Proc. of the National Conference on Communications (NCC), Jan 2011, Bangalore, India

Automatic Genre Classification of North Indian Devotional Music

Sujeet Kini, Sankalp Gulati and Preeti Rao
Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai 40076, India
Email: {kinisujeet, sankalpg, prao}@ee.iitb.ac.in

Abstract—**The automatic classification of musical genre from audio signals has been a topic of active research in recent years. Although the identification of genre is a subjective task that likely involves high-level musical attributes such as instrumentation, style, rhythm and melody, low-level acoustic features have been widely applied to the automatic task with varying degrees of success. In this work, we consider the genres of the music of northern India, in particular the devotional music sub-genres of bhajan and qawwali. Both are rooted in the framework of North Indian classical music and are similar in the sense of serving the identical socio-cultural function even if for different religious communities of the same region. Features representing timbre, as well as temporal characteristics in the form of tempo and modulation spectra of timbral features, are shown to be potentially effective discriminators as seen by classification experiments performed on a database of excerpts drawn from the two music genres.**

I. INTRODUCTION

North Indian classical music (known also as Hindustani classical music) is characterized by melodic and rhythmic structures based on well-founded principles. Several distinct categories of Indian vocal music derive essentially from this formal framework with individual, more local, regional and socio-cultural influences. The well-known classical vocal styles of dhrupad, khyal and thumri, among others, coexist with popular and devotional music genres [1, 2, 3]. Devotional music constitutes a prominent category whose performance is an important component of religious rituals across the country and all its religions. Bhajan and qawwali are two prominent sub-genres of North Indian devotional music, belonging to the Hindu and Islamic religious communities respectively. They are associated with similar socio-cultural settings involving the musical rendition of the corresponding religious texts in a manner that enhances audience participation and induces a collective spiritual fervor.

Although the qawwali song has its origins in 8th century Persia, the form as known today is heavily infused with Indian musical traditions. Both bhajan and qawwali share the tonal framework, including melodic ornamentation, characteristic of Hindustani vocal music. A conspicuous aspect of both musical forms is the strong rhythmic character which serves to reinforce the uttered religious texts through its regularity and repetitiveness thus effecting spiritual arousal and audience participation. The rhythmic forms used in bhajan and qawwali belong within the tala framework of Hindustani classical music [3]. In this work, we consider the distinctive features of the two mentioned devotional genres (which are indeed easily distinguishable acoustically by nearly anyone belonging to the Indian subcontinent) with a view to developing computable features to automatically classify audio segments belonging to one of the two.

Automatic genre classification from audio has been an area of active research due to its importance in music information retrieval systems. Low-level acoustic features have been widely applied to this task with varying degrees of success. Features based on the short-time signal spectrum as well as their temporal variation over longer durations have proved useful for audio classification involving a large number of music genres [4]. On the other hand, higher level music based features derived from pitch organization and rhythm are of great interest due to their musicological relevance. Recently timbral and percussion descriptors were used to classify polyphonic audio as Western or non-Western music [5]. Exploiting such high level features for MIR however is limited by the achievable quality of polyphonic transcription of the audio.

In this work, we explore the performance of timbral features, and associated modulation spectrum features, extracted from the audio signal in terms of the obtained separability of the two classes of Hindustani devotional music. The features are then evaluated on a data set via cross-validated classification error from train-test experiments using different classification frameworks.

II. ACOUSTIC DESCRIPTION

Both bhajan and qawwali performances are characterized by solo singing possibly alternating with chorus in the repetitive segments of the text. The accompanying instruments include the harmonium providing melodic accompaniment to the singing voice and to the chorus. Percussive instruments provide the rhythm continuously in the background. Bhajan songs employ tabla or dholak and cymbals while qawwali songs reinforce the dholak accompaniment with the clapping of hands. The presence of chorus is always a part of the qawwali song. While qawwali singers are invariably male, bhajans are likely to be sung equally often by female singers. The melody follows the raga system of Hindustani classical music in both sub-genres. The singing style of qawwali singers is vigorous and stressed, invoking excitement versus the relative calmness of bhajan singing.

Both genres have highly regular rhythms, reinforced by the percussive accompaniment adhering to the strongly periodic and hierarchic temporal framework of Hindustani tala [3]. The rhythmic patterns, in terms of basic sequences and timing of percussion strokes, tend to be different. The occupied tempo ranges are different but overlap to an extent as can be seen in Figure 1. The tempo of a segment of bhajan or qawwali, as indicated by the periodicity of the perceptually most salient beat of the music, is easily detected manually by listening to the polyphonic audio.

