International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 8
August 2017

A Comparative Approach for Analyzing Impact of Different Audio Features on Music Genre Classification

Ms. Sneha D Rodrigues
Thakur College of Engineering & Technology, Kandivali, Mumbai

Mr. Sanjeev Ghosh
Thakur College of Engineering & Technology, Kandivali, Mumbai

ABSTRACT– *With the advancement of technology in today's era, there is an utmost need for reliable music retrieval methods in order to organize and search through the large music archives that are available on the internet. Music genre classification is the most fundamental and essential component in music information retrieval (MIR) systems. An appropriate choice of music features and classifier is a crucial task for developing an accurate and efficient content-based classification system. In this work, a comparative analysis for four different set of features, viz. dynamic, timbre-texture, pitch and tonal features along with the statistical parameters is examined based on the performance of respective feature set. The performance evaluation is carried out on GTZAN musical database by using support vector machine (SVM) as a classifier. The experimental results show that out of all four set of features, better classification accuracy of* ***95.77%*** *is achieved for dynamic and timbre texture features.*

Keywords: Content-based classification, dynamic feature, feature extraction, music genre, pitch feature, statistical parameters, Support vector machines (SVM), timbre texture & tonal feature.

I. INTRODUCTION

Music is one of the most pleasing arts of human kind, which apart from mere entertainment has a vital role in our daily life. The evolution in digital music had a steady and placid progress, without much enthusiasm but made a huge impact on the precept of music consumption, sales and distribution overnight. It started with the standardization and ISO approval [ISO93] of the audio compression technique in 1993, i.e. MPEG-1 Audio Layer 3, which is generally referred as MP3 having an extension to music file as '.mp3'. The evolution of MP3 considerably reduced the storage size of audio files thereby requiring only a fragment of their original size on CD. Along with the rapid development of various affordable technologies, capable computers and Internet, easy content capturing/ storage and high speed transfer of music suddenly became possible without the need for any additional physical music medium. These recent advances in music field exploded the size of digital music collections from few privately owned CDs, to relatively huge amount of MP3 files stored on hard-disks and portable music devices.

With the possibility to access vast music archives almost anywhere and anytime it has become essential to offer end users, new ways to navigate and interact with these large music collections. This is where Music Information Retrieval (MIR) became an interesting and challenging topic in the field of research. Numerous methods for automatically measuring music similarities have already been evolved in the past years. However, existing techniques do not work effectively with the massive music collections accessible today.

The rest of the paper is organized as follows. Section 2 gives the idea about previously done work in the area of music genre classification followed by explanation of the general block diagram for the proposed method in section 3. Section 4 gives briefing of different features that can be extracted from the music signals. The overview of SVM classifier is explained in section 5. Section 6 deals with the experimental setup and database required for classification. The results and the corresponding analysis are covered in section 7 followed by section 8 that highlights the conclusion of the proposed work along with the scope for future work in area of MIR.

II. RELATED WORK

Music genre classification has gained importance in the field of MIR as genre being one of the most basic levels for differentiating various types of music. An appropriate choice of music features and classifier is a crucial task for developing an accurate and efficient content-based classification system.

In 2010, Ran Tao et al. proposed an approach for music genre classification [1] based on short-time timbre texture features that include Mel frequency cepstral coefficients (MFCC), zero crossing rates, spectral centroid, spectral roll-off, spectral flux, Root mean square energy (RMS) and 12 dimensional chroma vectors along with 7 different temporal descriptors. In this paper four different methods for representation of music clip were proposed and the efficiency of each method was analyzed by using support vector machine. The overall accuracy achieved was 78.6% on GTZAN music database.

Kirk Martinez et al. [2] extracts the timbre, rhythm and tempo features from audio file. An accuracy of 50% is achieved with the algorithm submitted to MIREX 2011 AMS. The accuracy was further enhanced to 84% by clustering the related genres.

Babu Kaji Baniya et al. [3] introduced a paper with additional harmonic feature along with timbral and rhythmic content features. The statistical parameters were also extracted for timbral texture features. A classification accuracy of 85.6% was obtained for GTZAN database by using Extreme learning machine (ELM) with bagging algorithm.

