Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 45 (2015) 186 – 196

Procedia Computer Science

International Conference on Advanced Computing Technologies and Applications (ICACTA-2015)

Evaluation of Music Features for PUK Kernel based Genre Classification

Santosh Chapaneri*[a], Renia Lopes[b], Deepak Jayaswal[c]

[a]*Assistant Professor,* [b]*ME Student,* [c]*Professor, Department of Electronics and Communication Engg.*
St. Francis Institute of Technology, University of Mumbai, India

Abstract

Since music conveys as well as evokes a wealth of emotions in a listener, there has been tremendous research and commercial development to automatically organize music using smart machine learning techniques. In this work, various features are extracted from the music signal for an effective representation to aid in genre classification. The feature set comprises of dynamic, rhythm, tonal, and spectral features comprising a total of 144 features. The size of feature set is further reduced to 39 features using correlation-based feature selection mechanism to remove the correlated features. Support vector machine classifier is used to train the genre classification system with a flexible Pearson Universal Kernel (PUK) that can adapt its behavior to various functions (from linear to Gaussian). The reduced feature set, consisting mostly rhythm and spectral features, significantly outperforms the complete feature set leading to an accuracy of 82% for classifying 5 genres.

Peer-review under responsibility of scientific committee of International Conference on Advanced Computing Technologies and Applications (ICACTA-2015).

Keywords: Music Genre; Pattern Classification; SVM; Pearson Universal Kernel

* Corresponding author.
E-mail address: santoshchapaneri@gmail.com

Peer-review under responsibility of scientific committee of International Conference on Advanced Computing Technologies and Applications (ICACTA-2015).
doi:10.1016/j.procs.2015.03.119

1. Introduction

Music plays an important role in modern information technology oriented society due to abundant availability of digital music on Apple's iTunes store, Amazon store, etc. hosting more than 2 million songs. The research area of Music Information Retrieval (MIR) has seen significant work in the past decade ranging from cover song detection, music recommendation, beat tracking, to audio fingerprinting. In this work, we focus on automatic music genre classification, since music genre is an important music descriptor that can be used to automatically organize large collections of music into playlists as per user's mood. Commercial applications such as Shazam, Pandora, Last.FM, and Spotify have been developed to aid this goal in the recent years. Applications are envisaged for smartphones where the music collection can be automatically grouped as per genre to enrich user's listening experience. From machine learning perspective, the task of genre classification is non-trivial due to the fact that the definition of genres are still ill-defined, as well as the music signals are time-varying complex signals requiring processing of high-dimensional values. Thus, a reduced feature representation is required for most work in MIR, followed by feature classification stage for genre classification problem.

1.1. Related Work

Mel-frequency cepstral coefficients (MFCC) features were extracted and Gaussian Mixture Model (GMM) classifier was used by Flexer[1] followed by a statistical evaluation of the genre classification experiment. Modulation spectral contrast features using octave music scale were proposed by Lee et al[2] as a modification of the conventional octave spectral contrast features. Musical surface and rhythm features were extracted by Tzanetakis[3] to develop GenreGram user interface for easy visualization of different genre songs. Support vector machine (SVM) with different kernels were used by Xu et al[4] for genre classification to achieve 93% accuracy; however, they used only 60 total samples with 15 samples for 4 genres, which is too small database in our opinion for training and testing the system. Daubechies Wavelet coefficient histograms were used by Tao et al[5] as discriminative features for 10 genres, along with the standard timbral and rhythm features. Block-level feature comprising of spectral contrast pattern, and correlation pattern among others were fused by Seyerlehner et al[6] using distance space normalization for various music genre datasets. Instance selection process was used by Miguel et al[7] prior to SVM classifier training resulting in faster classification with about 59% classification for 10 genres from Latin Music Dataset[8]. North Indian devotional music genres were classified by Sujeet et al[9] using tempo and modulation spectra of timbral features. Chromagram related features were used by Boonmatham et al[10] based on diatonic and chromatic scale for Thai music genre classification. Novel features using fluctuation patterns were proposed by Pampalk[11] using a sonogram representation of the music. The concept of compressive sampling was used by Chang et al[12] for both short-time and long-time music features, achieving high accuracies similar to non-negative tensor factorization approach[13] for 10-genre classification task. Rhythm patterns and rhythm histogram features were proposed by Lidy et al[14] as a means of psycho-acoustic transformations of the music signal. A self-adaptive harmony search technique was proposed by Huang et al[15] for selecting the most appropriate features from the set of conventional features. Aucouturier and Pachet[16] rightly claimed that it is only with efficient content management techniques (such as automatic genre classification) that millions of music songs from industry can be distributed to millions of users. However, there seems to be no clear identification of music features relevant in classifying various genres so far. In this work, we extract music features from 4 broad categories: dynamic, rhythm, tonal, and spectral to check the efficacy of each feature set for a 5 genre classification task.