III. Database

A total of 238 30-second duration excerpts equally distributed between two genres were extracted from the commercially available CDs. Bhajan excerpts are extracted from 109 different songs, belonging to 33 different artists, and that of qawwali from 103 songs, by 15 artists, ensuring a variety of compositions. The performances are typically of several minutes duration. The selected 30-second excerpt was obtained from approximately the middle of the performance with nearly constant tempo. The selected audio clips were filtered to limit the bandwidth to 8 kHz (considered sufficient for the subsequent processing). The tempo of each 30-second excerpt was extracted manually and automatically, as described next.

A. Tempo Estimation

Although a number of automatic tempo trackers have been proposed [6, 7], tempo estimation of music is an area of active research. A ground truth estimate of the tempo is therefore obtained for each song excerpt of our database for possible use in the subsequent feature extraction stage. The tempo of each clip is marked manually by a musician tapping along with the clip such that it corresponds to the most salient pulsation in the audio and is in accordance with the speed of the song. Any octave ambiguity in tapping rate is resolved by comparing the speed of the song with other songs and accordingly deciding the final beats per minute (BPM) value. In order to investigate the weaknesses of tempo trackers, if any, on the music of interest here, well-known, available automatic beat trackers based on detected onsets and pattern regularity are applied to obtain the automatically estimated tempo [6,7].

Figure 1 shows the distribution of the ground-truth tempos for the each class in the database. We note that the tempo range occupied by qawwali songs is shifted to the higher side as compared to the one by bhajan songs with limited overlap. The range of tempos observed would characterize the music as belonging to the “slow” to “medium” range in the context of Hindustani music. The automatically detected tempo estimates were evaluated for accuracy by considering a tempo estimate within 5% of the ground truth (GT) to be correct. Table 1 summarises the performance in terms of types of error. The column n-octave error, with n=0, shows number of correctly estimated tempo clips. Ellis’s tempo extractor (EL) appear to work well for qawwalis (101 correctly estimated tempos), but for bhajans, it fails almost half of the times. With Dixon’s tempo extractor (DX), only total 60 clips are correctly estimated. It generates +1 octave errors most often. Also we see that the tempo is detected very accurately on the qawwali data but not so on the bhajan songs. This difference in accuracy for the two genres may be attributed to the more complex rhythm patterns of bhajan as well as the occurrence of several slow tempo (< 60 BPM) pieces in the bhajan dataset.

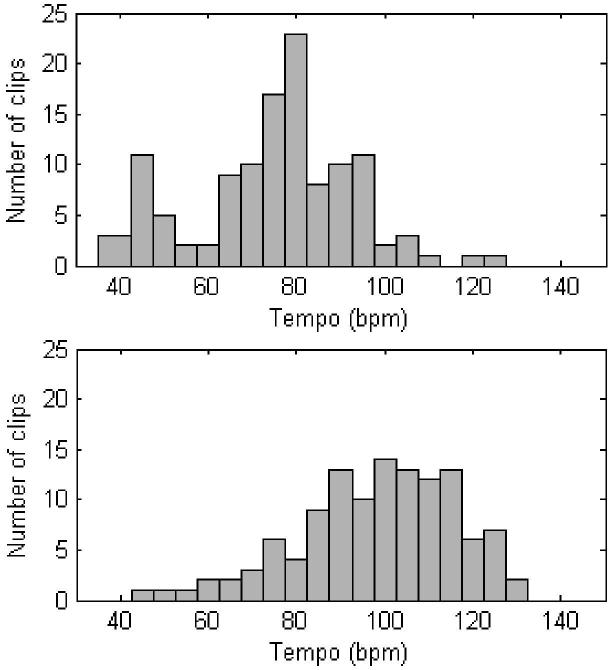

Figure 1. Tempo distribution for (a) bhajan songs and (b) qawwali songs of the database

IV. Feature Extraction

From the observed musical characteristics of bhajan and qawwali songs as presented in Sec. 2, it appears that features that capture the timbral texture, or rhythmic patterns in terms of temporal evolution of features, are likely to provide for sufficient separability of the two sub-genres in acoustic space. We discuss next the design of such features accompanied by methods for their extraction from the audio signal.

