HOW MANY BEANS MAKE FIVE? THE CONSENSUS PROBLEM IN MUSIC-GENRE CLASSIFICATION AND A NEW EVALUATION METHOD FOR SINGLE-GENRE CATEGORISATION SYSTEMS

Alastair J. D. Craft, Geraint A. Wiggins, Tim Crawford
Intelligent Sound and Music Systems, Centre for Cognition, Computation and Culture
Goldsmiths, University of London
{a.craft,g.wiggins,t.crawford}@gold.ac.uk

ABSTRACT

Genre definition and attribution is generally considered to be *subjective*. This makes evaluation of any genre-labelling system intrinsically difficult, as the ground-truth against which it is compared is based upon subjective responses, with little inter-participant consensus. This paper presents a novel method of analysing the results of a genre-labelling task, and demonstrates that there are *groups of genre-labelling behaviour* which are self-consistent. It is proposed that the evaluation of any genre classification system uses this modified analysis method.

1 INTRODUCTION

Several genre-classification systems have been proposed in the literature (surveyed in [1, §4.2]). There has not been a corresponding interest in the evaluation of such systems, or in the actual genre-labelling behaviour of people. Whilst it is generally accepted that genre labels are subjective, with little industry [6], or inter-participant [3] consensus, this is rarely, if ever, included in the evaluation of any genre-categorisation system: most evaluations assume some ground truth, be it industry defined or just the categorisation of the experimenter.

Despite the acceptance of the ground-truth problems there has been relatively little work on how people categorise genre (see §3). Whilst this partially excuses experimenters' reliance on some form of absolute ground truth in system evaluation, it does mean that the results of any studies are questionable, as there is little understanding of what such systems are trying to model. In addition the most commonly referenced ground-truth study, [7], is as yet unpublished, and was not designed as a study of inter-participant consensus, but rather of how much audio a participant needed to establish a consistent genre label [2].

The approach we propose towards ground truth in genre classification is as follows: ground truth is an artefact of an individual's response to music, not an artefact of the audio itself. Therefore the establishment of any ground truth will be the study of *responses* to music, and is therefore predominately a *cultural* study. Any *unity* of response is because of the widespread agreed nature of the musical cues to genre in particular pieces, but the expected response from a group of individuals will be a *diversity*.

If there is a diversity of responses in terms of genre labels to any particular piece, or set of pieces, the standard evaluation methodology that uses single genres as ground truth will, necessarily, not describe all the dataset adequately.

2 WHAT TO EXPECT OF GENRE LABELS?

We propose there are two main factors at play in how a listener assigns a single genre label to a piece of music: the number of musical cues associated with different genres in the piece; and the participant's knowledge and experience of the genres involved. The first of these factors is a feature of what the composer intended of the piece in question: as an intentional act a composer can draw upon stylistic elements from one, or more, genres. In cases of this type the assignment of *multiple* or compound genre labels is justified.

The second factor is a feature of the social and cultural background of any particular participant: in order for a person reliably to label a piece with a genre they have to be cognisant of the stylistic features of that genre. If they are not then the assigned label can not be seen as reliable: it could be randomly assigned, or else a number of different genres could be conflated under one genre label.

The second factor can also influence the outcome of a genre-labelling task where the pieces involved are influenced by the first: if a participant has no knowledge of the stylistic features of contributory genres then they will label inaccurately, as above; if a participant has knowledge of the stylistic features of only one of the influencing genres then they are most likely to assign that genre label to it; if they are knowledgeable in all the contributory stylistic features then they will have to choose which is most determinant. In this case a participant's relative knowledge of each genre style will give rise to different answers.

The evaluation of genre labelling systems is highly dependent upon the play of these two factors. For the evaluation of genre labelling systems that only use individual

genres these cultural factors must be eliminated.

Attempts to address these cultural issues have been proposed, and are surveyed in [1, §4.5]. However, these do not address the issue that the cultural metric could produce widely differing results for different groups of people.

3 HUMAN GROUND-TRUTH STUDIES

There have been relatively few studies of human genre-labelling behaviour. We are only aware of four.

3.1 Studies with incomplete coverage of the dataset

There are two studies ([5] and [4]) which give incomplete coverage of the dataset, and a full analysis is given in [1, §4.6.1]. The reliability of [5] is questionable, as insufficient details of the experiment are provided, and the dataset is not adequately covered by all participants. Similarly [4] has incomplete dataset coverage in one of the two conditions, and therefore its use in any understanding of the ground-truth is precluded. The complete coverage under the other condition allows an analysis like that described in §4 to be applied.

