

Music Genre Classification Revisited: An In-Depth Examination Guided by Music Experts

Haukur Pálmason¹, Björn Þór Jónsson^{1,2}, Markus Schedl³, and Peter Knees⁴

¹ Reykjavík University, Iceland

² IT University of Copenhagen, Denmark

³ Johannes Kepler University Linz, Austria

⁴ TU Wien, Austria

bjorn@ru.is

Abstract. Despite their many identified shortcomings, music genres are still often used as ground truth and as a proxy for music similarity. In this work we therefore take another in-depth look at genre classification, this time with the help of music experts. In comparison to existing work, we aim at including the viewpoint of different stakeholders to investigate *whether musicians and end-user music taxonomies agree on genre ground truth*, through a user study among 20 professional and semi-professional music protagonists. We then compare the results of their genre judgments with different commercial taxonomies and with that of computational genre classification experiments, and discuss individual cases in detail. Our findings coincide with existing work and provide further evidence that a simple classification taxonomy is insufficient.

Keywords: music genre classification, expert study, ground truth

1 Introduction

In the last 20 years, almost 500 publications have dealt with the automatic recognition of musical genre [15]. However, genre is a multifaceted concept, which has caused much disagreement among musicologists, music distributors, and, not least, music information retrieval (MIR) researchers [12]. Hence, MIR research has often tried to overcome the “ill-defined” concept of genre [10, 1]. Despite all the disagreement, genres are still often used as ground truth and as a proxy for music similarity and have remained important concepts in production, circulation, and reception of music in the last decades [3]. Their relevance for music perception is evidenced by studies that show the existence of common ground between individuals, e.g., [7, 13], their importance in users’ music similarity assessment [9], and their recognizability within fractions of seconds [4, 6]. As a result, genre classification remains a relevant task in MIR research [8, 11].

In comparison to work on optimising genre classification, work discussing ground truth for MIR, and in particular work discussing the viewpoint of different stakeholders, is scarce. For this reason, in this work, we investigate whether musicians and end-user music taxonomies agree on genre ground truth by comparing different commercial taxonomies and discussing individual cases in detail.

The remainder of this paper is organised as follows. We first discuss related work on defining and investigating genre ground truths (Section 2), then present

our study involving music experts (Section 3). Subsequently, the results of our study are discussed in detail, including a thorough review of selected artists and songs (Section 4). The paper is rounded off by concluding remarks (Section 5).

2 Related Observations on Genre Ground Truth

The original version of [4] from 1999 is much cited although it has been unavailable in print until the re-release in 2008. The authors chose the following 10 genres: blues, classical, country, dance, jazz, latin, pop, R&B, rap and rock. 52 university students representing “ordinary undergraduate fans of music” listened to excerpts from eight songs of each genre. The excerpts varied in length from 250 ms to 3,000 ms. Genre classification was taken from CDnow, BMG and Tower Records, the leading web based music vendors of the nineties. When listening to the 3,000 ms excerpts participants agreed with the ground truth about 70% of the time. When participants were only allowed to listen to 250 ms excerpts the accuracy varied greatly with genres, with less than 20% accuracy of blues songs, but over 70% accuracy of classical songs, with the average across all genres being 44%. A study with a small group of music theory majors revealed essentially the same results as with the non-musicians in the main study.

Lippens et al. [7] compared the results of automatic genre classification and human genre classification on the MAMI dataset. The MAMI dataset consists of 160 full length songs, originally classified into 11 genres. They concluded that due to various reasons this classification was not fit for automatic genre classification and therefore conducted a user study with 27 human listeners. Each participant listened to a 30-second-excerpt from all the songs and classified each song into one of six genres. The outcome from that study was as follows: 69 pop, 25 rock, 24 classical, 18 dance, 8 rap, and 16 other, with the genre “other” being used for songs that did not fit into any of the first five genres. The next step was to compare the selected genre of each participant with this new ground truth. The accuracy of the 27 participants ranged from 57% to 86% averaging at 76%. A subset of the MAMI dataset, called MAMI2, was then created. It included songs from the first five genres mentioned above, and only songs that had received 18 or more votes for their particular genre. This resulted in 98 tracks. The average classification accuracy of the participants for this dataset was 90%.