A. Timberal Features

Differences in timbre arise from the different instrumentation and the different styles of singing. Qawwali songs tend to be more dominated by chorus portions relative to bhajan songs. While the harmonium is the main melodic accompaniment in both genres and is often played relatively loud in qawwali songs, the percussion instruments differ in the more extensive use of hand clapping in qawwali. Both genres use the tabla or dholak with the addition of cymbals in bhajan songs. The audio thus has a complex structure due to the dense polyphony (co-occurring sources as well as rapidly time-varying sounds from the percussion). Features that capture the local spectral shape would be influenced by the overall texture of the polyphony. Temporal features that capture the cyclical variation of the timbre could potentially discriminate the rhythmic patterns including accented events in singing, and the sequence and timing of percussive strokes.

TABLE I. PERFORMANCE OF AUTOMATIC TEMPO ESTIMATORS [6,7] ON BHAJAN AND QAWWALI SONGS

Tempo Estimator	**Class**	**Octave Error**			
		n=-1	***n=0***	***n=1***	***other***
Ellis	Bhajan	2	46	50	21
	Qawwali	0	101	18	0
Dixon	Bhajan	0	5	74	40
	Qawwali	0	55	48	16

We consider short-time acoustic features that capture the shape of the spectral envelope over the bandwidth 0 to 8 kHz. These include the widely used Mel-frequency cepstral coefficients (MFCCs) and other measures of local frequency concentration such as spectral centroid, bandwidth, average energy, spectral roll-off, and zero-crossing rate (ZCR). The specified low-level features are computed as in standard practice from the waveform samples in each analysis window [8]. The duration of the analysis window is dictated by the assumed homogeneity of the signal properties within the window and is fixed in the present implementation at 23 ms. The characteristics of the feature vector over a longer term are obtained by feature summaries computed across texture window duration (typically on the order of a few seconds). The texture windows are non-overlapping. The feature summary could comprise simply the mean and variance of the feature across the texture window. On the other hand, a more detailed representation of the temporal variation of a feature over the texture window duration can be obtained via envelope modulation rates computed over different frequency bands of the modulation spectrum [4].

B. *Implementation*

Low level features are extracted from each analysis window of duration 23 ms with 50% overlap between windows to obtain a feature vector every 11.5 ms. The following features are used in the present work: the first 5 MFCCs, denoted as 5-MFCC, capturing the coarse shape of the frame-level magnitude spectrum. Other features that represent specific spectral characteristics are: zero crossing rate, spectral centroid, spectral bandwidth, spectral roll-off and short time average energy. The dimensionality of the feature vector is 10. For more reliable classification, feature parameters across analysis windows are grouped together to form a texture window of duration 3 s (more exactly, 2.97 s). Feature summaries for these larger segments are then computed from the frame-level feature sequences in one of two ways: mean-variance (MV) or envelope modulation (EM), to obtain eventually one feature vector per texture window.

The MV characterisation provides a vector of dimension 20 (feature means and feature variances) for each texture window duration. To obtain the EM features, the spectrum of the temporal trajectory of each of the frame-level features is computed over the texture window duration. The EM feature vector of dimension 40 is then made up of this modulation spectrum magnitude at DC, and the sum of magnitudes in the bin ranges corresponding roughly to the frequency bands 1-2 Hz, 3-15 Hz and 20-43 Hz. These bands approximately correspond to musical beat rate, speech syllabic rate and perceptual roughness respectively [4].

Since estimated tempos (either manual or automatic) are available for all the excerpts in the database, we also consider defining the texture window duration in a tempo dependent manner. With the texture window containing the same integer number of rhythm cycles across tempos, it is expected that the modulation features will indeed represent the rhythm-cycle-level pattern in a tempo normalised manner. The computation is carried out by varying the frame rate of the analysis windows within the tempo-dependent texture window so as to obtain the same fixed number of temporal sequence samples (256) in each case. In the present implementation, the texture window duration has been selected to cover 4 beat cycles at the specified tempo.

V. CLASSIFICATION EXPERIMENTS

The comparative performances of the feature sets for the genre classification task can be obtained via train-test cross-validation experiments on the available database. The generalizability of such a performance measurement is limited however due to the size of the database. Classification experiments are carried out separately using a Gaussian Mixture Model (GMM) classifier and a Support Vector Machine (SVM) classifier. The latter is a discriminative classifier unlike the former. We apply the above methods to measure the performance of the features proposed in Sec. 3 using the available database. Both mean-variance (MV) and envelope-modulation (EM) feature vectors are tested. Since ground-truth tempo is also available along with automatically estimated tempo for all the database songs, it is used to obtain tempo-dependent (TD) in addition to the tempo-independent (TI) texture windows. Further, classification performance is measured both at the texture-window segment level as well as at song level.