2. Music Features

Various features are extracted for each music signal, broadly categorized into computational features and perceptual features. The computational features do not convey any specific meaning about the genre of the song, but only describe the mathematical analysis of music signal. They are further categorized into dynamics and spectral feature set. The perceptual features mathematically represents the music properties based on the human auditory system. They are further categorized into rhythm and tonal feature set. In the following, we use the notation of x to

represent a discrete music signal sampled at frequency f_s with duration T in seconds and N in samples, and X the spectrum.

2.1. Dynamic Features

The dynamic feature set comprises of root-mean-square (RMS) energy, zero-crossing-rate (ZCR), and low energy rate. The RMS feature, which represents the global energy of the music signal, is computed by taking the root average of amplitude square of all samples. The ZCR rate is an indicator of noisiness of the music signal, and is computed by counting the number of times a signal crosses zero: ($x(n\text{-}1) < 0$ AND $x(n) > 0$), OR ($x(n\text{-}1) > 0$ AND $x(n) < 0$), OR ($x(n\text{-}1) \neq 0$ AND $x(n) = 0$). The low energy rate feature conveys the information about the proportion of silence or near-silence in the music signal. The threshold is set as the average energy determined from the energy envelope of signal and the low energy rate is computed as the fraction of frames which have energy less than this threshold.

2.2. Rhythm Features

Rhythmic features play a dominant role for differentiating music genres and are based on the time intervals between various note attacks, onsets, duration of notes, tempo, etc. The rhythm feature set comprises of beat histogram features, pulse clarity, fluctuation patterns (FP) features, and onsets features. The beats in music are the steady pulses that drives music forward and provides a temporal framework for a music piece to which a human taps along when listening to the music.

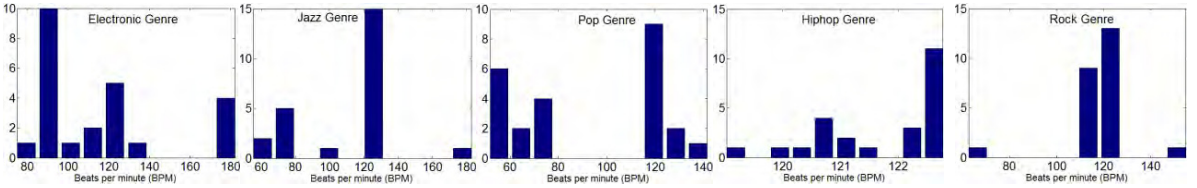

Fig. 1. Beat histograms of samples of electronic, jazz, pop, hiphop and rock genres (from left to right)

The tempo measured in beats per minute (BPM) is computed from the beat histogram (BH)[17]. The music signal is given to discrete wavelet transform (DWT) filter bank to decompose the signal into various octave bands. The envelope of each band is extracted and repeating events (pulses) are determined through auto-correlation. The peaks from the resulting periodogram are selected and converted into a beat histogram. Beat tracking is especially useful in applications such as DJ editing to match the tempo of two songs played on different decks. Fig. 1 shows the beat histogram for samples of 5 genres, where we observe a different distribution of the tempo measurement for these genres. Each bin indicates the relative strength of each rhythmic pulse within the music piece. The average tempo value is computed from the beat histogram, along with the beat sum (addition of all beats), strongest beat present in BH, and the strength of strongest beat.

Pulse clarity determines the strength of rhythmic periodicities and pulses in the music signal and conveys the information about how easily listeners can perceive the underlying pulsation in the music. It is estimated by the relative Shannon entropy from the auto-correlation function[18]. Fluctuation patterns (FP) describe the amplitude modulation of the loudness (dB) per each frequency band of the spectrogram[11]. The resolution of the modulation frequency is 30 in the range of 0 to 10 Hz, where the perception of fluctuation strength is maximum around 4 Hz and decreases gradually. The modulation frequencies are weighted according to a model of perceived fluctuation strength followed by Gaussian smoothing filter to enhance the patterns. The resulting FP is a two-dimensional matrix with rows indicating frequency bands, and columns indicating modulation frequencies, and each element indicates the fluctuation strength. The features computed from FP are the peak position, peak magnitude and centroid. Additionally, the maximum value (FP_Max), bass (FP_Bass), domination of low frequencies (FP_DLF), gravity (FP_G), and focus (FP_F) features are computed from the FP matrix, given by Eqs. (1 – 3). The bass is

computed as the sum of values in the two lowest frequency bands of FP. DLF is computed as the ratio between the sum of FP values in 4 highest and 3 lowest frequency bands. Gravity feature (FP_G) denotes the center of gravity of FP on the axis of modulation frequency, and usually models the effect of tempo, vibrato and tremolo in the music[11]. Focus feature (FP_F) denotes the distribution of energy in FP and is computed as the normalized mean of all FP values.

$$\text{FP\_Max} = \max(FP(:)); \quad \text{FP\_Bass} = \text{sum}(\text{sum}(FP(1:2,3:30))) \tag{1}$$

$$\text{FP\_DLF} = \text{sum}(\text{sum}(FP(1:3,:))) / \max(\text{sum}(\text{sum}(FP(9:12,:)))) \tag{2}$$

$$\text{FP\_G} = \frac{\sum_j j \sum_i FP(i,j)}{\sum_j \sum_i FP(i,j)}; \quad \text{FP\_F} = \text{mean}(FP(:)./\max(\max(FP(:)))) \tag{3}$$