Craft et al. [2] criticized how the MAMI2 dataset was created, and claimed that it was “not statistically well-founded”. Their argument was that the meaning of the genre “other” was undefined to the participants, resulting in different ways of using that genre: should participants only use it for songs that did not find a home in any of the other genres or should they also use it if a song features multiple genres? They examined the songs that did not make it into the MAMI2 dataset and found out that only one of these songs received 10 votes for “other”, one song received seven votes, but the remaining songs received five or fewer votes for the “other” genre. The authors then constructed a similarity graph of all songs in the MAMI dataset, where songs with similar *distribution of genre votes* were grouped together. It turned out that there were groups of tracks that spanned multiple genres, and there were genres that spanned multiple groups of similar tracks. The main conclusion of the paper was that it is unrealistic to try to create a genre classification dataset that is entirely unambiguous, since

real life datasets do not only contain unambiguous data. They proposed that all results from automatic genre classification systems should be weighted to reflect the amount of ambiguity of human classification of that same dataset.

The most commonly used dataset, GTZAN, introduced in the archetypal work in the field of genre recognition by Tzanetakis and Cook [16], contains 10 musical genres, namely: classical, country, disco, hiphop, jazz, rock, blues, reggae, pop, and metal. Despite its wide acceptance and reuse in subsequent studies, it exhibits inconsistencies, repetitions, and mislabeling as investigated in detail by Sturm [14].¹ Apart from challenging the notion of categorizing music pieces into exclusive genres, Sturm argues convincingly that the errors in the genre assignments make results stemming from different approaches non-interpretable and incomparable, as different machine learning algorithms are affected in different ways. Despite the already identified shortcomings of GTZAN, we investigated parts of this dataset with the help of musical experts to gain further insights.

3 User Study with Music Experts

For our user study, we asked music experts to classify selected tracks of the GTZAN dataset. To keep the workload low, we chose examples that were misclassified by a k -NN classifier (see below), since these seem to be difficult, mislabeled, or exhibit other particularities that justify a deeper investigation.

Using a new, very efficient k -NN classifier [5] with $k = 3$ on features consisting of MFCCs and spectral flatness measure (SFM) extracted through MARSYAS (<http://marsyas.info>) we reach a genre classification accuracy of 80.8% in a 10-fold cross validation setting, which closely matches the best results in the literature obtained using these particular features. That leaves 192 tracks misclassified, however, which are distributed over genres as illustrated in Figure 1.

To analyse these tracks in more detail we set up an experiment where 20 participants listened to the 192 wrongly classified songs. The participants are all active in the Icelandic music industry, either as musicians, producers, sound engineers, or DJs, and include both semi-professionals and professionals. More precisely, among the semi-professionals we included a singer/songwriter who has released two albums, but never received the recognition necessary to completely quit his day job, a DJ at a local club in Akureyri who also works at a computer store, a guitarist and singer in a wedding/club band who has a day job as a painter, and a music blogger who works in a factory during the day. The professionals includes a radio DJ at one of Iceland’s biggest radio stations, a guitarist and guitar teacher at the Akureyri School of Music, a drummer and drum teacher at the Akureyri School of Music, and a music producer and recording engineer.

Each participant received a list of the 10 genres of the GTZAN dataset² and then listened to the 30-second-clips for all 192 misclassified tracks, marking each track with the genre label that they felt best described that song. The listening environment was a quiet room with a high fidelity stereo system.

¹ George Tzanetakis, the author of the dataset, has repeatedly confirmed being aware of these issues, but has chosen not to correct them since the dataset has been used so many times and changing the files would render comparisons of results infeasible.

² The genre “classical” was included even though no track was classified as such.

