A Comparison of Human, Automatic and Collaborative Music Genre Classification and User Centric Evaluation of Genre Classification Systems

Klaus Seyerlehner[1], Gerhard Widmer[1,2], and Peter Knees[1]

[1] Dept. of Computational Perception, Johannes Kepler University, Linz, Austria
http://www.cp.jku.at
[2] Austrian Research Institute for AI, Vienna, Austria
http://www.ofai.at

Abstract. In this paper two sets of evaluation experiments are conducted. First, we compare state-of-the-art automatic music genre classification algorithms to human performance on the same dataset, via a listening experiment. This will show that the improvements of content-based systems over the last years have reduced the gap between automatic and human classification performance, but could not yet close this gap. As an important extension to previous work in this context, we will also compare the automatic and human classification performance to a collaborative approach. Second, we propose two evaluation metrics, called *user scores*, that are based on the votes of the participants of the listening experiment. This user centric evaluation approach allows to get rid of predefined ground truth annotations and allows to account for the ambiguous human perception of musical genre. To take genre ambiguities into account is an important advantage with respect to the evaluation of content-based systems, especially since the dataset compiled in this work (both the audio files and collected votes) are publicly available.

Keywords: genre classification, user centric evaluation

1 Introduction

Although genre definitions and annotations are somewhat subjective, genre categorizations or genre hierarchies are often used to organize large scaled music collections, as there seems to be some general consensus on genre annotations, at least to a certain degree. In music information retrieval (MIR), genre labels often serve as ground truth information, most notably to evaluate automatic genre classification systems, music similarity algorithms and music recommender systems. While publicly available genre classification datasets and also the annual ***M**usic **I**nformation **R**etrieval **E**valuation e**X**change* (MIREX) [3] make the numerous proposed systems more comparable to each other in terms of quality,

[3] http://www.music-ir.org/mirexwiki

there exists little work on making the evaluated systems comparable to human performance on the same task. To improve the comparability of automatic and human classification accuracy, we have conducted a listening experiment. This allows to compare the classification results of human listeners to those of state-of-the-art automatic genre classification algorithms. Furthermore, we will show that the collaborative result of the participants outperforms both automatic methods and individual human performance. While the collaborative result can be regarded as an upper bound on the achievable classification accuracy on this dataset, it also shows that collaborative techniques clearly outperform content-based approaches. Furthermore, the dataset containing both the full length tracks and the genre votes by the participants is publicly available from the first author's personal webpage. This will be useful to improve the evaluation of genre classification algorithms, because on the basis of such data one can define user centric evaluation metrics - so called *user scores*. The main advantage of user centric evaluation metrics is that one can account for genre ambiguities derived from the user votes whenever two automatic systems are compared.

The rest of the paper is organized as follows. First, in section 2 we report on the conducted listening experiment and point out the difference to the only related work by Lippens et al. [10]. In section 3 we then present the results obtained by the individual participants, briefly introduce five automatic classification methods and two collaborative approaches and compare the performance of these approaches to the performance of the individual participants. In section 4 we then discuss how the collected genre information can also be used to define two user centric evaluation metrics and present results for the automatic classification methods using the proposed evaluation criteria. Finally, we conclude on the obtained results in section 5.

2 The Listening Experiment

In general genre as an evaluation criterion is a well-discussed topic [6][4][18][3][11] and it is broadly accepted in Music Information Retrieval (MIR) as an evaluation criterion for content-based systems. Thus, there exist numerous publications focuses on comparing automatic systems to each other using genre information. There also exists some scientific work on evaluating the human abilities to classify music into genres. Most notably Gjerdingen et al. in [7] showed that humans are very fast at classify music into genres. About 300ms of audio are enough for humans to come up with the same categorization decision as with 3000ms of audio. Bella et al. in [2] investigated the human ability to classify classical music into sub-genres. Furthermore, Guaus et al. [8] study the effect of rhythm and timbre modifications on the human music genre categorization ability. They find that timbre feature provide more genre discrimination power than rhythm.

