PAPER TEMPLATE FOR ISMIR 2014

First author
Affiliation1
author1@ismir.edu

Second author
Retain these fake authors in submission to preserve the formatting

Third author
Affiliation3
author3@ismir.edu

ABSTRACT

Pitch chroma are a popular feature for many MIR tasks. Using the GTZAN data set we investigate the distributions of pitch chroma for 9 different genres, the degree to which these genres can be identified using these distributions and different strategies for achieving key-independence; namely transposition of the chroma according to its maximum value and 12-point FFT. We find that combining pitch chroma with commonly used MFCCs can lead to small increase in classification accuracy using a Support Vector Machine. Furthermore these results show that the imposing key-independence has a surprisingly small affect on performance.

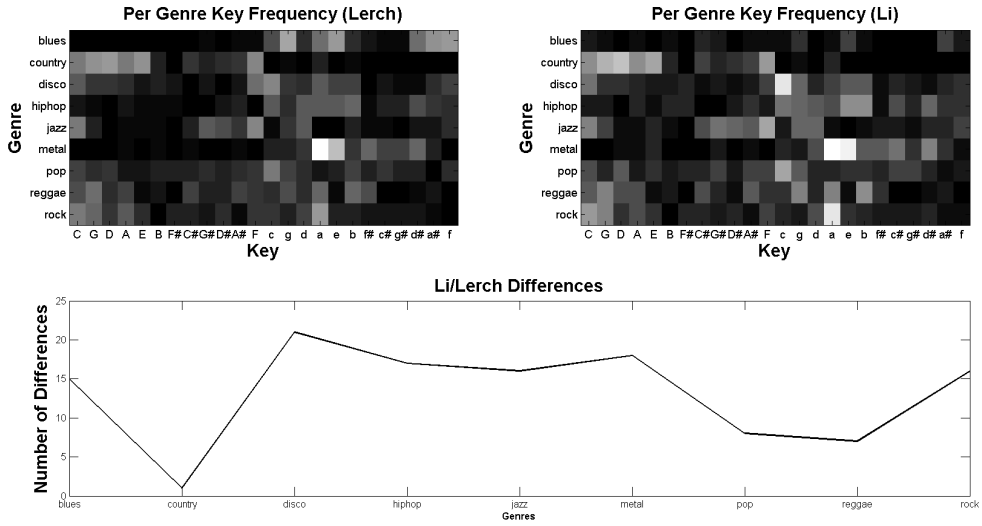
1. INTRODUCTION

Musical genre recognition is a well studied field in MIR [*ref]. As with any classification task, the feature used to summarize tracks is an extremely important concern. Previous work in this area has shown that timbral features, particularly Mel Frequency Spectrum Coefficients, are especially suited to the task of predicting genre. While MFCCs are suited to picking up on instrumental and timbral difference between genres, some work has shown that they are not totally independent of harmonic or tonal information [*ref li]. Here we investigate the extent to which tonal information can be used to discern genres y examining the distributions of pitch chroma within each of 9 different genres. First we examine the overall distribution of keys for each genre. Next different methods for transposing pitch chroma to a key-independent representation are introduced and we look at the chroma distributions using chroma box plots, inter-genre distance and mulch-dimensional scaling. In the final section we test the degree to which the chroma distribution separate genres by performing classification with an SVM using chroma and combined chroma/MFCC features.

2. DATA SET STATISTICS

The data set used was the GTZAN collection, which is popular for genre classification tasks. It consists of a collection of 1000 tracks divided into ten genres: Blues, Classical, Rock, Reggae, Pop, Metal, Rock, Jazz, Country and Hip Hop. First we examine the general layout of the data set, specifically with respect to the key distributions within genres. Key annotations for each track were produced manually and because of the difficulty of unambiguously estimating the key several were discarded, including basically all of the Classical genre.

A similar analysis was performed by Li [*ref], an examination of which reveals disagreements for several tracks – for example this analysis provided key labels for 98 tracks as opposed to 31 in the case of Li for the Blues genre. The most disagreements were seen in the Disco genre (21) and the fewest in Country (just 1 difference). The distribution of keys in each genre proved unsurprising: Jazz tracks tend to be clustered around flat keys which tend to be favored by trumpet and saxophone players. Blues are predominantly minor and mostly in the keys of G, E and B-flat. Country was almost exclusively in major keys which are convenient for the guitar; C-ma-



jor, G-major, D-major, A-major and E-major. The majority of Metal tracks are in either A-minor or E-minor, which are again well-suited to the electric guitar and bass and correspond to the two lowest open strings.