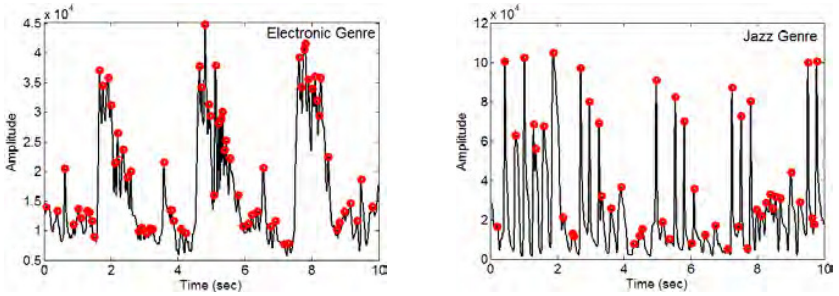

Fig. 2. Onset detection for electronic genre sample (left) and jazz genre sample (right)

Onsets are detected from the music signal, which represents the start times of perceptually relevant acoustic portions of the music. Onsets are determined with the following steps: a) the amplitude of the music samples is squared, b) windowing operation is performed for local analysis, c) differentiation and d) half-wave rectification is performed, to finally e) pick the peaks resulting in onsets. Fig. 2 shows the onsets detected for samples of electronic and jazz genre. The features determined from onsets are peak position, peak magnitude, attack time, and attack slope.

2.3. Tonal Features

Inspired by music theory, various features are computed from the pitch class profile (PCP) of the music signal. The tonal feature set comprises of chromagram peak position, peak magnitude, centroid, key clarity, key mode, and harmonic change detection function (HCDF) features. The 12 standard pitches in Western music are C, C#, D, D#, E, F, F#, G, G#, A, A#, and B. The chromagram is a spectrogram that shows the distribution of spectral energy among each of the 12 pitches by mapping the frequencies onto an octave music scale. To compute the chroma features, the audio signal is transformed to frequency domain using DFT followed by log-frequency filter-bank with linear center frequencies f_c on a log scale given by Eq. (4), where f_{min} is the minimum analysis frequency, β = bins per octave, Z = number of octaves, and k_{lf} = integer filter index between 0 to βZ-1. The chroma is determined by summing the amplitude values across octaves, resulting in chroma vector $C_f(b)$ given by Eq. (4), where X_{lf} is the log-frequency spectrum, z = integer octave index between 0 to Z-1, and b = pitch class (chroma index) between 0 to β -1.

$$f_c(k_{lf}) = f_{\min} \times 2^{\frac{k_{lf}}{\beta}}; \quad C_f(b) = \sum_{z=0}^{Z-1} \left| X_{lf}(b + z\beta) \right| \tag{4}$$

Fig. 3 shows the chroma histograms of the samples of 5 genres where we observe that electronic music mostly plays at selected few pitches but hiphop music contains significant contribution of all pitch classes. From the chromagram, the features extracted are peak position, peak magnitude, and centroid. Since there are 12 major scales and 12 minor scales in music, key clarity can be computed with cross-correlation of the chromagram to give the key strength associated with the best key candidate in the music signal. To estimate the modality of music (i.e. major or minor), the key mode is also computed by summing the key strength differences between all major keys and their relative minor keys. A 6-dimensional tonal centroid is computed, which corresponds to a projection of the pitch classes along circles of fifths, of minor thirds, and of major thirds[19], resulting in its flux denoted as harmonic change detection function (HCDF) ζ_n, given by Eq. (5). The transformation matrix Φ represents the basis of the 6-dimensional tonal space[19]. Formally, HCDF indicates the overall rate of change of the smoothed tonal centroid of the music signal.

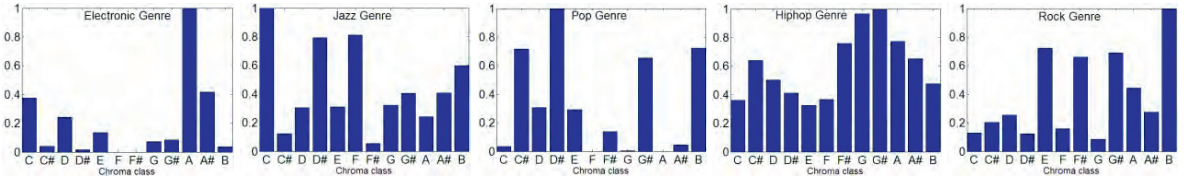

Fig. 3. Chroma histograms of samples of electronic, jazz, pop, hiphop and rock genres (from left to right)

$$\zeta_n(d) = \sum_{b=1}^{\beta} \Phi(d,b) C_n(b); \qquad \zeta_n = \sqrt{\sum_{d=0}^{5} \left[\zeta_{n+1}(d) - \zeta_{n-1}(d) \right]^2} \tag{5}$$

2.4. Spectral Features

Finally, features are extracted from the spectrum of music signal, and the spectral feature set comprises of standard statistical features including spectral centroid, spectral spread, spectral skewness, spectral kurtosis, spectral flatness, spectral flux, spectral brightness, spectral entropy, spectral roll-off (at 85% and 95%), spectral roughness, spectral irregularity, and spectral compactness, along with Mel-frequency cepstral coefficients (including delta and double-delta features).

