Orthodox church. The recordings were made niram by niram in successive sessions using a high-quality microphone, A few audio files can be accessed at https://sites.google.com/view/audiosamples-2020/. During experimentation, 60% files of the dataset are used for training, 10% is used for validation and the rest for testing.

3.2 Experimental set-up

MFCCs (39 dim comprising 13 dim MFCC, its delta and delta-delta features), timbral (T_{LF}, 4 dim) rhythmic (R_F, 3 dim) are frame-wise computed with a frame width of 40 ms and hop size of 10 ms and fused in feature-level to obtain 46-dimensional MTF. In the i-vector experimental phase, 100-dimensional i-vectors are computed using 128 mixture GMM from MFCC using Alize tool-kit [2]. UBM model is trained using features derived from the auxiliary database comprising audio file other than the files in the corpus. Auxiliary database, comprising 300 audio files (duration 25-35ms) of liturgical music category, is prepared in a studio environment. The songs from the training data are used for modelling the total variability matrix T by Eigen voice adaption. In the fusion scheme, track level aggregated timbral (T_{LF}) and rhythmic (R_F) features are concatenated with track-level computed i-vectors. Following the evaluation method widely used in the MIR tasks, we computed the precision and recall and the F1 measure as basic evaluation metrics for the performance.

4 Results and Analysis

The results are tabulated in Table 2. As per the table, the average classification accuracy of 42.66%, 48.70%, 50.00%, 52.60% and 83.76% are reported for SVM, DNN, i-vector

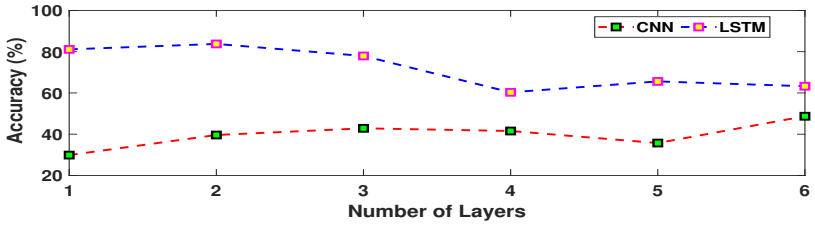

Fig. 2: Accuracy with number of layers for CNN and LSTM

framework, Mel-spectrogram-CNN and LSTM, respectively. It is worth noting that the LSTM outperforms other approaches with a significant margin. It is reasonable to say that time pattern capturing scheme is needed in order to recover more relevant information from temporal embedded musical traits [7]. Experiments show that the LSTM approach is promising for the given task, improving on the case where the dynamics are not taken into account, and a stationary characterization of the sequences is employed. LSTM utilized musical textural features to capture song dynamics effectively to perform oktoēchos classification. It is shown in [4] that the important music elements can be captured by i-vectors and may potentially benefit to the classification of music signal. A possible cause of the low value of accuracy in the given experimental set-up may potentially due to the inability to capture the rhythmic-temporal dynamics well with the given UBM framework. Besides, aggregation of musical texture features to track-level might have deteriorated the performance.

The performance with varying the number of layers of the network is shown in Fig. 2. For the CNN framework, the result improved, as the number of layers increased up to six and then saturated due to overfitting. It is due to the fact that as n increases, the model grows in-depth, and the upper layers find efficient feature representations that are invariant to small perturbations leading to better model generalization. The authors [14] emphasize the need for more training data in the visual representation-based approaches for the genre classification task. It is stated that CNN needs a large size of data to achieve better results since it is not successful enough for less data [9]. An elegant solution to this problem is data augmentation, by which deformations to a collection of annotated training samples results in additional training data. During LSTM approach, maximum accuracy is obtained for two layers as seen in lower-pane in Fig. 2. The proposed experiment validates the claim that temporal information has effectively been learned by MTF-LSTM framework. The experimental insights in [18] show that the performance of the system depends on the temporal architecture, which is basically designed by considering the musical domain knowledge.