Further, a comparative approach for music genre classification was proposed by Babu Kaji Baniya et al. [4] by using Extreme learning machine (ELM) with bagging algorithm as a classifier. The timbral texture and rhythmic content features were extracted along with the statistical descriptors applied on timbre texture features. The experimental analysis was carried out on two different databases viz. GTZAN and ISMIR2004 with different combination of features and the classification accuracy of 85.15% and 86.46% was achieved respectively.

In 2014, Babu Kaji Baniya et al. [5] proposed an approach for music genre classification by using support vector machine (SVM) with 10 fold cross validation. In this paper four different groups of features were extracted, viz. dynamic, rhythm, spectral and harmony. From these features five different statistical parameters, viz. 4th order central moment of each feature along with the covariance component was extracted. The increased size of feature vector was reduced on elimination of insignificant features by using Minimum Redundancy Maximum Relevance (MRMR) and Principle Component Analysis (PCA) feature reduction techniques. The overall accuracy achieved was 87.9% and 78.2% on application of MRMR and PCA respectively.

III. PROPOSED METHOD

In this work we try integrate four different set of features, viz. dynamic, timbre-texture, pitch and tonal features that are extracted from the music clip. Various statistical parameters like temporal mean, temporal skewness, temporal kurtosis and covariance are applied on the extracted features. The entire sets of values that are extracted from a music clip are represented in a form of template that is known as feature vector. The final classification task is obtained by using the trained support vector machine (SVM) classifier and the comparative analysis is obtained from confusion matrix. The overall functional block diagram of the proposed method is shown in Fig. 1.

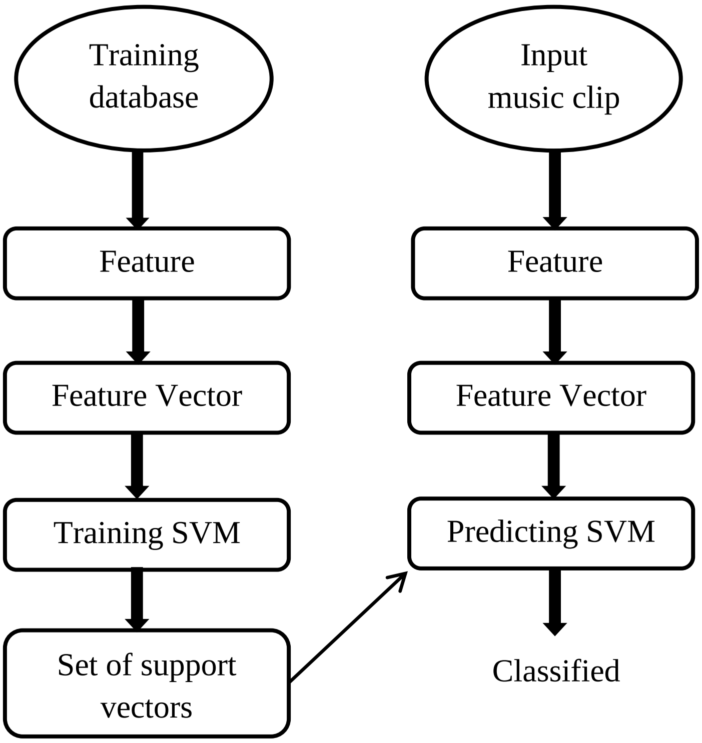


Fig. 1: Block diagram of the proposed method [1]

IV. FEATURE EXTRACTION

The objective of this step is to extract relevant and compact yet concise features vectors from the music clips that are capable of closely representing the music files in numerical format. The choice of features that need to be extracted for better efficiency is the challenging task in MIR systems.

The four types of feature sets are explained further;

2.1 DYNAMIC FEATURES

The dynamic features are one of the fundamental features related to music which contain the root mean square energy (RMS) and low energy [6]. Root mean square energy is the average energy that indicates the loudness of the music clip. It is computed as an intensity feature and is denoted as, $\mathrm{E_{rms}}$.

$$\mathrm{E_{rms}} = \sqrt{\frac{\sum_{n=0}^{N}[E(n)]^2}{N}} \qquad (1)$$

Where, E(n) is the energy at the sample value n and N is the total number of samples in the music frame [1].