However, there exists little work on **comparing** automatic to human performance on the same genre classification task. In [17] Soltau et al. mentioned that the genre confusions of a conducted listening experiment are similar to those of a proposed automatic system, but no evaluation to directly compare human to

automatic performance was conduced. The only work that really focuses on a comparison of human to automatic classification performance we are aware of is the work of Lippens et al. [10] and dates back to 2004. In [10], a listening experiment is conducted were 27 human listener manually classified a collection of 160 songs (the "MAMI dataset"), into 6 possible genres by listening to 30 seconds excerpts. The average performance of the participants (76%) is then compared to an automatic classification approach with a classification performance of 57%, and the baseline accuracy (26%). Unfortunately, the MAMI dataset and the survey data are not publicly available. To be able to also compare state-of-the-art systems to human classification, we decided to rerun a listening experiment quite similar to the one presented in [10]. In this listening experiment 24 persons were asked to do exactly the same task the machine was ask to solve, namely to categorize a set of songs into 19 genres. The participants of this survey were aged in between 20-40 and most of them had no specific musical background, but can be characterized as typical mainstream music consumers. The songs were drawn randomly from the *"1517-Artists"* dataset [15] in such a way that each genre is represented by 10 songs. The *"1517-Artists"* dataset itself consists of freely available songs from download.com[4][5] containing songs of both well-known and completely unknown artists. The genre labels were assigned by the artists of the songs. The genres and the number of tracks per genre of the subset used in the listening experiment are summarized in table 1. While it seems that just selecting 10 song per genre is at the lower bound for a descriptive subset of a genre, the number of songs that can be used in such a listening experiment is of course limited by the available human resources. In our case many of the participants of the listening experiments reported that it took them many hours to complete the survey and far longer as expected.

Comparing the conducted listening experiment presented in this paper to the listening experiment in [10] there are some important differences in the data, the design of the experiment, and the analysis of the results:

- **Unique Artists**
  To prevent artist effects and album effects [5], no two songs by one and the same artists are in the dataset used for the listening experiment. This is very important as artist and album effects can have a huge biasing influence on the obtained classification accuracies, especially on small datasets.
- **Number of Genres**
  The number of genres (19) in our listening experiment is significantly larger, and the musical scope is broader than in the MAMI dataset.
- **Equal number of tracks per genres**
  Each genre is represented by 10 representative songs, making this a balanced classification task that is not biased towards a popular, dominating genre like e.g. *"Pop&Rock"*.

[4] http://music.download.com/

[5] The `http://music.download.com/` began redirecting all artist pages and category doors to corresponding pages on their sister music site Last.fm on March 2009.

– **Explicit Genre Annotations**
  There exists a ground-truth genre label per songs that has been assigned by the artists that produced the songs via the music platform. The genre categories are the same as used by the music platform[6].
– **Publicly Available Data**
  The music files used in the presented experiment and the genre votes obtained through the listening experiment are both publicly available.[7] This will allow others to compare other methods not presented here to human performance in the future.
– **Collaborative Result**
  In section 3.3 the votes of the subjects are used to collaboratively estimate a song's genre. Thus, we are able to also compare the collaborative result of all subjects to both individual results as well as automatic classification systems.

It is important to note that we do not claim that the genre annotations of this dataset are particularly correct or that the genre taxonomy is perfectly consistent. In contrast we belief that genre and genre taxonomies by definition are ambiguous and inconsistent and good genre taxonomies need a careful design and should account for genre similarities [12]. However, it is important to see that for comparative evaluations like we perform in this paper annotation errors are not crucial as all evaluated approaches have to deal with the same annotation errors. With respect to genre inconsistencies we propose in section 4 to use so-called *user scores* as evaluation criteria, which allow to account for existing genre ambiguities.

The experiment was carried out as follows: Each participant was instructed to move the 190 anonymized full-length audio files into a set of folders representing the 19 genres, plus an extra folder *"other"* in case they had no idea what genre a song might belong to. Then a list of the files in the directory structure representing the genres was generated by a script and returned by each subject via e-mail. Finally, these files were parsed to obtain the votes of each individual.