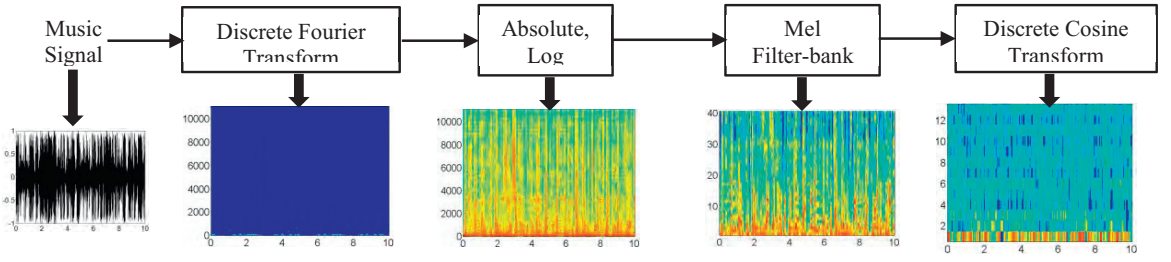

Fig. 4. Spectral computation of music signal leading to MFCC

Fig. 4 illustrates the various steps in computation of several spectral features leading to Mel-frequency cepstral coefficients (MFCC). The DFT is applied to windowed frames (of small duration in milliseconds) of music signal followed by log scaling, resulting in the power spectrum or spectrogram for the entire signal. The spectral centroid (first moment) refers to the center of mass of the power spectrum and perceptually indicates how “dark” or “bright”

the sound of music feels like. The spectral spread (second central moment) refers to the standard deviation of the spectrum. The spectral skewness (third central moment) refers to the measure of symmetry of the spectral distribution. The spectral kurtosis (fourth central moment) refers to the peakiness of the spectral distribution. The spectral flatness indicates whether the spectrum distribution is smooth/flat or spiky, and is computed as the ratio between the geometric mean and the arithmetic mean of the spectrum. The spectral flux refers to variability of the spectrum between frames and is computed as the distance between the spectra of each successive frames. The spectral brightness refers to amount of energy above a specific cut-off frequency. The spectral entropy refers to relative Shannon entropy, and is computed using the normalized power spectral density to determine the probabilities. If the probability distribution is towards flat, then the spectral entropy will be towards maximum, and minimum for a peaked distribution. The spectral roll-off refers to the frequency below which p% of the total spectral energy is accounted for (p = 85 and 95, in this work). The spectral roughness refers to the sensory dissonance due to beating phenomena between close frequency peaks, according to Plomp-Levelt curve[20]. The spectral irregularity (S_IR) refers to the degree of variation of the successive peaks in the spectrogram, and is computed as the sum of square of the difference in spectral amplitude, given by Eq. (6) where $X(k)$ denotes the spectrum of each frame using N point FFT. The spectral compactness (S_CT) refers to noisiness of the music signal and is computed by Eq. (7) by summing over all frequency bins of FFT spectrum.

$$\mathrm{S\_IR} = \sum_{k=1}^{N} \big(X(k) - X(k+1)\big)^2 \Big/ \sum_{k=1}^{N} X^2(k) \tag{6}$$

$$\mathrm{S\_CT} = \sum_{k=2}^{N-1} \left(\log(X(k) - \frac{\log(X(k-1)) + \log(X(k)) + \log(X(k+1))}{3} \right) \tag{7}$$

For the computation of MFCC features, the spectrum is converted to a perceptually motivated Mel frequency scale. Discrete cosine transform is applied to the Mel-filtered log spectrum from which only the first 13 de-correlated coefficients are extracted, as illustrated in Fig. 5. Further, to consider the temporal evolution of MFCC features, delta (DMFCC) and double-delta (DDMFCC) features are computed, referring to velocity and acceleration[21].

Table 1. Features extracted from music signal

Feature Set	#	Features
Dynamics	1-2	RMS energy (*m*, *std*)
	3-4	ZCR (*m*, *std*)
	5	Low Energy Rate (*m*)
Rhythm	6-7	Tempo (*m*, *std*)
	8-9	Beat Sum (*m*, *std*)
	10-11	Strongest Beat (*m*, *std*)
	12-13	Strength of Strongest Beat (*m*, *std*)
	14	Pulse Clarity (*m*)
	15-16	FP Peak (pos, mag) (*m*)
	17	FP Centroid (*m*)
	18-22	FP Statistics (Maximum, Bass, DLF, Gravity, Focus)
	23-24	Onsets Peak (pos, mag) (*m*)
	25-28	Rhythm Attack (time, slope) (*m*, *std*)
Tonal	29-32	Chromagram Peak (pos, mag) (*m*, *std*)
	33-34	Chromagram Centroid (*m*, *std*)
Spectral	41-42	Centroid (*m*, *std*)
	43-44	Spread (*m*, *std*)
	45-46	Skewness (*m*, *std*)
	47-48	Kurtosis (*m*, *std*)
	49-50	Flatness (*m*, *std*)
	51-52	Flux (*m*, *std*)
	53-54	Brightness (*m*, *std*)
	55-56	Entropy (*m*, *std*)
	57-60	Roll-off (85% and 95%) (*m*, *std*)
	61-62	Roughness (*m*, *std*)
	63-64	Irregularity (*m*, *std*)
	65-66	Compactness (*m*, *std*)
	67-92	MFCC (*m*, *std*)
	93-118	DMFCC (*m*, *std*)
	119-144	DDMFCC (*m*, *std*)