2.2 PITCH FEATURE

A music composition made by various musical instruments or sung by a singer is a spectrum of signals with different frequencies. Pitch is the feature which gives the frequency information of such music signals. The pitch is estimated by applying short-time Fourier transform (STFT) and analyzing a series of spectrum. The harmonic product spectrum that is used to calculate pitch of the music signal is defined as [7].

$$P_n\left(e^{j\omega}\right) = \prod_{r=1}^{k} \left|X_n\left(e^{j\omega r}\right)\right|^2 \qquad (2)$$

Where, $X_n(e^{j\omega r})$ is the spectrum of a windowed frame. The log harmonic product spectrum is stored as a log frequency [7].

$$\widehat{P_n}\left(e^{j\omega}\right) = 2\sum_{r=1}^{k} \log\left|X_n\left(e^{j\omega r}\right)\right| \qquad (3)$$

2.3 TIMBRE TEXTURE FEATURES

Timbral features are a set of components that are used to differentiate mixture of sounds that have similar or rather same pitch, rhythm and loudness contents [8]. In order to extract timbral features, the music clips are first divided into a set of short-time frames that are statistically stationary by applying a windowing function to the original music clip at fixed intervals. Further the timbral texture features are computed on each frame [9]. Briefing of almost all timbral features that can be extracted from music clips is given below;

Mel-frequency Cepstral Coefficients (MFCC)is an efficient feature that represents the perfectly represents the human auditory system [10].The first 13 coefficients are selected after implementation of MFCC extraction process shown Fig. 2 for music classification task.

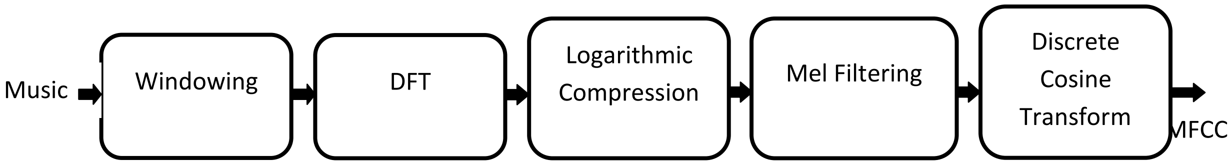


Fig. 2: MFCC extraction process [11]

Zero crossing rate (ZCR) is the number of time domain zero crossings within a frame. The ZCR provides the measure of noisiness of the music signal. Also the periodic music signals tend to have a small value of zero crossings than the noisy signals. The zero crossing rate for a particular frame x_r of length N, denoted as ZCR_r is computed by the formula [4];

Spectral centroid determines the point in the spectrum where most of the energy is concentrated and is correlated with the dominant frequency of the signal. Where, $M_t[n]$ is the magnitude of the STFT at frame t and frequency bin n, Spectral centroid is calculated as [4];

$$ZCR_r = \frac{1}{2}\sum_{n=1}^{N} \left|\mathrm{sgn}(x_r[n]) - \mathrm{sgn}(x_r[n-1])\right| \qquad (4)$$

$$C_t = \frac{\sum_{n=1}^{N} M_t[n] * n}{\sum_{n=1}^{N} M_t[n]} \qquad (5)$$

Spectral roll-off (R_t), is the frequency below which 85% of the signal's energy is accumulated and is given as [4];

$$\sum_{n=1}^{R_t} M_t[n] = 0.85 * \sum_{n1}^{N} M_t[n] \qquad (6)$$

Where, $\mathrm{M_t[n]}$ is the magnitude of the STFT at frame t and frequency bin n.

Spectral flux is the measure of change of spectral shape i.e. the measure of variation in the value of spectrum between the adjacent frames.

$$\mathrm{F_t} = \sum_{\mathrm{n1}}^{\mathrm{N}} (\mathrm{N_t[n]} - \mathrm{N_{t-1}[n]})^2 \qquad (7)$$

Where, $\mathrm{N_t[n]}$ and $\mathrm{N_{t-1}[n]}$ are the normalized magnitude of the STFT at the present frame t, and the previous frame $\mathrm{t-1}$ respectively[4].

Bandwidth is the magnitude-weighted average of the difference between the spectral components and the frequency centroid. It is a quantitative measure of the range of frequencies over which the power or density spectrum is concentrated [7].

2.4 TONAL FEATURES

Tonality refers to the quality of music signal that interprets the relationship between notes, chords and keys according to the pitch distribution.