3. PITCH CHROMA AND KEY INDEPENDENCE

3.1 Pitch Chroma Extraction

Pitch chroma are a commonly used feature for Music Information Retrieval [*ref] which show the relative amount of intensity in each pitch class or note by grouping the magnitude of frequency bands of the Fourier Transform across multiple octaves. Each track in the labeled collection was down-sampled to 10kHz and pitch chroma vectors were extracted block-wise over a three octave range, starting from C = 130.8 Hz. Pitch chroma were then normalized using the 1-norm:

$$v_{norm} = \frac{v}{\|v\|_1} = \frac{v}{\sum_n v_n} \quad (1)$$

After extracting pitch chroma for each block we can also average them over whole tracks to obtain an "averaged pitch chroma" for each track, sometimes called a Pitch Histogram [**ref]. After averaging using the median for each chroma-bin each song is then represented by a single 12-dimensional chroma vector.

3.2 Key Independence

A problem with using a pitch chroma approach is that the overall tonality of each song may dominate any between-genre differences; because the distribution of major and minor keys are not the same for each genre, songs might be classified according to their key using this method. For example pitch chroma extracted from minor tracks would be classified as Metal, since the averaged pitch chroma for Metal would be obtained by averaging pitch chroma extracted from predominantly minor tracks. This is an inherent problem in using pitch chroma as opposed to more often used timbral features like MFCCs and in order to account for this we processed major and minor tracks separately. Another consideration revealed in the analysis of the data set is that the the key-labels are not uniformly distributed within each key. For example the majority of songs in the keys of A and E-minor are in the Rock and Metal genres. To account for this we investigated several techniques for obtaining a key-independent representation.

3.2.1 Transposition by Max

First the index of the maximum value of the chroma across all bins is found. Then, the chroma is transposed using a circular shift such that the index of the maximum is equal to one.

3.2.2 Fourier Transform

This method involves the use of the Fourier transform. Firstly, for a given pitch chroma vector the mean over all bins is calculated and is subtracted from each component. We then take the magnitude of each Fourier coefficient. Because transpositions in the chroma are encoded in the phase component, by throwing them out this quantity will be independent of transposition (and therefor of key).

3.2.3 Transposition by Ground Truth

The ground truth labels for each song is used to transpose each chroma using a circular shift so that the bin corresponding to the key label is in the first position (ie for a song in the key of A-major/-minor, the chroma is cyclically shifted so that the bin for the pitch class A is in the first index).

4. BOX PLOTS AND GENRE DISTANCES

The key-sorted songs in each genre were averaged to get a mean chroma for each track. For each genre we further take the median over all the songs to arrive at the final key-independent chroma for that genre. We also include the Major and Minor profiles found in [*ref] and [*ref] for comparison.

4.1 Box Plots

For each genre box plots were formed using the previously calculated median per chroma bin. For each plot the horizontal line represents the median for that bin and the boxes the region given by the second and third quartile ranges (outliers have been omitted for clarity). Here we have plotted the six most populated genres for both major and minor. Overall each genre roughly exhibits the expected major pattern with some exceptions, for example Jazz. The Jazz distribution is noticeably more flat than other genres – it doesn't exhibit the same peakiness at the root and fifth for example and the relative ratios of each interval are noticeably more uniform. This agrees with the classification results of Section 5, where Jazz was one of the least confused major genres.

For the minor examples, the distributions look much more uniform with relatively more mass in non-diatonic bins – for example the flat-second (C#) and-sharp fifth (G#). genres.