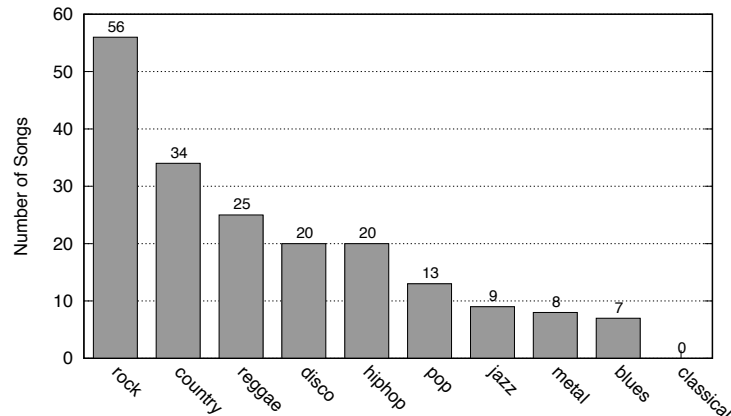


Fig. 1. Genre distribution of the 192 misclassified songs used in our user study.

While the dataset contains no information about artists or song names, listening to the songs reveals several interesting facts, including two mistakes in its creation. First, a live version of the song “Tie your mother down” by Queen is included twice, once labelled as rock and once labelled as metal. During k -NN search both versions get wrongly classified since the version labelled rock gets all its votes from the version labelled metal and vice versa. Second, one reggae sound clip is faulty, with only 6 seconds of music and 24 seconds of loud noise. It is interesting to note that the k -NN classifier labels this noise as pop.

The set often includes several songs by the same artists. Out of these 192 songs 7 songs are by Sting, 6 by Jethro Tull and 4 by Rolling Stones. Other artists that have multiple songs include Black Sabbath, Led Zeppelin, Beastie Boys, Bob Marley, Willie Nelson, Alanis Morissette, Vince Gill and Guns’n’Roses. All Sting songs are from his first two albums “Dream of the Blue Turtles” and “Bring on the Night” where Sting uses famous jazz musicians including Branford Marsalis on saxophones and Kenny Kirkland on pianos. All these Sting songs are classified as rock by Tzanetakis, whereas participants were divided between pop and jazz. The Jethro Tull songs included such diverse songs as “Happy and I’m smiling”, “Bungle in the Jungle” and “Life is a love song”. Again were all songs considered rock by Tzanetakis. Most were also classified as rock by participants, while some were classified as pop. Many commented that “Life is a love song” is really folk or acoustic, but no such genre is included in the set.

4 Results

4.1 Comparison with GTZAN Ground Truth and k -NN Classifier

Table 1 compares the manual classification of the music experts to the ground truth of the GTZAN dataset. The table shows that the agreement between individual participants and the ground truth ranges from 52.6% to 69.8%, and is

Participant Agreement	GTZAN ground truth		k -NN classifier	
	Songs	Percentage	Songs	Percentage
Lowest	101	52.6%	15	7.8%
Highest	134	69.8%	34	17.7%
Median	112	58.3%	25	13.0%
Average	113.4	59.1%	24.9	13.0%
Majority vote	122	63.5%	24	12.5%

Table 1. Participants’ agreement with the ground truth and the k -NN classifier. Agreement is the share of songs participants described with the same genre as the ground truth given by the GTZAN dataset (3rd column) and as the prediction of the classifier (rightmost column).

on average 59.1%. Agreement is defined as the share of songs participants annotated with the same genre as the ground truth given by the GTZAN dataset (3rd column) and as the prediction of the classifier (5th column). The table reports results at different levels of agreement/types of users, e.g., the user with the lowest agreement in row "Lowest"; analogously for the other rows. For instance, the participant (out of the 20) who agreed the least with the ground truth agreed with the classification of 101 songs (out of the 192), or 52.6%, and he agreed with the classification of k -NN on only 15 songs (out of the 192).

These numbers should not be compared with the results from [4] and [7] since the dataset used for this experiment was specifically chosen because the automatic genre classifier was not able to classify the songs correctly. It is, however, interesting to note the low agreement rate on these songs. Table 1 also shows the number of tracks where the participants agreed with the results of the k -NN classifier; remember that these tracks were all wrongly classified by the automatic classification. It is interesting that for nearly 27% of the tracks, participants agreed neither with the GTZAN ground truth nor the k -NN classifier.