3 Human, Automatic and Collaborative Classification

3.1 Human Classification

The collected information from the listening experiment is represented as a set T of tuples $t = (u_t, s_t, \hat{g}_t, g_t)$, where u_t (1 to 24) identifies the participant and s_t (1 to 190) the rated song. The ground truth genre of the song s_t is denoted $g_t \in G$, where G is the set containing the 19 ground truth genres. $\hat{g}_t \in G^+$ represents the genre predicted by participant u_t. G^+ is the set of genres plus the *"other"* category. The classification accuracy of subject u with respect to the

[6] music.download.com
[7] www.seyerlehner.info

Genre	#tracks
Blues	10
Country	10
Hip-Hop	10
Jazz	10
New Age	10
Reggae	10
Classical	10
Folk	10
Latin	10
Rock & Pop	10
Alternative & Punk	10
Electronic & Dance	10
R&B & Soul	10
World	10
Vocals	10
Children's	10
Easy Listening	10
Comedy & Spoken Word	10
Soundtracks & More	10
total	**190**

Table 1. *Genre distribution of the songs used in the listening experiment.*

given ground truth annotation is then given by

$$acc_u = \frac{\sum_t^{\{t \in T | u_t = u\}} \hat{g}_t == g_t}{|\{t \in T | u_t = u\}|} \tag{1}$$

A look at Figure 1 shows that there is a huge variation in the performance of individual participants. Obviously the individual results heavily depend on the musical knowledge of the individuals. The worst participant exhibits a classification accuracy of 26%, which is still far better than the baseline (guessing), which would be 5%. The classification rate of the best individual is 71%. The average classification accuracy obtained by the participants is 55%, the median is also 55%. Figure 1 visualizes the classification accuracies achieved by the individual participants sorted from the worst to the best participant.

Aggregating the individual results of all users yields the overall classification result. Figure 2 shows the confusion matrix with respect to the ground truth. Altogether 55% of all song-genre assignments of the participants were correct. However the performance depends on the genre. While some genres seem to be well-defined (e.g. *"Comedy&Spoken Word"*, *"Electronic&Dance"*, *"Hip-Hop"*), there is almost no agreement among the participants for the genres *"Folk"* and *"Vocals"*. For the other genres the participants agree to a certain extent. The most significant genre confusions are *"Folk"* - *"Vocals"*, *"Alternative&Punk"* - *"Rock&Pop"*, *"EasyListening"* - *"NewAge"*, *"Country"* - *"Folk"*, *"Blues"* -

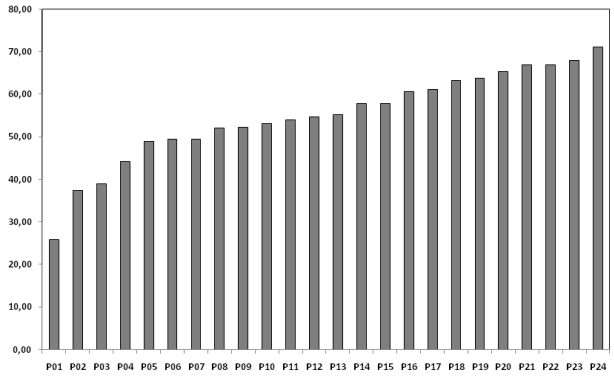

Fig. 1. *Ordered classification accuracies of the participants.*

"Jazz", *"Reggae"* - *"Hip-Hop"* and *"Latin"* - *"EasyListening"* and vice versa. These confusions indicate genre ambiguities, but can also be interpreted as some sort of genre similarities. Also, many genre pairs are never or extremely rarely confused, which implies that it is very easy for humans to distinguish these genres. Based on the user votes one can define the genre-song voting matrix $V = (v_{g,s})$, where $v_{g,s}$ denotes the number of times the participants voted for genre g given song s:

$$v_{gs} = \sum_{t}^{\{t \in T | s_t = s\}} \hat{g}_t == g \tag{2}$$

The genre-song voting matrix is visualized in figure 4. One can even visually see that the majority of the participants agree with the ground truth information for most of the songs. In contrast to the confusion matrix, the genre-song voting matrix visualizes the classification result for each song separately and is a compact representation of the results of the listening experiment. To further analyze the votes one can define the number of different genres $D(s)$ the participants have assigned to a specific song s:

$$D(s) = \sum_{g}^{G^+} v_{gs} > 0 \tag{3}$$

Figure 3 (left) shows a histogram of the number of different genres $D(s)$ the user voted for. Although there are 20 options to choose from, in general the participants did not vote for more than 8 different genres. This indicates that some genres are not relevant at all for some songs. Furthermore we can identify the most frequently estimated genre, the second most frequently estimated genre and so on, for each song. Then we can aggregate the number for votes for the k (1 to 20) most frequently estimated genre over all songs. The percentage of the accumulated votes relative to the total number of votes is visualized in figure