A. *Cross-validated Classification Accuracy*

Train-test classification is performed at the token-level (i.e. texture window level) and at the song level in 10-fold cross validation. First, the entire dataset of songs is divided into ten disjoint subsets, then nine of these are used to train the classifier, and remaining one set is used for testing. This is repeated nine times with different testing sets. The GMM classifier and SVM classifier are individually tested for the classification task. The Cluster [9] and libSVM [10] library packages are used for training the GMM and SVM classifiers respectively. The optimal order full-covariance matrix GMM is determined automatically for each class by the minimum description length (MDL) criterion which obtains a trade-off between training data likelihoods and classifier parameter dimensionality [9]. In case of the SVM classifier radial basis function kernel is used and classifier parameters C and γ are varied to obtain the best classification results using a grid search approach

Each token in the testing fold is classified by both GMM and SVM classifiers. Then a class label is assigned to an entire song clip using the majority rule applied to the classified tokens corresponding to that song (with aggregated probability across tokens applied to resolve any ambiguity).

B. Results

In Figure 1 we can see that bhajan and qawwali classes dominate separate tempo ranges; consequently tempo is one of the key features for classification. Table II shows results of the song-level classification with GMM classifier based on tempo alone. It can be seen that, with ground truth tempo the clips are identified well above chance as expected. Effect of more number of correctly identified clips by Ellis's tempo estimator is reflected in its higher classification accuracy than that of Dixon's tempo estimator. Classification performance with tempo feature may be improved by adding timbral features as we investigate next.

Table III(a) shows classification accuracy for GMM classifier with randomly selected non-overlapping training and testing dataset. The row TI 3S represents performance tempo-independent texture window of length 3 s, while other three rows are for tempo-dependent texture window with each of the three tempo detectors mentioned in Sec. 3. To remove the possible biasing of the trained models by specific artists' recording characteristics (rather than by the characteristics of the genre), artist filtering of the dataset is done. Artist filtering is carried out by ensuring that the artists represented in the training set and testing set are non-overlapping. Table III(b) shows summary of results after artist filtering for the GMM classifier. It is observed that the classification performance decreases with artist filtering as would be expected. The higher accuracies with EM summary over the MV summary indicate that the temporal evolution of timbre features plays an important role in the genre discrimination. The differences in temporal patterns may be attributed to rhythm pattern differences in the two genres.

Unlike the tempo-independent features, the tempo-dependent (TD) features, due to the inherent time normalization, lack the explicit description of the tempo. To compensate for this, the tempo of the clip was tagged on to the TD feature vector. The resulting performances are shown in Table IV(a). We see that the accuracies with ground-truth (GT) tempo included improve across the 3 feature sets with respect to the same feature sets in Table III(b). In the case of automatically detected tempo (EL and DX), the improvement is not as consistent which may be attributed to the unreliability of the tempo value.

Finally, the feature sets of Table IV (a) are tested in a different classifier framework, namely the SVM classifier with the results shown in Table IV (b). The classifier parameter values of C and γ are chosen so as to maximize the 5 fold cross-validated accuracy over training set. The optimization is carried out using a grid search. We note large improvements in classification accuracy with respect to GMM classifier accuracy with the same feature sets. With the SVM classifier, the largest dimensioned feature set (MV+EM) seems to achieve consistently higher accuracies over the other two feature sets.

TABLE II. CLASSIFICATION ACCURACY USING TEMPO AS THE FEATURE FOR INDIAN DEVOTIONAL MUSIC DATASET

Tempo Estimator	**Accuracy (%)**
Ground Truth	76.89
Ellis	71.42
Dixon	62.61

TABLE III. SUMMARY OF RESULTS FOR GMM CLASSIFIER WITH (A) RANDOMLY CHOSEN TRAINING SET (B) WITH ARTIST FILTERED TRAINING SET

Texture window type	**Clip Accuracy (%)/Token Accuracy(%) for types of feature summaries**		
	MV	***EM***	***MV+EM***
TI-3S	92.02/85.84	93.70/87.85	94.54/87.89
TD-GT	92.86/85.19	90.76/86.00	89.92/86.16
TD-EL	92.86/85.92	89.92/85.64	89.50/85.67
TD-DX	90.76/81.83	88.24/82.88	91.18/83.30

(a)