We then created a new ground truth using the majority vote of the 20 participants. As the last line of Table 1 shows, this new ground truth agrees with the GTZAN ground truth for 63.5% of the 192 tracks, and with the k -NN classifier for 12.5% of the 192 tracks; for 24% of the tracks the new ground truth agrees with neither. To examine closer how much the participants agreed, we used this new majority vote as ground truth. Table 2 confirms that there is

Participant Agreement	Songs	Percentage
Lowest	121	63.0%
Highest	166	86.5%
Median	153	79.7%
Average	150.3	78.3%

Table 2. Participants’ agreement with “majority vote” ground truth. Agreement is defined as in Table 1.

Genre	GTZAN ground truth		k -NN classifier	
	Songs	Percentage	Songs	Percentage
Blues	6	85.7%	1	14.3%
Country	29	85.3%	1	2.9%
Disco	10	50.0%	6	30.0%
Hiphop	18	90.0%	1	5.0%
Jazz	7	77.8%	2	22.2%
Metal	1	12.5%	4	50.0%
Pop	12	92.3%	1	7.7%
Reggae	16	64.0%	4	16.0%
Rock	23	41.1%	4	7.1%
Total	122	63.5%	24	12.5%

Table 3. Majority vote agreement, by genre, with GTZAN and k -NN classifier. Agreement is defined as in Table 1.

considerable variation in the way our participants classified the songs, with the highest agreement with the new majority vote ground truth being 166 songs, or 86.46%. However, we also see from this table that overall there is more individual agreement with this new majority vote ground truth than the original GTZAN ground truth, so there seems to be a number of songs that everyone believes are wrongly classified in the original ground truth.

As a more detailed analysis, Table 3 shows a comparison of the majority vote for each genre to both the original ground truth and the results from our k -NN classification. As the table shows, participants agree strongly with ground truth for pop, hiphop, blues, country and jazz. Reggae and disco have moderate agreement, while rock and specially metal have very low agreement.

4.2 Comparison with the “World Out There”

In order to compare the classification of ground truth, k -NN and our participants to that of the world out there, we selected 15 songs randomly from the songs in the dataset that we recognized. We then looked at how these songs are classified on iTunes, allmusic.com and last.fm.

Apple’s on-line media store, iTunes, only classifies albums so songs actually can have multiple genre classifications if they are featured on more than one album. Two songs in our set, David Bowie’s “Space Oddity”, and Jethro Tull’s “Life is a love song” fall into this category, where both are classified as pop in one place, and rock in another place.

Allmusic.com is a music reference web page with album and artist critique. Allmusic.com classifies artists into genres and styles, where genres are usually very broad, such as “Pop-Rock” but styles are narrower. We report the genre and the two top styles of each artist.

Last.fm is an Internet radio station that allows users to tag songs. Tags can be any text that listeners use to describe songs. Most popular tags are displayed