4.2 Genre Distances

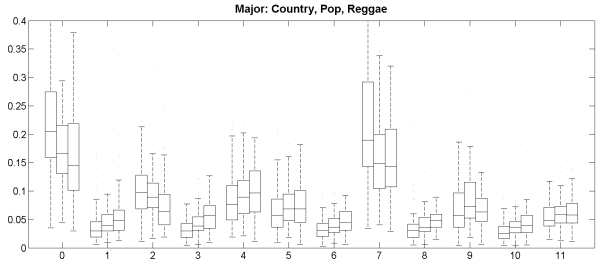
To see how distinct genres are with respect to their pitch chroma, distance between each genre chroma were calculated using Euclidean, Manhattan and KL-Divergence. The following figures show these between-genre distances using Manhattan distance. Genres for which the number of examples were less than 30 are grayed out. The labels are as follows: B is blues, C country, D disco, H hip-hop, J jazz, M metal, P pop, Rg Reggae, Rk is rock, K is the Krumhansl key profile and T the Temperley profile. Some facts revealed are that according to this metric, the most mutually distinct genres when considering Major tracks are Country and Jazz while the the most similar are Rock and Pop. For Minor tracks, Reggae and Pop are most similar while Reggae and Rock are the most distinct.

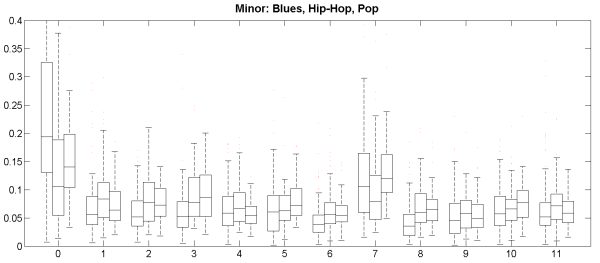
	B	C	D	H	J	M	P	Rg	Rk	K	T
B	0	0.29	0.31	0.47	0.33	0.23	0.28	0.32	0.23	0.33	0.37
C	0.29	0	0.17	0.46	0.28	0.28	0.13	0.22	0.12	0.27	0.33
D	0.31	0.17	0	0.29	0.13	0.24	0.14	0.11	0.13	0.13	0.21
H	0.47	0.46	0.29	0	0.18	0.3	0.35	0.27	0.38	0.2	0.25
J	0.33	0.28	0.13	0.18	0	0.24	0.17	0.15	0.21	0.12	0.18
M	0.23	0.28	0.24	0.3	0.24	0	0.25	0.23	0.2	0.19	0.33
P	0.28	0.13	0.14	0.35	0.17	0.25	0	0.14	0.1	0.18	0.23
Rg	0.32	0.22	0.11	0.27	0.15	0.23	0.14	0	0.15	0.1	0.15
Rk	0.23	0.12	0.13	0.38	0.21	0.2	0.1	0.15	0	0.18	0.26
K	0.33	0.27	0.13	0.2	0.12	0.19	0.18	0.1	0.18	0	0.15
T	0.37	0.33	0.21	0.25	0.18	0.33	0.23	0.15	0.26	0.15	0

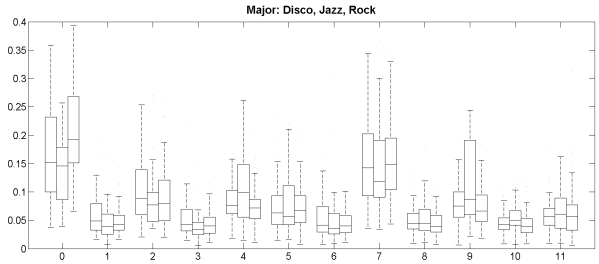
Table 1. Major genre distances, Manhattan distance.

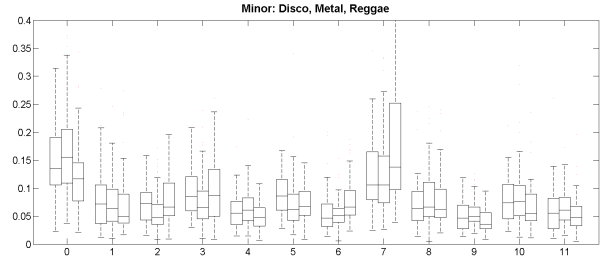
	B	C	D	H	J	M	P	Rg	Rk	K	T
B	0	0.22	0.21	0.27	0.16	0.17	0.17	0.25	0.24	0.25	0.36
C	0.22	0	0.3	0.45	0.25	0.3	0.29	0.26	0.24	0.39	0.42
D	0.21	0.3	0	0.22	0.19	0.12	0.07	0.12	0.13	0.14	0.19
H	0.27	0.45	0.22	0	0.27	0.22	0.21	0.26	0.33	0.18	0.27
J	0.16	0.25	0.19	0.27	0	0.21	0.14	0.2	0.22	0.24	0.28
M	0.17	0.3	0.12	0.22	0.21	0	0.11	0.19	0.16	0.17	0.29
P	0.17	0.29	0.07	0.21	0.14	0.11	0	0.1	0.12	0.13	0.21
Rg	0.25	0.26	0.12	0.26	0.2	0.19	0.1	0	0.12	0.18	0.22
Rk	0.24	0.24	0.13	0.33	0.22	0.16	0.12	0.12	0	0.21	0.26
K	0.25	0.39	0.14	0.18	0.24	0.17	0.13	0.18	0.21	0	0.16
T	0.36	0.42	0.19	0.27	0.28	0.29	0.21	0.22	0.26	0.16	0