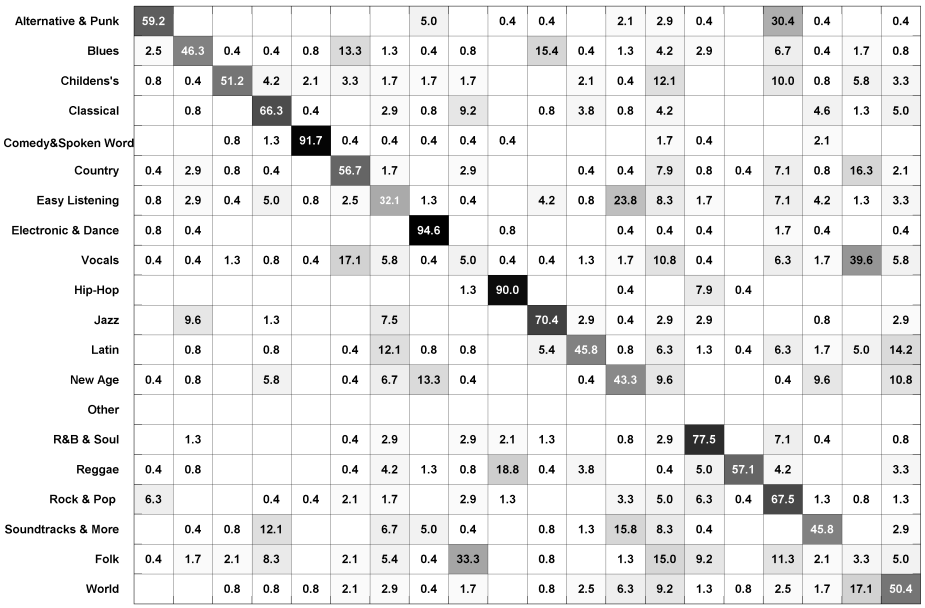

Fig. 2. *Confusion matrix of the classifications resulting from the experiment with respect to the ground truth annotation. Entry i, j is the percentage of user votes that predicted class j when the true class was i.*

3 (right). Consistently with the histogram in figure 3 all votes are within the 12 most frequently estimated genres. In general there exists a strong consensus among the participants on a song's genre. The most frequently predicted genre for each song is responsible for 64% of all votes. The two most frequently predicted genres of each song, together represent 80% of all votes (see figure 3). Therefore, we can conclude that the majority of the participants strongly agree on just one or two possible genre assignments for most of the songs.

3.2 Automatic Classification

To compare human to automatic classification performance we will use five different automatic classification methods. The choice of the evaluated approaches contains classical, well-known and state-of-the-art systems. Only complete genre classification systems as proposed in the literature are evaluated. Thus, the evaluated systems extract different feature sets and are based on different classification approaches. Two of the evaluated classification systems (SG-NN and RTBOF-NN) are based on nearest neighbor classifiers. The other three algorithms (GT-SVM, BLF1-SVM, BLF2-SVM) are based on a support vector machine classifier. The reported classification accuracies are obtained via leave-one-out cross-validation. The automatic classification methods are briefly described below.

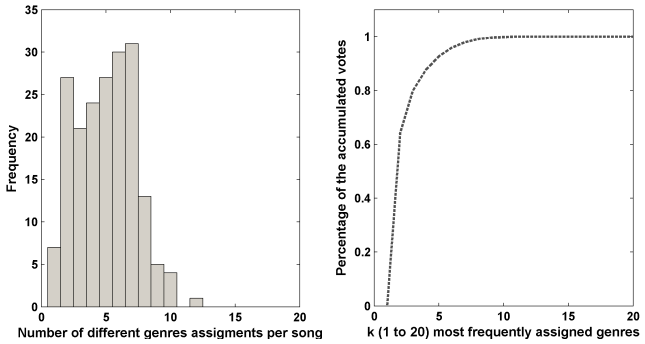

Fig. 3. *Histogram of the number of different genres per song the participants have voted for (left) and percentage of the accumulated number of votes for the k most frequently assigned genres per song (right).*