Artist Song	GTZAN	iTunes	allmusic.com	last.fm	k-NN genre / %	participants genre / %	
Ani DiFranco Cradle and all	rock	alternative	pop/rock folk urban folk	folk female.voc indie	country jazz blues	24 pop 22 jazz 18 blues	85 15
Billy Joel Movin' out	rock	rock	pop/rock singer/songwr. soft rock	classic rock pop soft rock	disco rock country	35 pop 21 rock 20 country	75 25
Bob Seger Against the wind	rock	rock	pop/rock rock'n'roll hard rock	classic rock rock soft rock	disco country reggae	24 pop 23 country 16 rock	65 25 10
David Bowie Space Oddity	rock	pop/rock	pop/rock hard rock glam rock	classic rock glam rock british rock	country disco rock	37 pop 23 rock 19 reggae	75 20 5
Jethro Tull Life is a love song	rock	pop/rock	progressive blues-rock hard rock	country classic rock 70s	disco rock country	43 pop 16 rock 15 blues	85 10 5
Led Zeppelin D'yer Mak'er	rock	rock	pop/rock blues blues-rock	classic rock rock reggae	pop disco rock	26 reggae 23 rock 23 pop	50 35 15
Simply Red Freedom	rock	pop	pop/rock soul adult.cont.	pop rock easy	disco rock blues	28 pop 25 jazz 21 disco	55 25 15
Sting Consider me gone	rock	rock	pop/rock adult cont. cont. pop/rock	rock jazz pop	pop hiphop reggae	31 jazz 15 pop 14 blues	50 45 5
Jimmy Cliff Many rivers to cross	reggae	reggae	reggae reggae-pop roots reggae	reggae soul jamaica	classical jazz country	26 pop 23 classical 22 reggae	70 25 5
Marcia Griffiths It's Electric	reggae	reggae	reggae dancehall roots reggae	funk dance party	pop disco hiphop	60 pop 23 disco 5 hiphop	65 30 5
Cher Believe	pop	pop	pop/rock dance-pop adult. cont.	pop dance 90s	disco pop reggae	26 disco 24 pop 23 hiphop	60 35 5
Madonna Music	pop	pop	pop/rock dance-pop adult.cont.	pop dance electronic	hiphop pop jazz	32 pop 22 hiphop 18 disco	65 25 15
Guns'n'Roses Live and let die	metal	rock	pop/rock hard rock heavy metal	rock hard rock cover	rock metal disco	54 rock 38 metal 4 blues	75 20 5
Living Colour Glamour Boys	metal	rock	pop/rock alt. metal alt. pop/rock	rock funk rock 80s	hiphop metal disco	34 rock 28 pop 22	60 40
Willie Nelson Georgia on my mind	country	country	country trad.country progr. country	country classic country folk	blues country classical	38 country 29 blues 16 jazz	40 30 20
Beastie Boys Fight for your right	hiphop	hiphop/rap	rap pop/rock alt. pop/rock	hip-hop 80s rock	metal hiphop rock	59 rock 18 metal 17 hiphop	70 25 5

Table 4. Comparison of ground truth, iTunes, allmusic.com and last.fm, k -NN classification and participants' voting.

on the website. We report the three most popular tags, omitting all tags that include artists or song names.

Table 4 shows the comparison of the ground truth, k -NN classification, our participants' voting, iTunes, allmusic.com and last.fm. The table shows that iTunes agrees with the ground truth of the dataset in most cases, or 12 for out of the 15 songs, if we count the two songs that have both pop and rock classification in iTunes. The allmusic.com genre label is very broad, and in 12 out of the 15 songs this genre label is pop/rock. This goes for songs classified as pop, rock or metal by the ground truth. The table also shows that in 8 songs the k -NN classification has the correct genre in 2nd place, and in 2 songs the correct genre comes in 3rd place. From this small sample our participants only agree with the ground truth for 2 songs which is quite far from their agreement for the whole 192 songs. The participants have the correct genre in 2nd place in 6 songs, and in 3rd place in 3 songs.

4.3 Discussion of Particular Songs

We now discuss in order each of the 15 tracks from Table 4 in more detail, both the song itself as well as the various classifications.

Ani DiFranco's "Cradle and all" has a very strong acoustic guitar presence and this is without a doubt the reason why our k -NN program classifies the song as country. Many country songs have this same sound character. We see folk mentioned both at allmusic.com and last.fm, which also is a genre characterized by the acoustic guitar, but our ground truth does not include this genre. iTunes uses the alternative genre for this song, but this genre is very ill-defined. Our participants classify the song as a pop song, with several of them commenting that they would use folk, or acoustic pop, if either was available.

Billy Joel's "Movin' out" can hardly be classified as a disco song, although it has the dry 70's drum sound. Yet our k -NN classifier classifies it as disco, as too many other rock songs, with rock coming in second. 75% of our participants classify it as a pop song with the remaining votes going to rock. Both allmusic.com and last.fm use terms such as soft rock, which we believe is a synonym for pop in many people's mind.

Bob Seger's "Against the wind" is on all three websites considered a rock song. However, our solution does not have rock in the top three places, whereas 2 of our 20 participants classified the song as a rock song. The song has several elements of a classic country song including the acoustic guitar, the piano playing, and the vocal harmonies. This is one of the rock songs which our k -NN classifier classifies as a disco song, which is plainly wrong. We believe that if multiple genres were to be used, then pop, rock and country should all be used.