35-38	Key Clarity & Mode (*m*, *std*)
39-40	HCDF (*m*, *std*)

For the various features extracted, we compute the mean (*m*) and standard deviation (*std*) for a reduced description of each feature contour. Thus, overall 144 features are extracted from each music signal as shown in Table 1. Although the extracted 144 features are able to capture relevant aspects of the music signal, it is desirable to reduce the number of features to avoid the curse of dimensionality for the classifier. Further, it is desirable to remove any correlation between the extracted features, since the correlated features will mislead the classifier during training. The aim of feature selection mechanism is to maximize inter/intra class ratio, i.e. maximize the feature variance between various genres and minimize the feature variance for same genre. We use Correlation-based Feature Selection (CFS) technique[22] to reduce the feature set size, since it can determine only those useful features that are correlated with or predictive of the genre class. The Best First search method is used for CFS that searches the space of feature subsets with greedy hill climbing technique. At each stage, a local optimum choice of feature subset is found to determine the globally optimum feature set. From the 144 extracted features, CFS resulted in 39 features as shown in Table 2, and these 39 features are given as representation of each music signal to the classifier. Fig. 5 shows the box plots of selected features across 5 genres, where each box has a line at the lower quartile, median, and upper quartile values. The whiskers extend to the most extreme feature value within 1.5 times the inter-quartile range. Outliers are feature values beyond the whiskers and marked with '+'. We observe that hiphop genre has distinguishing features relative to other genres, and the median values of features are mostly distinct for all genres.

Table 2. Reduced features by CFS feature selection technique

Feature Set	#	Features
Dynamics	1	RMS energy (*std*)
	2	ZCR (*std*)
Rhythm	3	Beat Sum (*m*)
	4	Strongest Beat (*std*)
	5	Strength of Strongest Beat (*std*)
	6	FP Centroid (*m*)
	7	Pulse Clarity
	8-12	FP Statistics (Maximum, Bass, DLF, Gravity, Focus)
Tonal	13	Chromagram Peak (pos) (*std*)
	14	Key Clarity (*m*)
	15-16	Key Mode (*m*, *std*)
Spectral	17	Centroid (*std*)
	18	Skewness (*m*)
	19	Entropy (*m*)
	20	Flatness (*m*)
	21-22	Roughness (*m*, *std*)
	23	Irregularity (*m*)
	24	Compactness (*m*)
	25-29	MFCC 1 (*m*), 2 (*std*), 4 (*m*), 5(*m*), 9 (*std*)
	30-34	DMFCC 1, 2, 4, 10, 13 (*std*)
	35-39	DDMFCC 2, 4, 6, 10, 12 (*std*)

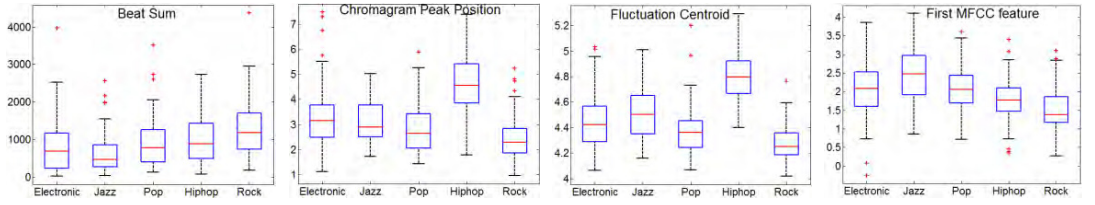

Fig. 5. Box plots of selected features: Beat Sum, Chromagram Peak Position, Fluctuation Centroid, and 1st MFCC (from left to right)