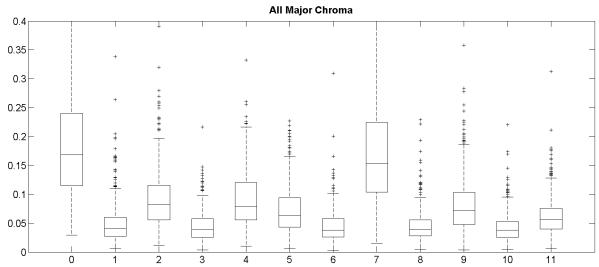
Table 2. Minor genre distances, Manhattan distance.

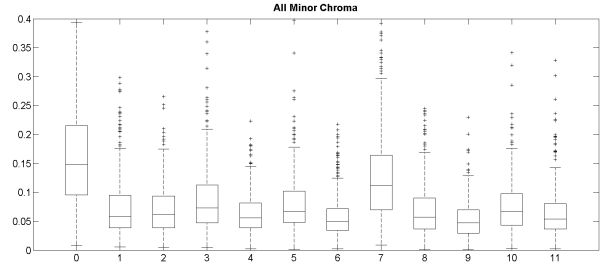


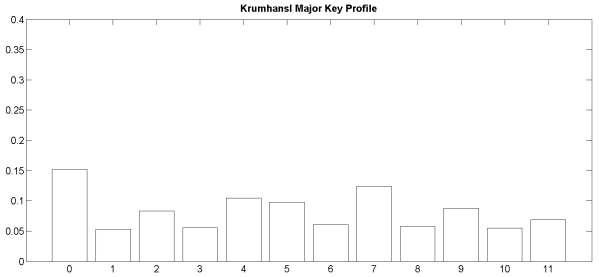


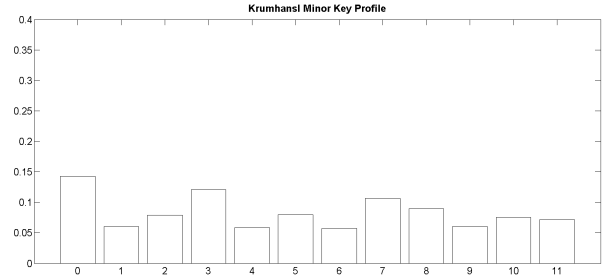












5. CLASSIFICATION

Here we investigate the use of pitch chroma for the purposes of genre classification. Multi-class classification was performed using a Support Vector Machine, specifically the LibSVM implementation in MATLAB[*ref]. The SVM hyper-parameters (C, γ) representing the margin and kernel widths were tuned using a grid-search and 5-fold cross validation on a separate stratified split of the data. Best performance was achieved using a non-linear SVM with Radial Basis Function kernel.

We tested using (*i*) the whole key-labeled data set and (*ii*) the major/minor subsets individually, in order to control for the differences in major/minor distribution between keys. 10-fold cross validation accuracy was calculated using raw pitch chroma, max-sorted pitch chroma, Fourier pitch chroma and ground truth sorted pitch chroma. We compare the results to commonly used MFCC features as well as a combination of MFCC and chroma features. MFCC features were calculated as follows: for each block we compute 12 MFCC coefficients, then take the mean and standard deviation over the whole song. These features are then concatenated to form a 24 dimensional feature vector.