Single Gaussian (SG-NN) The *Single Gaussian Nearest Neighbor Classifier* (SG-NN) is based on the so-called Bag of Frames (BOF) approach [1]. Each song is modeled as a distribution of Mel Frequency Cepstrum Coefficients (MFCCs). A single multivariate Gaussian distribution is used to model the distribution of MFCCs of a song. To identify the nearest neighbors the Kullback-Leibler (KL) divergence between two models is computed. This approach is a fast and popular variant proposed by Levy et al. [9] of the classic timbre based audio similarity measure.

Rhythm Timbre Bag of Features (RTBOF-NN) The *Rhythm Timbre Bag of Features Nearest Neighbor Classifier* (RTBOF-NN) is a state-of-the-art music similarity measure proposed by Pohle et al. in [13]. This measure ranked first in the MIREX 2009 music similarity and retrieval task and has proven to be statistically significantly better than most of the participating algorithms. In contrast to the classic Single Gaussian approach this RTBOF-NN Classifier reflects the current state-of-the-art in nearest neighbor classification. Basically, it has two components – a rhythm and a timbre component. Each component, rhythm and timbre, consists of a distribution model over local spectral features. The features, described in [13], are complex and incorporate local temporal information over several frames. Because of its components we will call this approach Rhythm Timbre Bag Of Features (RTBOF) in our evaluations.

Block-Level Feature (BLF-SVM) The Block-Level Feature Support Vector Machine approach (BLF-SVM) is a genre classification algorithm based on block-level features. An earlier version of this algorithm [14] participated in the MIREX 2009 Audio Genre Classification task and took rank 14 out of 31. However, no statistically significant difference to the winning algorithm was found. This approach will be denoted BLF1-SVM. Additionally, we also evaluate an im-

proved variant of this algorithm, which we call BLF2-SVM here. This algorithm includes three novel block-level features (Spectral Contrast Pattern, Correlation Pattern and Variance Delta Spectral Pattern). For a detailed description of these new feature we refer to [16]. This improved approach is expected to perform comparably to the state-of-the-art methods in genre classification.

Marsyas (MARSYAS-SVM) The Marsyas (Music Analysis, Retrieval and Synthesis for Audio Signals) framework[8] is an open source software that can be used to efficiently calculate various audio features. For a detailed description of the extracted features we refer to [19]. This algorithm has participated in the MIREX Genre Classification task from 2007 onwards, and the features as well as the classification approach have been the same over the years. We use the framework to extract the features exactly as for the MIREX contest (MARSYAS version 0.3.2). Then we use the WEKA Support Vector Machine implementation to perform cross-validation experiments. This method is closest to the automatic approach by Lippens et al. [10] and should help to make our experiment more comparable to this previous experiment.

3.3 Collaborative Classification

In this section we present two straight-forward collaborative classification approaches (CV and CSS-NN) based on the users' aggregated votes.

Collaborative Voting (CV) The Collaborative Voting (CV) approach is simple. The genre most participants have voted for is the predicted genre of a song. This method basically combines the individual classification results of the participants following the majority rule like a meta-classifier.

Collaborative Filtering (CF-NN) The Collaborative Filtering Nearest Neighbor Classifier (CF-NN) is related to an item-based collaborative filtering approach. Each song is represented by its voting profile, which corresponds to the column vector of a song in the genre-song voting matrix (see figure 4). One can then derive song similarities by comparing the voting profiles of the songs. To compare song profiles the *city-block* distance (l_1 *norm*) was used in our experiments. The song similarity information can then be used to perform nearest neighbor classification.