3.2 Studies With Complete Coverage of the Dataset

3.2.1 Perrot and Gjerdingen (1999) [7]

In this much cited but still-unpublished study, 52 college students were presented with short excerpts of music of various lengths, and asked to categorise the excerpts into one of ten genres. In one of the conditions the participants' agreement with the classification provided by web-based CD-vendors was around 70%. The exact details of the experiment are given in [1, §4.6.2].

The 'results' of this study are those most frequently cited in the literature, and the 70% accuracy is often used as some sort of benchmark. Whilst the research and results may well be sound the experiment can only be viewed as anecdotal, since it is unpublished and the experimental data is unavailable [2].

3.2.2 Lippens et al. *(2004) [3]*

Lippens *et al.* [3] conducted a human labelling of the MAMI dataset, a collection of 160 full tracks of music. The tracks came annotated with 11 musical genres, but some of the genres were very poorly represented (*i.e.* had very few tracks) or were heterogeneous. As a result the authors conducted a user study with 6 genres (Pop, Rock, Classical, Dance, Rap and Other) to confirm that their definitions were consistent among different subjects [3, §3.1].

27 participants listened to the central 30 seconds of each track m and independently chose one genre s out of the 6 possibilities. In evaluating the genre-classification system the authors only used tracks that satisfied two conditions: the elected genre could be any but 'other'; the number of votes for the elected genre had to be greater than 18. The tracks which satisfied these conditions were used to construct a second dataset: MAMI2.

4 ANALYSIS OF LIPPENS *ET AL.*

The original analysis of these results is given in [3]. They required, for the purposes of their evaluation, for each track to have a unique genre. This has two adverse effects: an artificial worsening of the performance of human participants in the MAMI dataset; and a misunderstanding of the actual structure of the data.

We have re-analysed the data from this study. With reference to this new analysis we illustrate and address these two issues. [1]

4.1 Tracks With Multiple Elected Genres

There were 4 tracks which had multiple elected genres. For these four tracks only one category was used in the original analysis as the category for the track. If either of the genres that satisfy $G^{(m)}_{\max}$ [2] is allowed as the genre for a track then the percentage corresponding classification across the whole dataset is 77.3%, with standard deviation of 6.1% and minimum and maximum values of 59.38% and 86.88% respectively. This is different to the statistics presented in the original paper for the MAMI dataset. The ambiguity over these four tracks was not highlighted in [3], although they were removed in the MAMI2 dataset.

4.2 The *'Other'* Problem

In constructing the MAMI2 dataset [3] discarded any data where $Q^{(m)}_{\max} < 18$, [3] as these are "mainly tracks with many votes for 'other', and little consensus among the human listeners" [3, §3.1]. The handling of these tracks to form the MAMI2 dataset was not statistically well-founded.

4.2.1 What Does 'Other' Mean?

The main problem with the 'other' category is that its meaning is undefined to the participants. It can be interpreted in several ways: the category to assign to any track that does not fall into any of the other categories; the category to assign to any track that a particular participant does not know the genre of; or the category to assign to any track that uses features of two, or more, genres.

The first and third of these served the purposes of the experimenters well, as they eliminates tracks that do not fall into any genres, and those that fall into hybrid genres respectively. However, the second causes problems as it introduces noise into the data by demanding an ambiguous answer from the participants.

If, for instance, there is a two-way split between 'other' and another genre the possible interpretation of this data is dependant upon which of the first two interpretations are in play above. Those who voted for 'other' may not know the genre of the piece, or else realise that it is in a sub-genre of the alternate genre. It is therefore possible that tracks with many votes for 'other' *could* be strongly categorised if a different question had been asked of the participants.

[1] For the benefit of the reader we have used the same notation as used in the original study where appropriate.

[2] The genre with the maximum number of votes for a particular track

[3] $Q^{(m)}_{\max}$ is the number of votes for the elected genre for track m.