David Bowie's "Space Oddity" features the acoustic guitar very much, and this is without a doubt the reason it gets classified as a country song by k -NN. All websites use the rock genre, sometimes with specific sub-genres of rock for this song, although iTunes classifies it as pop when it is a part of Bowie's "Singles collection" album. Most participants in our study classified the song as a pop song, with rock coming in second. The reggae classification of one participant must be a mistake, since there is not a single reggae element in the song. Just as it is difficult to pinpoint the boundaries between rock and metal, it is also very difficult to pinpoint exactly the difference between pop and rock.

Jethro Tull's catalog of songs is extremely diverse, so classifications on artists level are not going to be very accurate. Allmusic.com classifications of blues rock or hard rock hardly describe "Life is a love song" well. Last.fm tags of progressive, classic rock and 70's are more accurate, although less popular tags, such as folk rock describe the song better, in our opinion. Our participants classified it as pop, with rock coming in second, and one participant using the blues genre. The song is very acoustic, with acoustic guitars, mandolins and a flute. As with some other acoustic songs it gets a considerable number of votes from country songs in k -NN classification.

Led Zeppelin is of course one of the greatest rock bands in history, so it does not come as a surprise that "D'yer Mak'er" is classified as a rock song by ground truth, iTunes, and two most popular last.fm tags, with allmusic.com using blues and blues rock. Blues is indeed where the roots of Led Zeppelin lie. 50% of our participants and a considerable number of last.fm users want to classify this song as a reggae song, and it cannot be denied that indeed it has much more reggae feel than "Many rivers to cross". At the same time it features some pop elements, reflected for instance in its instrumentation.

Simply Red's "Freedom" is classified as rock by ground truth. This time we are not surprised with the disco classification of k -NN since the song has in our opinion more disco elements than rock elements, including the guitar sound and the prominent strings. The rhythm, although not the standard disco beat, also resembles disco, with very prominent bongo drums and tambourines. A vast majority of participants classify the song as a pop song, thereby agreeing with iTunes and the most popular last.fm tag (where rock comes in second).

Sting's "Consider me gone" is one of the songs he recorded with several famous jazz musicians. Our participants have almost the same number of votes for jazz and pop for this song, with one person considering it a blues song. None mentioned the rock genre used by the ground truth, iTunes, and last.fm. We notice, however, that last.fm also has both pop and jazz tags, while allmusic.com concentrates on the adult-contemporary label. This is one of these songs where it is very difficult to say that one particular genre is correct.

Jimmy Cliff's "Many rivers to cross" is yet another one of those difficult songs. Websites and ground truth agree on defining the song as a reggae song, but the song does not include any trademark reggae features, such as the off-beat rhythm. Instead it has some classical characteristics, such as the prominent church organ sound. Jimmy Cliff is one of those artists that has merged reggae and pop music successfully, and as with Marcia Griffiths this song is perhaps not very representative for him. Most of our participants classify this as a pop song, with classical coming in second.

Marcia Griffiths' "It's electric" is an example of a song that perhaps does not represent the artist very well, and therefore there is inconsistency between genres that are created by artist or album classifications and genres that are created by song classification. Both k -NN and our experiment participants classify this as a pop song, with disco and hip-hop coming in 2nd and 3rd, respectively. Last.fm tags include funk, dance and party which can be said to be closer to the pop, disco, hip-hop, categories than reggae assigned by both iTunes and allmusic.com. However, some of our participants commented that Marcia Griffiths is known as

a reggae artist, but they still could not classify this particular song as a reggae song.

Cher's "Believe" features the infamous disco drum beat where the high-hat opens on every offbeat. Most of the instruments are obviously programmed, which makes the sound different from the classic 70's disco songs. Participants agree with k -NN in classifying this as a disco song, but both put pop in second place with the difference in votes in the k -NN classification being very low. Perhaps the style dance-pop used by allmusic.com describes it best, but what is dance-pop other than a combination of disco and pop?