3.4 Comparison

In figure 5 the classification results of the automatic methods, the collaborative approaches and the individual results of the participants are visualized together, sorted according to the achieved accuracy. Clearly, the content-based

[8] http://marsyas.info

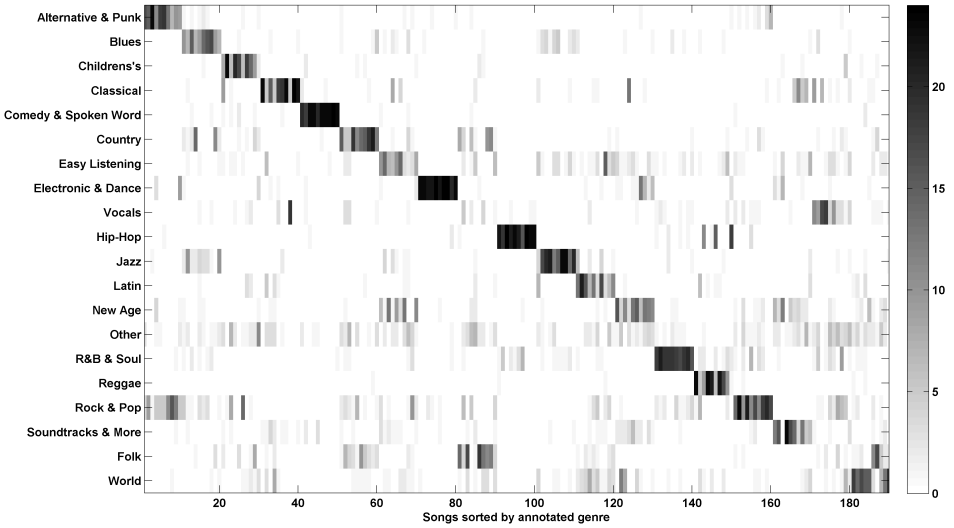

Fig. 4. *Visualization of the genre-song voting matrix. Tracks are sorted according to the ground truth genre.*

approaches perform worse than most of the participants, whereas the collaborative approaches achieve high classification accuracies and outperform most of the participants. The observation that collaborative approaches do better than most individual humans can be explained by the fact that these type of algorithms better reflect the group opinion, which is the aggregated knowledge of many individuals. Therefore, the group as a whole has a broader musical knowledge than any individual, as each person is typically familiar with some but not with all genres of a classification dataset.

Comparing the best content-based approach (BLF2-SVM) to the best collaborative approach (CF-NN) it turns out that the latter achieves almost double the classification accuracy of the content-based approach. Taking a look at the various content-based method, we can see that there exist clear differences. The classical timbral similarity measure performs worst, just outperforming the worst participant. The classic MARSYAS-SVM approach does not perform much better, which slightly contradicted our expectations.[9] Both recent methods RTBOF and BLF2-SVM show an improvement in classification accuracy over the 'classic' approaches. This indicates that the improvements in automatic classification reduced the gap between human and automatic classification, but still there exists a difference of about 10 percentage points between the best automatic method and the average human participant. Furthermore, based on the obtained re-

[9] Interestingly, when performing a 10-fold cross-validation instead of leave-one-out, we get comparable results for MARSYAS-SVM and BLF1-SVM. This effect is yet to be investigated.

sults we can define an upper bound on the achievable classification accuracy for automatic methods on this dataset. Clearly because of inconsistencies of the classification taxonomy and possible annotation errors none of the evaluated methods will ever reach perfect classification accuracy. However, as all evaluated methods have to deal with these problems the classification result of the CF-NN approach can be interpreted as an upper bound for automatic methods on this dataset.

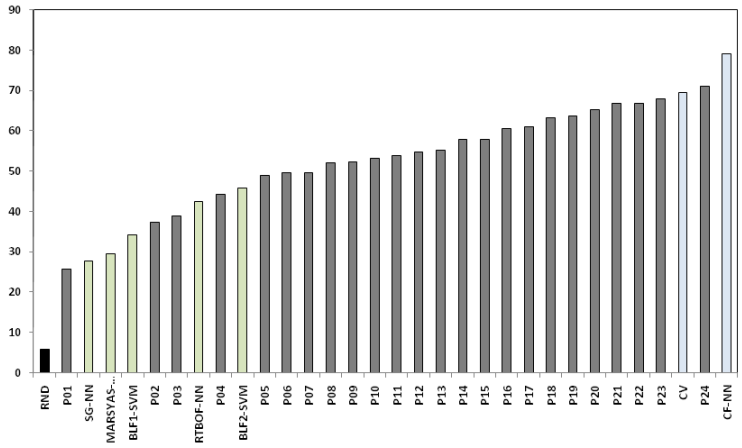

Fig. 5. *Comparison of classification results of the individual participants, automatic methods and the collaborative approaches.*

4 Evaluation based on User Data

One of the main disadvantages of using the classification accuracy as evaluation criterion is that such experiments heavily depend on the quality of the ground truth annotations. To improve the quality of the ground truth one can of course ask an expert to define the genre annotations, but still the evaluation would just depend on a single opinion and as already pointed out there will always exist some annotation errors due to the inconsistency of the genre taxonomy itself.