3. Feature Classification using PUK Kernel

The majority of music classification tasks such as genre classification, music segmentation, and chord recognition typically use k-nearest neighbor (KNN), support vector machines (SVM), and Gaussian mixture model (GMM). SVMs have been shown to perform best for music related features as input[5]. Thus, we use the supervised SVM classifier using the implementation of sequential minimal optimization (SMO) algorithm, implemented in WEKA (Waikato Environment for Knowledge Analysis) data mining toolkit[23]. In its standard form, SVM is a binary classifier that separates the features of two classes with an optimal hyperplane such that a) the largest possible instances of same class belong to same side of the hyperplane, and b) the distance of either class features from the hyperplane is maximum. The output class prediction of SVM for an unseen test instance $\mathbf{z}$ is either $y+ = 1$ or $y- = -1$, given by the decision function $\text{pred}(\mathbf{z}) = sgn(\mathbf{wz} + b)$, where b is the bias term, and the hyperplane $\mathbf{w}$ is determined with the help of Lagrange optimization given by Eq. (8), subject to the constraints $0 \le \alpha_i \le C$, where C is the cost penalty parameter for outliers, $i = 1, 2, \ldots, l$, and $\text{sum}(\alpha \cdot y) = 0$. The kernel function K maps the feature vector $\mathbf{x}$ to a high-dimensional space where linear separation may be possible. Various kernels have been used in SVM that satisfy Mercer's condition: linear kernel, polynomial kernel, radial basis function (RBF) kernel, and sigmoid kernel. Typically, a k-fold cross validation is performed to optimally determine the kernel parameters as well as the cost parameter C.

$$w(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j); \qquad K(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{\left[1 + \left(\frac{2\sqrt{\left\| \mathbf{x}_i - \mathbf{x}_j \right\|^2} \sqrt{2^{(1/\omega)} - 1}}{\sigma} \right)^2 \right]^{\omega}} \tag{8}$$

In this work, we use the Pearson VII Universal Kernel (PUK)[24] which is a generic kernel adapting the behavior of various common kernels mentioned above. The choice of PUK kernel for SVM ensures that SVM will perform accurately even if the kernel function chosen is not appropriate. The choice of kernel parameters for PUK allows it to exhibit behavior of linear, higher-order polynomials as well as Gaussian kernels[24]. Furthermore, the PUK kernel K given by Eq. (8) satisfies the condition of being symmetric kernel matrix, and positive semi-definite. The parameters ω and σ control the Pearson-width (half-width) and the tailing factor of the peak, respectively. Thus, by varying these two parameters, flexibility can be achieved for the SVM to function properly. For the purpose of multi-class genres, one-vs.-one SVM technique is used where pair-wise classification is done using binary SVM and the final decision is taken with the help of majority voting. An internal 10-fold cross validation is done to optimally determine the SVM parameters via a grid-search technique.

4. Experimental Results

We evaluate the various features extracted above for classification of 5 popular genres, viz. Electronic, Jazz, Pop, Hiphop, and Rock, for 500 music samples (100 samples/genre) from the freely available Benchmark database[25]. Each sample is 10 second duration drawn from a random position of the corresponding song, encoded using mp3 at a sampling rate of 44.1 kHz. The dataset is divided into 60% training and 40% testing samples. The cross-validation is performed on the training samples to choose the optimum parameters for PUK kernel and all experiments are done using the reduced feature set from Table 2. Fig. 6 illustrates the classification accuracy (%) for individual feature sets mentioned in Sec. 2 for 5 genres, where we observe that the overall performance achieved by feature sets follow the trend: Spectral > Rhythm > Tonal > Dynamics. Better accuracies are obtained for hiphop genre due to its distinct beat patterns, and pop genre suffers poor performance because it is often confused with rock genre[26].

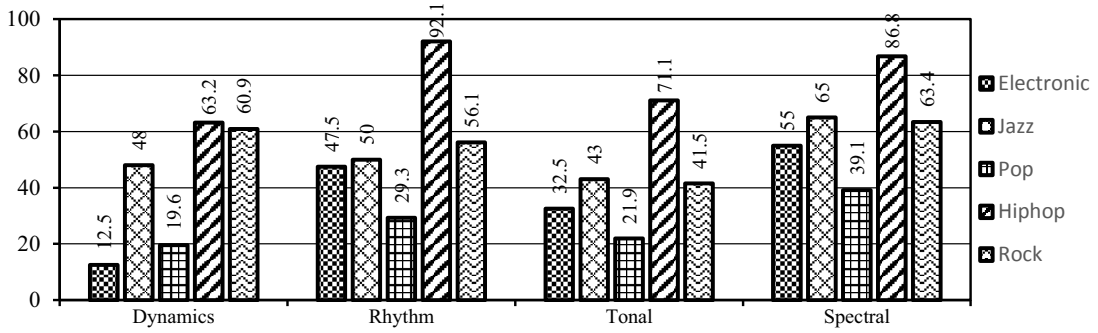

Fig. 6. Classification accuracy for individual feature sets (values in %)