The normalized confusion metrics of LSTM is plotted in Fig. 3. Class-wise classification accuracy of all nirams are greater than 70% for LSTM. Niram 5 and niram 7 report accuracy greater than 90%. Class wise accuracy can be examined from the bar plot given in Fig. 4 from all phases. The significant improvement in class-wise accuracy of nirams 1, 3, 7, and 8 over CNN based framework can be seen from the plot. The performance can potentially be improved using data augmentation and proper choice of architecture. The performance metrics precision, recall and F1 score for all the five

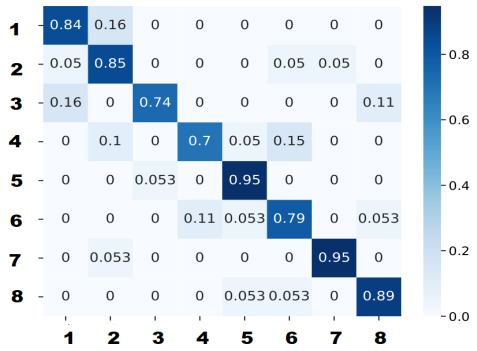

Fig. 3: Normalized Confusion Matrices for MTF-LSTM

approaches are given in Table 3. The average F1 measure of 0.43, 0.50, 0.50, 0.52, 0.84 are reported for SVM, DNN, i-vector-DNN, CNN and LSTM, respectively. The high values of precision, recall and F1 score show the significance of LSTM for the proposed task. Fig. 5 visualizes the output vectors produced by the snippets for the last

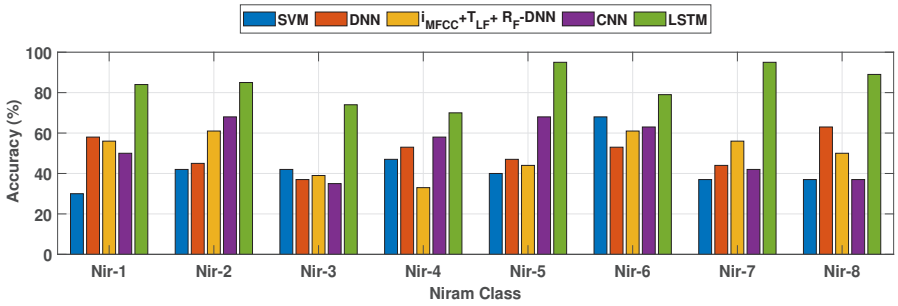

Fig. 4: Class-wise performance for entire phases of the experiments

dense layer of the trained LSTM network using t-SNE. Note that there is good clustering (as represented with colour) and a general separation of different classes for LSTM. It is important to note the effectiveness of LSTM in the proposed task without using any modelling data or augmentation data as that of i-vector or CNN methodologies. Since the results show the promise of temporal pattern learning, other frameworks have to be experimented to investigate the potential of the proposed approach.

5 Conclusion

Oktoēchos classification is addressed in this paper. The performance of the proposed approaches is evaluated using a newly created corpus of Liturgical music in Malayalam.

Table 3: Precision (P), recall (R), and F1 measure

SL.No	**Colour**	**MFCC+T_{LF}+R_F-SVM**			**MFCC+T_{LF}+R_F-DNN**			**$\mathbf{i}_{\mathbf{MFCC}}$+$T_{LF}$+$R_F$-DNN**			**Mel-spectrogram-CNN**			**MFCC+ T_{LF}+R_F- LSTM**		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	Niram-1	0.55	0.30	0.39	0.35	0.58	0.44	0.48	0.56	0.51	0.42	0.50	0.45	0.80	0.84	**0.82**
2	Niram-2	0.40	0.42	0.41	0.36	0.45	0.40	0.46	0.61	0.52	0.52	0.68	0.59	0.74	0.85	**0.79**
3	Niram-3	0.25	0.42	0.31	0.32	0.37	0.34	0.54	0.39	0.45	0.70	0.35	0.47	0.93	0.74	**0.82**
4	Niram-4	0.64	0.47	0.55	0.71	0.53	0.61	0.46	0.33	0.39	0.69	0.58	0.63	0.88	0.70	**0.78**
5	Niram-5	0.62	0.40	0.48	0.53	0.47	0.50	0.40	0.44	0.42	0.54	0.68	0.60	0.86	0.95	**0.90**
6	Niram-6	0.43	0.68	0.53	0.62	0.53	0.57	0.52	0.61	0.56	0.55	0.63	0.59	0.75	0.79	**0.77**
7	Niram-7	0.33	0.37	0.35	0.50	0.35	0.41	0.59	0.56	0.57	0.47	0.42	0.44	0.95	0.95	**0.95**
8	Niram-8	0.54	0.37	0.44	0.80	0.63	0.71	0.60	0.50	0.55	0.44	0.37	0.40	0.85	0.89	**0.87**
	Macro	0.47	0.43	**0.43**	0.52	0.48	**0.50**	0.50	0.50	**0.50**	0.54	0.53	**0.52**	0.85	0.84	**0.84**