Madonna's "Music" is a very electronic song. Most, if not all instruments are electronic in nature and programmed instead of being "hand-played". It has this in common with most hiphop songs, in addition to some strange vocal effects. However, in our opinion it lacks the hiphop beat to be classified as a hiphop song. We see that our participants agree with ground truth, iTunes and last.fm most popular tag, in classifying it as a pop song, and indeed pop is the genre with the second most votes in k -NN. Allmusic.com uses dance-pop which also describes the song very well.

Guns'n'Roses version of the Wings hit "Live and let die" is considered a metal song by ground truth. iTunes, k -NN, participants and last.fm all agree on rock, while the first style at allmusic.com is hard rock, with metal coming in second for both k -NN and our participants. It is difficult to say where the boundaries lie between rock and metal. This song does include a large dose of overdriven guitars, which does characterize metal, but in our opinion the overall sound and feel is much more rock.

Living Colour's "Glamour Boys" is classified as hiphop by k -NN with metal and disco in 2nd and 3rd place, respectively. Ground truth considers this a metal song, while participants, iTunes and the most popular last.fm tag agree on rock. Some participants commented that indeed the verse with its clean guitar sound of the song is a pop verse, while the chorus with its overdriven guitar and more aggressive voice is more rock oriented. This caused some of them to have problems deciding which genre to use. In the end it was 60/40 for rock against pop.

"Georgia on my mind" has been recorded by many artists. With Willie Nelson being a country icon, iTunes, which classifies albums, and allmusic.com, which classifies artists, use the country genre for his version of this song. The three most popular tags at last.fm are country, traditional country and folk. The fourth most popular tag (not counting the tag Willie Nelson) is blues. k -NN classifies the song as blues with country coming in second place, while this is reversed for our participants. The song, in our opinion, is more of a blues song than a country song, but Willie Nelson does of course bring some country flavor to it.

Beastie Boys' "Fight for your right" would probably never be classified as a hiphop song by people that heard it the first time and did not know that Beastie Boys are a hiphop/rap band. The instrumentation and rhythm are those of a typical rock/metal song, with loud overdriven guitars, and simple bass and drum beats. The vocals are the only thing that resemble rap music. k -NN strongly classifies this as metal with hiphop and rock coming in 2nd and 3rd, while 70% of our participants classify it as rock, and 25% as metal. One participant classified it as hiphop.

4.4 Impact of Ground Truth Definition on Classification Accuracy

Having seen that the participants in our ground truth experiment had in many ways different opinions on which genre songs in the GTZAN dataset should belong to, we decided to change the ground truth of the songs where the majority vote of participants differs from the ground truth. Recall from Table 1 that the majority vote results from the experiment agrees with the ground truth for 122 songs of the 192 that were incorrectly classified by the k -NN classifier, meaning that we changed the ground truth of 70 songs. Table 1 shows us that out of these 70 songs, the results of the user experiment agrees with the results from our k -NN classification for 24 songs.

After re-running k -NN classification experiments with the updated ground truth, to our surprise, the classification accuracy did not improve much: it went from 80.8% to 81.5%, meaning only 7 more songs were correctly classified, despite the ground truth for 24 songs being changed to exactly as the k -NN classifier had previously classified them. Additionally, 86 tracks had the correct genre in 2nd place, for a total of 90.1% in 1st or 2nd place. This is an increase of only one song compared with the unmodified ground truth.

The reason for this limited improvement is that in many cases the vote difference of the k -NN classifier between the genres in 1st and 2nd place is very low, so several songs that were correctly classified when using the unmodified ground truth definition changed to being incorrectly classified using the modified ground truth definition. It is also worth pointing out that we only had the participants of our experiment listen to the songs that were originally incorrectly classified. If we were to actually change the ground truth in order to make each genre more coherent we would need to perform a larger-scale study to investigate the entirety of 1,000 songs.