We evaluate the performance of combinations of feature set to determine their contributions in maximizing the classification accuracy. Table 3 shows the accuracy of various feature sets for 5 genre classification. Using only 33 rhythm and spectral features, the accuracy is 78.5% with an improvement of 1.5% using additional 4 tonal features and a further improvement of 2% with additional 2 dynamic features. The maximum accuracy of 82% is obtained by considering all 39 features from all feature sets, whereas with taking all 144 features, the overall accuracy obtained is only 65%. The area under curve (AUC) of the receiver operating characteristics (ROC) is 0.84 for all 144 features, whereas AUC is 0.92 for the reduced feature set. The proposed system performs better than existing work (though on a different database) where the classification accuracy was reported to be 75.6%[1] and 62%[3] for 5 genre classification task. Table 4 shows the confusion matrices with precision and recall values obtained on 200 test samples by considering all features as well as reduced feature set. Due to dimensionality reduction by CFS technique, the overall accuracy, precision, and recall values are significantly increased. Using AdaBoost boosting technique with SVM classifier, the accuracy increased only marginally to 82.3%. From all the 5 genres, music samples of pop suffers most confusion due to unclear meaning and definition of the pop genre, whereas hiphop genre outperforms all other genres since it contains distinct intense rhythmic beats and rapping parts. Also, PUK kernel based SVM outperformed other classifiers for the music genre classification task under consideration: J48 Decision Tree (58%), k-Nearest Neighbors (65%), Random Forest (65%), Naïve Bayes (70%), SVM with Polynomial kernel (72.5%), Multi-Layer Perceptron (77.5%), SVM with radial basis function (RBF) kernel (78%), and SVM with PUK kernel (82%).

Table 3. Classification results for combination of feature sets (D: Dynamic, R: Rhythm, T: Tonal, S: Spectral)

Feature Sets	Accuracy	No. features
S + D	62.5 %	24
S + R	**78.5 %**	**33**
S + T	70 %	26
D + R	62 %	13
D + T	44 %	6
R + T	64.5 %	15

Feature Sets	Accuracy	No. features
D + R + S	78.5 %	35
D + T + S	74 %	28
D + R + T	67 %	17
R + T + S	**80 %**	**37**
D + R + T + S	**82 %**	**39**

Table 4. Confusion matrices for all 144 features (left), and reduced 39 features (right) (E: Electronic, J: Jazz, P: Pop, H: Hiphop, R: Rock)

	E	J	P	H	R	Recall
E	**23**	12		2	4	56.1
J	1	**32**	2	2	1	84.2
P	4	11	**15**	3	10	34.9
H	1	2		**32**		91.4
R	3	2	7	3	**28**	65.1
Prec.	71.9	54.2	62.5	76.2	65.1	

	E	J	P	H	R	Recall
E	**32**	2	3	1	2	80
J	4	**30**	3	2	1	75
P	2	2	**30**	4	3	73.2
H	1	1		**36**		94.7
R			4	1	**36**	87.8
Prec.	82.1	85.7	75	81.8	85.7	

5. Conclusions

In this work, 500 samples of 5 music genre (electronic, jazz, pop, hiphop, and rock) were obtained from the benchmark database[25], and music features were extracted for each in 4 categories: dynamic, rhythm, tonal, and spectral. Since no single feature is discriminative enough for accurate classification of multiple genres, various combinations were explored to maximize the accuracy. Dimensionality reduction technique such as CFS plays an important role in removing redundant features and improving the overall accuracy. The PUK kernel based SVM classifier outperformed other classifier techniques due to its flexibility to adapt to various functionalities (from linear to Gaussian) depending on the input feature data. The proposed music genre classification system is capable of classifying 5 genres with a maximum accuracy of 82% with features mostly from the rhythm and spectral category. For future work, genre relations can be enforced with multi-level hierarchy, since not all genres are equidistant (i.e. genres do not have a flat hierarchy). For example, soft rock songs are closer to pop songs than heavy metal songs. Also, higher-level dimensions such as temporal integration of features with multi-variate auto-regressive modelling[27] and Gaussian process regression[28] can be explored from an auditory perspective.

Acknowledgements

Parts of MIRToolbox[29], jAudio[30] and Music Analysis (MA) toolbox[31] were used in this work.