The **12-dimensional Chromagram** is the measure of energy distribution along the pitches that is shown by harmonic pitch class profile (HPCP). The chroma information is obtained by decomposing the human auditory system perception of pitch into tone height (i.e. octave number) and chroma (i.e. pitch class) [1].

The **6-dimensional tonal centroid vector** refers to the projection of chords and it can be extracted from the 12-dimensional chromagram [6].

2.5 FEATURE DESCRIPTORS

The short-time low level features that are extracted within a frame, that have to be small enough according to the assumption that the signal for a short amount of time is stationary. However it is more relevant to govern the variation of features over a few separate frames. This can be accomplished by mapping the short frames into large texture windows. For the analysis of such variations between the consecutive frames, the following statistical parameters of music are taken into account;

The **arithmetic mean** is obtained by summing all the parameters in the feature vector and dividing that sum by the total number parameters in the vector [4].

Temporal skewness is the measure of the asymmetry of the distribution of feature values around its temporal mean value [4].

Temporal kurtosis is the measure of the flatness or peakedness of the distribution of feature values around its temporal mean value [4].

Covariance is measured between the two random variables or features to analysis the relationship between the random variables [4].

2.6 FEATURE VECTOR

To effectively classify music signals based on genre, the extracted features from the original music clip are represented in a form of a template known as feature vector. In this paper every music clip is represented by one feature vector that consists of 41 different features values.

V. SVM CLASSIFIER

The music classification based on genre level is implemented by using Support vector machine (SVM) classifier. It is a supervised binary classifier which performs classification by constructing an optimal hyper-plane that separates the training data points into two parts by, thereby predicting which part the test data falls into. However the use of this binary classifier can be extended for multi-class music genre classification by adopting any one of the two strategies, i.e. one-versus-one strategy and one-versus-all strategy [1].

In this work, one-versus-one strategy is applied by training SVM classifier for each pair of genre class, i.e. by constructing N*(N-1)/2 classifiers for N number of classes. The multi-class classification is achieved by evaluating the result from each possible pair of class and thereby assigning a new instance to the class that wins the largest number of votes.

VI. EXPERIMENTAL REQUIREMENTS

The prerequisite setup and parameters required for performing and evaluating the classification task are detailed below;

6.1 DATABASE

The performance evaluation of the proposed method is done on GTZAN database which was collected by G. Tzantakis [8]. It consists of 1000 songs divided into 10 music genres, viz. Blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae and rock. Every class of genre has 100 music clips of 30 seconds duration each. All music clips are in .au format (22050Hz, 16bits, mono). In our study, we have divided the music clips into short-time frames having 512 samples each.

6.2 EVALUATION PARAMETERS

A comparative analyzes is done for studying the impact of different set of features in classification task. The experiments are carried out by using MATLAB version 2014 for 3 different feature groups as tabulated below;

Table 1.Feature group

Sr. No.	Feature Group	Features included	No. of feature values/feature vector
1	FG1	Dynamic + Timbre Texture features	19
2	FG2	FG1 + Pitch + Tonal features	38
3	FG3	FG2 + Statistical descriptors	41

The classification task is performed by using SVM as a classifier and on the basis of K-fold cross validation (where k = 2, 5 and 10-fold), performance of each feature group is evaluated.

The parameter selected for evaluation is classification accuracy, where accuracy is defined as;

$$\frac{Number\ of\ correctly\ classified\ music\ clips * 100}{Total\ number\ of\ music\ clips\ in\ the\ database} \qquad (8)$$

VII. RESULTS AND ANALYSIS

In this work, we first analyzed the impact of K-fold cross validation (where k = 2, 5 and 10-fold) in music genre classification task by using SVM on FG1 given Table 1. The experiment results on music clips divided into 10 genres from GTZAN database are obtained by taking average of respective classification accuracy for a particular K-fold value. The details of the same are tabulated in Table 2. as below;

Table 2.Classification accuracy for different K-fold values on FG1

value of K-fold	Blues	classical	country	disco	Hip-hop	jazz	Metal	pop	reggae	rock	Avg %CA
2	78	95.5	79.5	70	77.5	76	87.5	86	72	63	78.5
5	93.8	97.6	91.4	90.4	89.6	92.4	94.4	94.4	89.4	87.2	92.06
10	97.2	98.8	95.8	94.7	95.2	95.7	94.9	97	94.6	93.8	**95.77**

From Table 2.it can be noted that classification accuracy (CA) for 10-fold cross validation applied on FG1 is greater than 2-fold and 5-fold cross validation.