5.1 Results and Discussion

The SVM performance is summarized in Tables * and *. One fact revealed in this analysis is the relatively small performance increase observed by sorting the chroma, even when using the ground truth directly. By sorting the chroma any information on the genre given by the key has been removed (for example as a prior estimate based on the key distributions in each genre) which would, it is hoped, reveal any structural differences. However whether or not this technique improves performance depends on the initial distribution of pitch chroma; suppose genre groups are perfectly separable given their key-independent chroma distributions and then for each song we transpose it into a new key with some related probability. Since cyclic-transpositions correspond to reflections this process is unlikely to make the data significantly more or less separable, which may explain the small increase in performance when using sorted chroma. Splitting the data into major and minor results in slightly improved accuracy, however in this case we have added information by using the ground truth and realistically harmony estimation should be done automatically which would be an additional source of error.

Unsurprisingly MFCCs vastly outperformed pitch chroma features alone. However, combining features results in a slight performance increase in overall accuracy of around 4-5% indicating that the pitch chroma distributions contain at least some genre-relevant information.

Feature	Major	Minor	All
Raw	35.35 ± 2.53	37.90 ± 1.39	35.04 ± 1.97
Max	37.24 ± 2.35	34.72 ± 2.21	**35.91 ± 1.65**
FFT	37.74 ± 2.29	36.36 ± 2.58	32.36 ± 2.08
GT	**40.33 ± 2.04**	**39.66 ± 3.33**	33.83 ± 0.92

Table 3. 10-fold CV accuracies using pitch chroma.

Feature	All
MFCC	55.01 ± 2.96
MFCC+Raw	59.31 ± 2.08
MFCC+Max	59.14 ± 1.7
MFCC+FFT	57.93 ± 1.66
MFCC+GT	**59.94 ± 2.02**

Table 4. 10-fold CV accuracies using MFCC and pitch chroma.

	B	C	D	H	J	M	P	Rg	Rk	Total
B	0	0	0	0	0	0	1	1	0	2
C	0	51	1	0	5	0	3	6	9	75
D	0	12	7	0	0	0	2	7	6	34
H	0	2	2	1	1	0	2	1	1	10
J	0	8	1	0	26	0	0	1	5	41
M	0	0	0	1	0	0	0	2	0	3
P	0	17	2	0	0	0	4	7	5	35
Rg	0	11	4	0	1	0	2	7	18	43
Rk	0	20	4	0	2	0	2	4	13	45

Table 5. Confusion Matrix: Major ground-truth sorted chroma.

	B	C	D	H	J	M	P	Rg	Rk	Total
B	51	0	1	3	4	14	2	0	1	76
C	2	0	0	0	0	1	0	1	0	4
D	1	0	14	8	0	12	4	3	2	44
H	3	0	2	32	0	13	1	4	0	55
J	12	0	1	0	8	0	0	0	0	21
M	14	0	4	14	0	34	3	1	1	71
P	7	0	8	7	0	15	1	1	1	40
Rg	2	0	2	8	0	9	3	10	1	35
Rk	13	0	4	1	0	7	1	5	3	34

Table 6. Confusion Matrix: Minor ground-truth sorted chroma.

	B	C	D	H	J	M	P	Rg	Rk	Total
B	54	3	3	2	1	3	5	4	5	80
C	5	55	6	1	0	1	3	3	6	80
D	4	5	40	8	2	2	9	3	7	80
H	0	0	9	44	13	3	2	9	0	80
J	1	1	1	12	58	1	3	0	3	80
M	3	0	4	3	2	63	0	1	4	80
P	2	9	10	7	6	0	41	5	0	80
Rg	4	4	3	12	1	0	3 4	9	4	80
Rk	4	11	18	0	3	7	4	4	29	80

Table 7. Confusion Matrix: MFCC + ground truth chroma features.

6. CONCLUSION

7. REFERENCES

[1] E. Author: “The Title of the Conference Paper,” *Proceedings of the International Symposium on Music Information Retrieval*, pp. 000–111, 2000.

[2] A. Someone, B. Someone, and C. Someone: “The Title of the Journal Paper,” *Journal of New Music Research*, Vol. A, No. B, pp. 111–222, 2010.

[3] X. Someone and Y. Someone: *Title of the book*, Editorial Acme, Porto, 2012.