References

1. A. Flexer. Statistical Evaluation of Music Information Retrieval Experiments. J New Music Research 53:2, pp. 113-120, Jun 2006.
2. C. Lee, J. Shih, K. Yu, and J. Su. Automatic Music Genre Classification Using Modulation Spectral Contrast Feature. IEEE Intl. Conf. Multimedia and Expo, pp. 204-207, Jul 2007.
3. G. Tzanetakis, and P. Cook. Music Analysis and Retrieval Systems for Audio Signals. J American Society for Information Science and Technology, 55:12, pp. 1077-1083, Oct 2004.
4. C. Xu, N. Maddage, X. Shao, F.Cao, and Q. Tian. Musical Genre Classification using Support Vector Machines. IEEE Intl. Conf. Acoustics, Speech and Signal Processing, pp. 429-432, Apr 2003.
5. T. Li, M. Ogihara, and Q. Li. A Comparative Study on Content-Based Music Genre Classification. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 282-289, Aug 2003.
6. K. Seyerlehner, G. Widmer, and T. Pohle. Fusing Block-Level Features for Music Similarity Estimation. Intl. Conf. Digital Audio Effects (DAFx-10), pp. 1-8, Sep 2010.
7. M. Lopes, F. Gouyon, A. Koerich, and L. Oliveira. Selection of Training Instances for Music Genre Classification. IEEE Intl. Conf. Pattern Recognition, pp. 4569-4572, Aug 2010.
8. C. Silla, A. Koerich, and C. Kaestner. The Latin Music Database. Intl. Conf. Music Information Retrieval, pp. 451–456, Sep 2008.
9. S. Kini, S. Gulati, and P. Rao. Automatic Genre Classification of North Indian Devotional Music. IEEE National Conf. Communications (NCC), pp. 1-5, Jan 2011.
10. P. Boonmatham, S. Pongpinigpinyo, and T. Soonklang. Musical-scale Characteristics for Traditional Thai Music Genre Classification. IEEE Intl. Computer Science and Engineering Conf., pp. 227-232, Sep 2013.
11. E. Pampalk, A. Flexer, and G. Widmer. Improvements of Audio-Based Music Similarity and Genre Classification. Intl. Society for Music Information Retrieval Conf. (ISMIR), pp. 628-633, Sep 2005.
12. K. Chang, J. Jang, and C. Iliopoulos. Music Genre Classification via Compressive Sampling. Intl. Society for Music Information Retrieval Conf. (ISMIR), pp. 387-392, Aug 2010.
13. Y. Panagakis, and C. Kotropoulos. Music Genre Classification via Topology Preserving Non-negative Tensor Factorization and Sparse Representations. IEEE Intl. Conf. Acoustics, Speech and Signal Processing (ICASSP), pp. 249-252, Mar 2010.
14. T. Lidy, and A. Rauber. Evaluation of Feature Extractors and Psycho-Acoustic Transformations for Music Genre Classification. Intl. Society for Music Information Retrieval Conf. (ISMIR), pp. 34-41, Sep 2005.
15. Y. Huang, S. Lin, H.Wu, and Y. Li. Music Genre Classification Based on Local Feature Selection using a Self-Adaptive Harmony Search Algorithm. Elsevier Data & Knowledge Engineering vol. 92, pp. 60-76, Jul 2014.
16. J. Aucouturier, and F. Pachet. Representing Musical Genre: A State of the Art. Journal of New Music Research 32:1, pp. 83–93, Mar 2003.
17. G. Tzanetakis, G. Essl, and P. Cook. Human Perception and Computer Extraction of Musical Beat Strength. Intl. Conf. Digital Audio Effects (DAFx-02), pp. 257-261, Sep 2002.
18. O. Lartillot, T. Eerola, P. Toiviainen, and J. Fornari. Multi-Feature Modeling of Pulse Clarity: Design, Validation And Optimization. Intl. Society for Music Information Retrieval Conf., pp. 521-526, Sep 2008.
19. C. Harte, M. Sandler, and M. Gasser. Detecting Harmonic Change in Musical Audio. Proc. ACM Workshop on Audio and Music Computing Multimedia, pp. 21-26, Oct 2006.
20. Plomp, and Levelt. Tonal Consonance and Critical Bandwidth. J of the Acoustical Society of America, pp. 548-560, Apr 1965.
21. S. Chapaneri. Spoken Digits Recognition using Weighted MFCC and Improved Dynamic Time Warping. Intl. J. Computer Applications 40:3, pp. 6-12, Feb 2012.

22. M. Hall. Correlation-based feature selection for discrete and numeric class machine learning. Proc. 17th Intl. Conf. on Machine Learning, pp. 359–366, Jul 2000.
23. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. The WEKA Data Mining Software: An Update. ACM SIGKDD Explorations 11:1, pp. 10-18, Jun 2009.
24. B. Ustun, W. Melssen, and L. Buydens. Facilitating the Application of Support Vector Regression by using a Universal Pearson VII Function based Kernel. Elsevier J Chemometrics and Intelligent Laboratory Systems, vol. 81, pp. 29-40, Mar 2006.
25. H. Homburg, I. Mierswa, B. Moller, K. Morik, and M. Wurst. A Benchmark Dataset for Audio Classification and Clustering. Intl. Society for Music Information Retrieval Conf. (ISMIR), pp. 528-531, Sep 2005.
26. M. Müller, D. Ellis, A. Klapuri, and G. Richard. Signal Processing for Music Analysis. IEEE J Selected Topics In Signal Processing 5:6, pp. 1088-1110, Oct 2011.
27. A. Meng, P. Ahrendt, J. Larsen, and L. Hansen. Temporal Feature Integration for Music Genre Classification. IEEE Tran. Audio, Speech, and Language Processing 15:5, pp. 1654-1664, Jul 2007.
28. K. Markov, and T. Matsui. Music Genre and Emotion Recognition using Gaussian Processes. IEEE Access, vol. 2, pp. 688-697, Jul 2014.
29. O. Lartillot, and P. Toiviainen. MIR in Matlab (II): A Toolbox for Musical Feature Extraction from Audio. Intl. Society for Music Information Retrieval Conf. (ISMIR), pp. 237–244, Sep 2007.
30. D. McEnnis, C. McKay, I. Fujinaga, and P. Depalle. jAudio: A Feature Extraction Library. Intl. Society for Music Information Retrieval Conf. (ISMIR), pp. 600-603, Sep 2005.
31. E. Pampalk. A MATLAB Toolbox to Compute Similarity from Audio. Intl. Society for Music Information Retrieval Conf. (ISMIR), pp. 254-257, Oct 2004.