4.2.2 Votes For 'Other' in Tracks That Aren't in MAMI2

Figure 1 shows the distribution of votes for 'other' in tracks not in MAMI2. Although many of the tracks have many votes for 'other', many have few, or no votes for 'other'—the modal value is 0, and 36% of these tracks have less than 6 votes for 'other'. This is contrary to the statement that "these are mainly tracks with many votes for 'other' "[3, §3.1]

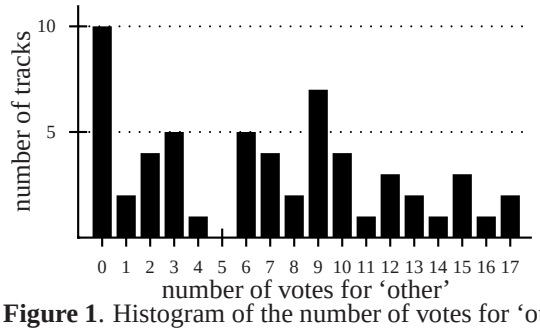

Figure 1. Histogram of the number of votes for 'other' in tracks not in MAMI2 dataset

The votes given for tracks that did not make it into the MAMI2 dataset, excluding those where $G(m) =$ 'other' are shown in Figure 2.

As can be seen from the data there is little subjectivity in the votes for some of these tracks: votes are typically distributed between two genres, rather than more randomly assigned across several genres. In many of the cases where there is a 3-genre split either one of the genres is 'other', which illustrates the possibility that 'other' is being used as a wild-card, or else one of the genres has one or two votes, and therefore may be a participant error. The majority of the tracks are voted for in ways similar to other tracks: whilst an individual vote may be subjective, these results *collectively* are only subjective if a unique genre is required to represent each track.

4.3 How many genres are identified?

The authors of [3] required a dataset which was unambiguous in order to evaluate their system. However, the method they used to create this dataset was not sufficient to guarantee unambiguous placement of tracks into unique genres.

In order for the desired criteria to be achieved we propose two necessary conditions:

1. In order for track m to form part of the set of tracks representative of a particular genre G it has to be voted for in a way that is *dissimilar* to the way tracks in a different genre are voted for.

2. In order for a track m to form part of the set of tracks representative of a particular genre G it has to be voted for in a way that is *similar* to the way tracks in the same genre are voted for.

The authors of [3] informally satisfied the first of these conditions, but tacitly assumed that satisfying the first condition would, necessarily, satisfy the second. We have used the Kolmogorov-Smirnov test for these purposes, which

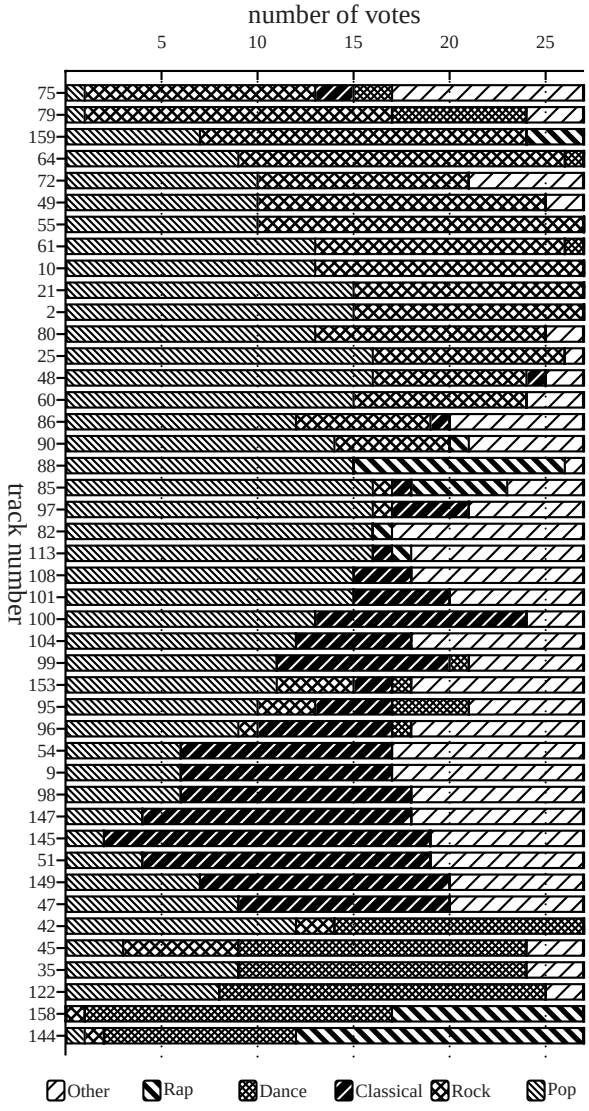

Figure 2. Distribution of votes for tracks with $Q_{\max}^{(m)} < 18$ and $G(m) \neq$'other'

determines whether two underlying probability distributions differ, based upon finite samples. It provides a measure of similarity between the two probability distributions.