To overcome these limitations we propose to perform a user centric evaluation by aggregating the collected genre votes of the participants of the listening experiment. Thus, the ground truth is no longer based on a single opinion, but on the aggregated opinions of all the participants regarding the genre affinity of a given song. This way we can not only use the obtained data from the listening experiment to make automatic classification methods comparable to human classification performance, but this information can also be used to account for

genre ambiguities whenever genre classification is used in an evaluation, as already proposed in [3] and [10]. The basic idea for such a quality measure is straight-forward: If even humans are unsure about a genre label then it will be hard for the machine to get the label right.

To reflect these uncertainties of the genre annotations in a quality measure, a *user score* is defined similarly to [10]. A user score measures the agreement of the predictions of an automatic method with the genre assignments of the humans participating in the listening experiment. Thus, any algorithm can collect points for each song s in the dataset according to the agreement with the user votes. In particular, for each song $s \in S$ the classification of the algorithm into genre $\hat{g}_s \in G$ is rated by the number of times this genre was voted for ($v_{\hat{g}_s,s}$) relative to the number of times the participants voted for the most frequently predicted genre ($\max(\{v_{g,s}|g \in G\})$).

$$\text{US1} = \frac{1}{|S|} \sum_{s}^{s \in S} v_{\hat{g}_s,s} / \max(\{v_{g,s}|g \in G\}) \tag{4}$$

Extending the idea in [3], another straight-forward definition of a *user score* — this score is denoted US2 — is to take the number of collected points relative to the maximum number of points one can obtain on the dataset.

$$\text{US2} = \sum_{s}^{s \in S} v_{\hat{g}_s,s} / \sum_{s}^{s \in S} \max(\{v_{g,s}|g \in G\}) \tag{5}$$

The difference of the two scores is that for US1 each song contributes equally, whereas for US2 it is more important to correctly predict songs where the participants agreed pretty much on a single genre. One important advantage of both user scores is that they no longer rely on the ground truth annotation, but are solely based on the user ratings. By definition both scores are in the range between 0 and 1.

Approach	**US1**	**US2**	**acc.**
BLF2-SVM	0.5615	0.5080	0.4579
RTBOF-NN	0.4352	0.3827	0.4253
BLF1-SVM	0.3672	0.3382	0.3421
MARSYAS-SVM	0.3217	0.3031	0.2953
SG-NN	0.3156	0.2791	0.2779
RND	0.0578	0.0673	0.0584

Table 2. *Comparison of the user scores (US1, US2) and the classification accuracy (acc.) obtained for the automatic approaches presented in section 3.2.*

Table 2 summarizes the user scores and the classification accuracy for the automatic classification methods presented in section 3.2. To our knowledge this

is the **first comparison** of automatic classification methods also accounting for genre ambiguities in the literature. The ranking of the analyzed algorithms is the same for all quality criteria. However, taking genre ambiguities into account clearly changes the evaluation result. For example the difference between the BLF2-SVM and the RTBOF-NN is relatively bigger for the users scores compared to the classification accuracy. An improvement of a user score over the classification accuracy reveals that the misclassified songs are not classified into an arbitrary, completely unrelated genre, but into a genre that users find similar, or tend to confuse also. We advocate this method for future evaluations of genre classifiers, whenever appropriate data are available.