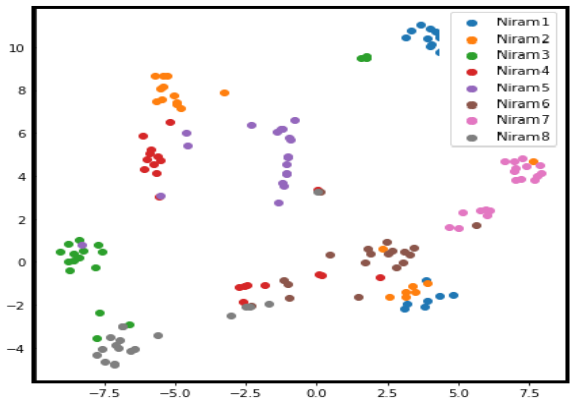

Fig. 5: t_{SNE} plot from LSTM

The evaluation shows the potential of MTF-LSTM framework in Oktoecẖos classification with an average classification accuracy of 83.76%. Since the Greek liturgy and Gregorian chant also share similar musical traits with Syrian tradition, the musicological insights observed can potentially be applied to those traditions as well.

References

1. Baniya, B.K., Ghimire, D., Lee, J.: Automatic music genre classification using timbral texture and rhythmic content features. Proc. of 17th Int. Conference on Advanced Communication Technology pp. 434–443 (2015)
2. Bonastre, J.F., Wils, F., Meignier, S.: Aliźe, a free toolkit for speaker recognition. in Proc. of Interspeech **1**, 737–740 (01 2005)
3. Choi, K., Fazekas, G., Sandler, M., Cho, K.: Convolutional recurrent neural networks for music classification. in Proc. of IEEE Int. Conference on Acoustics, Speech and Signal Processing pp. 2392–2396 (2017)
4. Dai, J., Xue, W., Liu, W.: Multilingual i-vector based statistical modeling for music genre classification. Proc. of Interspeech pp. 459–463 (2017). https://doi.org/10.21437/Interspeech.2017-74
5. Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P.: Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech, and Language Processing **19**, 788–798 (2011)
6. Eghbal-zadeh, H., Lehner, B., Schedl, M., Widmer, G.: I-vectors for timbre-based music similarity and music artist classification. in Proc. of 16th Int. Society for Music Information Retrieval Conference pp. 554–560 (2015)
7. Garcia-Garcia, D., Arenas-Garcia, J., Parrado-Hernandez, E., Diaz-de Maria, F.: Music genre classification using the temporal structure of songs. in Proc. of IEEE Int. Workshop on Machine Learning for Signal Processing (2010)
8. Ghosal, D., Kolekar, M.H.: Music genre recognition using deep neural networks and transfer learning. in Proc. of Interspeech pp. 2087–2091 (2018)
9. Kaya, M., Bilge, S.H.: Deep metric learning: A survey. Symmetry **11**(9), 1–26 (2019)