5 Conclusion

We have seen through a number of experiments that the evaluation of the results from automatic genre classification systems is not as simple as it might seem. This confirms the findings of prior work which already took a critical view on genre classification and genre ground truth. Just because the classification of a given song does not agree with a given ground truth classification does not necessarily mean it is wrong. Given the subjective nature of genre classification, and how artists sometime merge two or multiple known genres, there are many situations where two or more genres might be appropriate for a given song. We have also seen that changing the ground truth increased our accuracy for 7 songs out of the 1,000. We conclude that in order to create a working automatic genre classification system much more emphasis has to be put on the ground truth creation and analysis, and evaluation of the results of such systems need to be much more than simply calculating a percentage of how many of the top genres agree with a given ground truth. We agree with [2] that one good way of such evaluations could be to weight the results from such systems to reflect the amount of human classification ambiguity of the same dataset.

Acknowledgements

Supported by the Austrian Science Fund (FWF): P25655 and the Austrian FFG: BRIDGE 1 project *SmarterJam* (858514).

References

1. Aucouturier, J.J., Pachet, F.: Representing musical genre: A state of the art. *Journal of New Music Research* 32(1), 83–93 (2003)
2. Craft, A., Wiggins, G., Crawford, T.: How many beans make five? the consensus problem in music-genre classification and a new evaluation method for single-genre categorisation systems. In: *Proceedings of the 8th International Symposium on Music Information Retrieval (ISMIR)*. Vienna, Austria (September 2007)
3. Drott, E.: The End(s) of Genre. *Journal of Music Theory* 57(1), 1–45 (2013)
4. Gjerdingen, R.O., Perrott, D.: Scanning the dial: The rapid recognition of music genres. *Journal of New Music Research* 37(2), 93–100 (2008)
5. Guðmundsson, G.P., Jónsson, B.P., Amsaleg, L.: A large-scale performance study of cluster-based high-dimensional indexing. In: *ACM Multimedia Workshop on Very-Large-Scale Multimedia Corpus, Mining and Retrieval*. Florence, Italy (2010)
6. Krumhansl, C.L.: Plink: “Thin Slices” of Music. *Music Perception: An Interdisciplinary Journal* 27(5), 337–354 (June 2010)
7. Lippens, S., Martens, J.P., De Mulder, T., Tzanetakis, G.: A comparison of human and automatic musical genre classification. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2004)
8. McKay, C., Fujinaga, I.: Musical Genre Classification: Is It Worth Pursuing and How Can It Be Improved? In: *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR)*. Victoria, BC, Canada (October 2006)
9. Novello, A., McKinney, M.F., Kohlrausch, A.: Perceptual Evaluation of Music Similarity. In: *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR)*. Victoria, BC, Canada (October 2006)
10. Pachet, F., Cazaly, D.: A Taxonomy of Musical Genre. In: *Proceedings of Content-Based Multimedia Information Access (RIAO) Conference*. Paris, France (2000)
11. Scaringella, N., Zoia, G., Mlynek, D.: Automatic Genre Classification of Music Content: A Survey. *IEEE Signal Processing Magazine* 23(2), 133–141 (March 2006)
12. Schedl, M., Flexer, A., Urbano, J.: The neglected user in music information retrieval research. *Journal of Intelligent Information Systems* 41, 523–539 (December 2013)
13. Seyerlehner, K., Widmer, G., Knees, P.: A Comparison of Human, Automatic and Collaborative Music Genre Classification and User Centric Evaluation of Genre Classification Systems. In: *Detyniecki, M., Knees, P., Nürnberger, A., Schedl, M., Stober, S. (eds.) Adaptive Multimedia Retrieval: Context, Exploration, and Fusion, LNCS, vol. 6817*. Springer (2011)
14. Sturm, B.L.: An Analysis of the GTZAN Music Genre Dataset. In: *Proceedings of the 2nd International ACM Workshop on Music Information Retrieval with User-centered and Multimodal Strategies (MIRUM)*. Nara, Japan (October–November 2012)
15. Sturm, B.L.: The State of the Art Ten Years After a State of the Art: Future Research in Music Information Retrieval. *Journal of New Music Research* 43(2), 147–172 (2014)
16. Tzanetakis, G., Cook, P.: Musical Genre Classification of Audio Signals. *IEEE Transactions on Speech and Audio Processing* 10(5), 293–302 (July 2002)