5 Conclusions

Based on the evaluation results presented in section 3.4, we can conclude that there is some progress with respect to automatic genre classification methods, reducing the gap between automatic methods and human classification. However, the best performing automatic method in our experiment still performs about 10 percentage points worse than the average human participant. Furthermore, we could also show that the collaborative approach outperforms both automatic methods as well as individual human performances. Thus, collaboratively collecting meta-information about music e.g. via a music platform is a very powerful method and is also the clear trend in the music business. For content-based methods this implies that they are only beneficial in situations where no other data is available – for instance, in cold start situations, or in special application scenarios where no access to collaboratively collected meta-data is possible. Additionally, with respect to the evaluation of content-based systems we have proposed two user centric evaluation criteria. The proposed user-scores no longer depend on a single ground truth annotation, but on the aggregate opinion of the participants of the conducted listening experiment. One advantage of the proposed user-scores is that they account for genre ambiguities which will help to improve the evaluation of automatic classification systems in future, especially since the whole dataset (including both the audio files and the collected votes) is publicly available.

6 Acknowledgments

This research was supported by the Austrian Research Fund (FWF) under grant L511-N15. We especially thank all the participants of the listening experiment. It took many of them hours to complete the survey.

References

1. Aucouturier, J.J., Defreville, B., Pachet, F.: The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music. The Journal of the Acoustical Society of America (2007)

2. Bella, S.D., Peretz, I.: Differentiation of classical music requires little learning but rhythm. Cognition (2005)
3. Craft, A., Wiggins, G.A., 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: Proc. Int. Sym. on Music Information Retrieval (ISMIR-07) (2007)
4. Ellis, D., Whitman, B., Berenzweig, A., Lawrence, S.: The quest for ground truth in musical artist similarity. In: Proc. of the 3rd international Conference on Music Information Retrieval (ISMIR-02) (2002)
5. Flexer, A., Schnitzer, D.: Album and artist effects for audio similarity at the scale of the web. In: Proc. of the 6th Sound and Music Computing Conference (SMC-09) (2009)
6. Geleijnse, G., Schedl, M., Knees, P.: The quest for ground truth in musical artist tagging in the social web era. In: Proc. of the 8th International Conference on Music Information Retrieval (ISMIR-07) (2007)
7. Gjerdingen, R., Perrott, D.: Scanning the dial: The rapid recognition of music genres. Journal of New Music Research (2008)
8. Guaus, E., Herrera, P.: Music genre categorization in humans and machines. In: 121th AES Convention (2006)
9. Levy, M., Sandler, M.: Lightweight measures for timbral similarity of musical audio. In: AMCMM '06: Proceedings of the 1st ACM workshop on Audio and music computing multimedia. pp. 27–36. Santa Barbara, California, USA (2006)
10. Lippens, S., Martens, J., Mulder, T.D., Tzanetakis, G.: A comparison of human and automatic musical genre classification. In: Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP-04) (2004)
11. McKay, C., Fujinaga, I.: Musical genre classification: Is it worth pursuing and how can it be improved? In: Proc. of the 7th Int. Conf. on Music Information Retrieval (ISMIR-06) (2006)
12. Pachet, F., Cazaly, D.: A taxonomy of musical genre. In: Proc. of Content-Based Multimedia Information Access Conference (RIOA) (2000)
13. Pohle, T., Schnitzer, D., Schedl, M., Knees, P., Widmer, G.: On rhythm and general music similarity. In: Proc. of the 10th International Society for Music Information Retrieval Conference (ISMIR-09) (2009)
14. Seyerlehner, K., Schedl, M.: Block-level audio feature for music genre classification. In: online Proc. of the 5th Annual Music Information Retrieval Evaluation eXchange (MIREX-09) (2009)
15. Seyerlehner, K., Widmer, G., Knees, P.: Frame level audio similarity - a codebook approach. In: Proc. of the 11th International Conference on Digital Audio Effects (DAFx-08) (2008)
16. Seyerlehner, K., Widmer, G., Pohle, T.: Fusing block-level features for music similarity estimation. In: Proc. of the 13th International Conference on Digital Audio Effects (DAFx-10) (2010)
17. Soltau, H., Schultz, T., Westphal, M., Waibel, A.: Recognition of music types. In: Proc. of the 23th IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP-98) (2010)
18. Sordo, M., Celma, O., Blech, M., Guaus., E.: The quest for musical genres: Do the experts and the wisdom of crowds agree? In: Proc. of the 9th International Conference on Music Information Retrieval (ISMIR-08) (2008)
19. Tzanetakis, G., Cook, P.: Musical genre classification of audio signal. IEEE Transactions on Audio and Speech Processing (2002)