10. Lartillot, O., Eerola, T., Toiviainen, P., Fornari, J.: Multi-feature modeling of pulse clarity: Design, validation and optimization. in Proc. of the 9th Int. Conference on Music Information Retrieval pp. 1–5 (2008)
11. Laurier, C., Grivolla, J., Herrera, P.: Multimodal music mood classification using audio and lyrics. in Proc. of Seventh IEEE Int. Conference on Machine Learning and Applications pp. 688–693 (2008)
12. Li, T., Ogihara, M., Li, Q.: A comparative study on content-based music genre classification. in Proc. of 26th Int. ACM Conference on Research and Development in Information Retrieval pp. 282–289 (2003)
13. Li, T., Ogihara, M., Li, Q.: A comparative study on content-based music genre classification. in Proc. of the 26th Annual Int. ACM Conference on Research and development in information retrieval pp. 282–289 (2003)
14. Liua, C., Fengb, L., Liuc, G., Wangd, H., Liub, S.: Bottom-up broadcast neural network for music genre classification. Pattern Recognition Letters pp. 1–7 (2019)
15. Palackal, J.: Oktoechos of the syrian orthodox churches in south india. Ethnomusicology **48**, 229–250 (2004)
16. Pesek, M., Leonardis, A., Marolt, M.: An analysis of rhythmic patterns with unsupervised learning. Applied Science pp. 1–22 (2020)
17. Pons, J., Lidy, T., Serra, X.: Experimenting with musically motivated convolutional neural network. in Proc. of Int. Workshop on Content-Based Multimedia Indexing pp. 1–5 (2016)
18. Pons, J., Serra, X.: Randomly weighted cnns for (music) audio classification. in Proc. of IEEE Int. Conference on Acoustics, Speech and Signal Processing pp. 336–340 (2019)
19. Rajan, R., Kumar, A.V., Babu, B.P.: Poetic meter classification using i-vector-mtf fusion. In: INTERSPEECH (2020)
20. Rajan, R., Murthy, H.A.: Music genre classification by fusion of modified group delay and melodic features. In: 2017 Twenty-third National Conference on Communications (NCC). pp. 1–6 (2017). https://doi.org/10.1109/NCC.2017.8077056
21. Rajan, R., Raju, A.A.: Poetic meter classification using acoustic cues. In: 2018 International Conference on Signal Processing and Communications (SPCOM). pp. 31–35 (2018). https://doi.org/10.1109/SPCOM.2018.8724426
22. Rajan, R., Raju, A.A.: Deep neural network based poetic meter classification using musical texture feature fusion. In: 2019 27th European Signal Processing Conference (EUSIPCO). pp. 1–5 (2019). https://doi.org/10.23919/EUSIPCO.2019.8902998
23. Salamon, J., Rocha, B., Gomez, E.: Musical genre classification using melody features extracted from polyphonic music signals. in Proc. of IEEE Int. Conference on Audio, Speech, and Signal Processing pp. 81–85 (2012)
24. Shao, X., Xu, C., Kankanhalli., M.S.: Unsupervised classification of music genre using hidden Markov model. in Proc. of IEEE Int. Confernce on Multimedia and Expo, **3**, 2023–2026 (2004)
25. Sukhavasi, M., Adappa, S.: Music theme recognition using CNN and self-attention. preprint arXiv:1911.07041 (2019)
26. Tang, C.P., Chui, K., Yu, Y., Zeng, Z., Wong, K.: Music genre classification using a hierarchical long short term memory model. in Proc. of Int. Conference on Information Retrieval, Japan pp. 521–526 (2018)
27. Vysanethu, P.: Musicality makes the malankara liturgy musical (moran etho 2). St.Ephrem Ecumenical Research Institute, Kottayam, Kerala, India (2004)
28. Wulfing, J., Riedmille, M.: Unsupervised learning of local features for music classifications. in Proc. of Int. Society for Music Information Retrieval Conference. pp. 139–144 (2012)
29. Zhong, J., Hu, W., Soong, F., Meng, H.: DNN i-vector speaker verification with short, text-constrained test utterances. in Proc. of Interspeech pp. 1507–1511 (2017). https://doi.org/10.21437/Interspeech.2017-1036