On the similar lines, the experimental results for FG2 and FG3 from Table 1. are evaluated. The details are tabulated in Table 3. and Table 4.

Table 3.Classification accuracy for different K-fold values on FG2

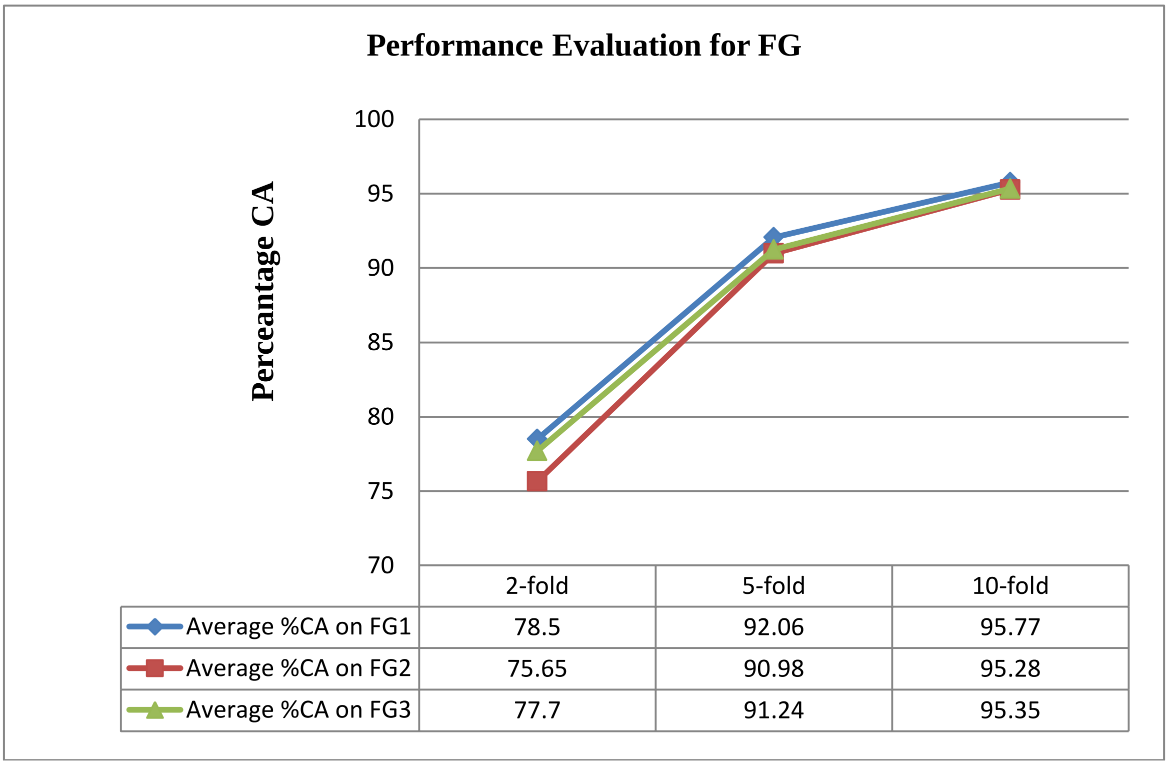
value of K-fold	blues	classical	country	disco	Hip-hop	jazz	Metal	pop	reggae	Rock	Avg %CA
2	72.5	86	77.5	71.5	75	73	83	83	69	66	75.65
5	91.6	95.2	91.2	90	87.8	90.6	93.2	94	89	87.2	90.98
10	96.4	97	95.5	95.1	92.2	95.6	96.4	97.3	94.5	92.8	**95.28**

Table 4.Classification accuracy for different K-fold values on FG3

value of K-fold	blues	classical	country	disco	Hip-hop	jazz	Metal	pop	reggae	Rock	Avg %CA
2	78	88	77.5	71.5	75	76	84.5	84.5	71	71	77.7
5	91.4	96	90.8	89	88.6	91.6	94	94	89	88	91.24
10	96.2	97.1	95.5	95.5	91.3	96	96.1	96.2	94.9	93.7	**95.35**

From the above tables, it can be stated that 10-fold cross validation in SVM results in overall improvement of classification accuracy.