Texture window type	**Clip Accuracy (%)/Token Accuracy(%) for types of feature summaries**		
	MV	***EM***	***MV+EM***
TI-3S	84.45/80.04	90.76/84.66	90.34/84.20
TD-GT	87.39/80.93	86.97/82.27	86.13/81.94
TD-EL	86.55/79.53	87.39/83.19	85.29/82.63
TD-DX	85.29/77.44	87.82/81.76	86.97/80.29

(b)

TABLE IV. SUMMARY OF RESULTS WITH TIMBRAL FEATURES AND TEMPO AS A FEATURES WITH ARTIST FILTERED TRAINING SET FOR (A) GMM CLASSIFIER (B) SVM CLASSIFIER

Texture window type	**Clip Accuracy (%)/Token Accuracy(%) for types of feature summaries**		
	MV	***EM***	***MV+EM***
TD-GT	86.55/81.66	87.82/82.67	86.97/82.43
TD-EL	87.82/82.44	86.55/83.59	86.97/83.10
TD-DX	84.87/77.86	86.55/81.06	86.13/78.61

(a)

Texture window type	**Clip Accuracy (%)/Token Accuracy(%) for types of feature summaries**		
	MV	***EM***	***MV+EM***
TD-GT	89.92/85.76	91.18/86.32	92.86/87.50
TD-EL	89.92/85.79	92.44/87.22	92.44/87.84
TD-DX	88.66/84.38	91.60/85.43	92.86/87.02

(b)

An analysis of the misclassifications across feature sets and classifiers revealed that bhajans misclassified as qawwalis by timbral features corresponded generally to female singers. The confusion probably arises from timbral similarity with qawwali songs rendered with high-pitched, stressed male voices. For qawwalis misclassified as bhajans, it was observed that vigorous drumming and handclapping were absent, and so was the stressed singing style and heavy chorus.

VI. CONCLUSIONS

The task of finding computable acoustic features that capture the distinctions between audio of two genres of North Indian music was considered. The two genres, bhajan and qawwali, are prominent categories of devotional music and match to some extent in several acoustic aspects as well. Both comprise of solo and chorus singing with similar instrumentation and strong rhythm that is explicitly manifested in the percussive accompaniment. The differences in the texture of the polyphony are captured by low-level acoustic features describing the short-term spectral envelope shape. It is observed, from the classification experiments reported that the temporal behavior of the low-level features as computed from estimated tempo of the song together with short-time modulation spectra are useful in discriminating between audio from the two classes. A by-product of the study is the finding that available automatic tempo estimators show relatively poor accuracies on this music.

The results of this study can usefully contribute to a hierarchical genre classification system for Indian music where the features are tailored to the sub-genres at every level of the hierarchy. Further, the work indicates the potential of a larger study of audio descriptors including high-level musicology based attributes for Indian classical music. Signal processing methods for reliable feature extraction from audio including methods for more accurate automatic tempo detection for different music genres are suggested for future work.

REFERENCES

[1] R. Qureshi: Sufi music of India and Pakistan: sound, context, and meaning in Qawwali, Cambridge University Press, 1986.

[2] Chandra, David, Music of India. [Online]. http://chandrakantha.com

[3] M. Clayton: Time in Indian music: rhythm, metre, and form in North Indian rãg performance, Oxford University press Inc., New York, 2000.

[4] M. F. Mckinney and J. Breebaart, "Features for Audio and Music Classification," Proceedings of the International Symposium on Music Information Retrieval (ISMIR), pp. 151-158, 2003.

[5] M. Haro, P. Herrera, "From Low-level to Song-Level Percussion Descriptors of Polyphonic Music," 10th International Society for Music Information Retrieval Conference (ISMIR), pp. 243-248, 2009.

[6] Daniel P.W. Ellis, "Beat Tracking by Dynamic Programming," Journal of New Music Research, Special Issue on Beat and Tempo Extraction, Vol. 36, No. 1, pp. 51-60, 2007.

[7] S. Dixon, "Automatic Extraction of Tempo and Beat from Expressive Performances," Journal of New Music Research, Vol 30, No. (1), pp. 39-58, 2001.

[8] G. Tzanetakis, P. Cook, "Musical Genre Classification of Audio Signals," IEEE Transactions on Speech and Audio Processing, Vol. 10, No. 5, pp. 293-302, 2002.

[9] C. A. Bouman, Cluster: An Unsupervised Algorithm for Modeling Gaussian Mixtures. [Online]. http://www.ece.purdue.edu/~bouman, 1997.

[10] C-C. Chang and C-J. Lin: LIBSVM : a library for support vector machines. [Online] http://www.csie.ntu.edu.tw/~cjlin/libsvm