Further Fig. 3. relates the impact of different feature groups on the classification task.



	2-fold	5-fold	10-fold
Average %CA on FG1	78.5	92.06	95.77
Average %CA on FG2	75.65	90.98	95.28
Average %CA on FG3	77.7	91.24	95.35

Fig. 3: Impact of individual feature group on classification accuracy.

The comparative analysis of different feature group shows that FG1, that includes dynamic and timbre texture features performs much better than the other two groups in all K-fold values of cross validation.

Table 5.comparesour proposed method with previous approaches in terms of average classification accuracy in the GTZAN database.

Table 5.Comparative analysis

Reference	Average classification accuracy (%)
Ran Tao [1]	78.60%
Kirk Martinez [2]	84.00%
Babu Kaji Baniya [3]	85.60%
Babu Kaji Baniya [4]	85.15%
Babu Kaji Baniya [5]	87.90%
Proposed method	**95.77%**

From the comparative analysis, it is clear that our proposed method gives better classification results than the other approaches.

VIII. CONCLUSION

In this work, we considered four different set of features, viz. dynamic, timbre texture, pitch and tonal features along with four different features descriptors, namely, mean, temporal skewness, temporal kurtosis and covariance. The comparative evaluation for performance of different feature groups as mentioned in Table 1 was carried out by using SVM along with K-fold cross validation technique. Experimental results show that 10-fold cross validation by SVM gives better results for every feature group. Further it can be concluded that FG1 that consist of dynamic and timbre texture features could achieve highest classification accuracy of 95.77% than other two feature groups. As, the proposed method concentrated only on GTZAN database with total 1000 music clips and 10 genres, this work can be further extended for various other databases having large numbers of songs. Also music genre classification based on mix-genre level can be implemented. The work can be further used to get better music retrieval results based on genre classification in the field of MIR.

REFERENCES

[1] Ran Tao, Zhen yang Li, Ye Ji, "Music genre classification using temporal information and support vector machine", Research gate, Jan 2010, in press.

[2] Kirk Martinez, Franz de Leon. A music genre classifier combining timre, rhythm and tempo models. In Proceedings of the TENCON 2012 IEEE region 10 Conference, 1324-9689.

[3] Babu Kaji Baniya, Deepak Ghimire, Joonwhoan Lee. Evaluation of different audio features for musical genre classification. In 2013 IEEE workshop on Signal processing systems.

[4] Babu Kaji Baniya, Deepak Ghimire, Joonwhoan Lee. Automatic music genre classification using timbral texture and rhythmic content features. In ICACT transactions on advanced communications technology (TACT), Vol. 3, Issue 3, May 2014.

[5] Babu Kaji Baniya, Joonwhoan Lee, Deepak Ghimire. Audio feature reduction and analysis for automatic music genre classification. In Proceedings of 2014 IEEE International conference on Systems, Man and Cybermetics.

[6] Zhe Wang, Jingbo Xia. The analysis and comparison of vital acoustic features in content-based classification of music genre. In 2013 International conference on Information Technology and Applications.

[7] Xi Shao, Changsheng Xu, Mohan S Kankanhalli. Applying neural network on the content-based audio classification. In ICICS-PCM 2003.

[8] Nicolas Scaringella, Giorgio Zioa, and Daniel Mlynek. Automatic genre classification of music content. In IEEE Signal processing magazine, March 2008.

[9] Tao Li, Mitsunori, Qi Li. A comparative study on content-based music genre classification. In SIGIR 2003.

[10] Jia-Ching Wang, Jhing-Fa Wang, Cai-Bei Lin, Kun-Ting Jian, Wai-He Kuok. Content-based audio classification using support vector machines and independent component analysis. In 18th

[11] Teng Zhang Ji Wu, Dingding Wang, Tao Li, "Audio retrieval based on perceptual similarity", 2014 International Conference on Collaborative Computing Networking, Applications and Work sharing, 22-25 Oct. 2014

[12] Deepa P. L., K. Suresh, "An optimized feature set for music genre classification based on support vector machine", IEEE, 2011

[13] G. Tzanetakis. GTZAN Genre Collection.