Figure 3 shows the similarity between all tracks in the MAMI dataset as a similarity matrix, sorted so that tracks that have a similar distribution of genre votes are grouped together, within coarse boundaries that reflect $G(m)$ for each track. White areas show groups of tracks that are similar, black areas show tracks that are dissimilar. As can be seen in this graph there is strong structure in the ways people voted for different tracks.

Groups of tracks that are all similar to each other form squares that run along the main diagonal. Every actual genre in the study has such a group of tracks. However, there are groups of tracks which are similar but which span more than one genre and some global genre boundaries include more than one group of similar tracks.

From Figure 3 we infer that the participants identified 8 genres, based upon the assumption that tracks in the same genre will be voted for in a similar way, irrespective of

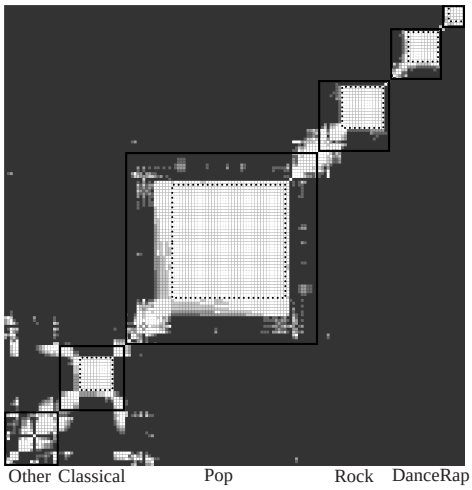

Figure 3. Similarity matrix for data collected by Lippens *et al.*. Heavy black boxes indicate genre majority votes, dotted black boxes indicate self-similar groupings not ambiguous with other genres.

how those votes are spread across the different choices.

4.4 Inclusion/exclusion criteria

If we wish to have an human-annotated test set with unambiguous genre-labels then tracks which are found to be equivalent to others in a different genre must be discarded. Similarly, any minor sub-genres within a genre that are identified should be discarded, as they betray the influence of other genres: there are 13 tracks in MAMI2 that should have been discarded.

5 TOWARDS AN IMPROVED TEST SET AND EVALUATION METHODOLOGY

Whilst it may be useful to have a genre labelling of a dataset that is entirely unambiguous, it is also unrealistic: a practial categorisation system needs to be able to deal with *all* tracks, not just unambiguous data.

We propose that the results of a genre classification system should be weighted to reflect the amount of ambiguity given to the genre labelling in a human task. Higher penalties should be incurred for misclassification of those tracks which the human participants unambiguously classify. If the system were to misclassify a track which is ambiguous in terms of a single genre the system should not be penalised for categorising into any of those genres. However, it should be penalised for categorising the track into any genre that is not one of those genres. This method would adequately cover around 85% of the MAMI dataset.

6 CONCLUSIONS

This paper has reviewed some important issues inherent in the evaluation of the genre classification task. We argue that researchers have not paid sufficient attention to the evaluation of the task as to the classification methods used, relying frequently on *ad hoc* methods of evaluation, typically against an anecdotal result [7]. However, at the same time authors have frequently stated that musical genre is inherently subjective, which calls such a method of evaluation to task.

We have proposed that there are cultures of genre-labelling behaviour, and that the structure of these cultures of behaviour needs to be better understood in order to evaluate properly the results of any system that models genre-labelling. We are currently undertaking a number of experiments to analyse these cultures of practice.

7 ACKNOWLEDGMENTS

We would like to thank all those with whom we have informally discussed these issues, in particular David Lewis and Daniel Müllensiefen, who also provided helpful feedback on this paper. We would also like to thank Matthias Varewyck for providing us with the data originally reported in [3] and Bob Gjerdingen for his clarification of various issues arising from [7].

8 REFERENCES

[1] Alastair J. D. Craft. The role of culture in the music genre classification task: human behaviour and its effect of methodology and evaluation. PhD Transfer Report, 2007. `http://www.doc.gold.ac.uk/~map01ac/transfer.pdf`.

[2] Robert O. Gjerdingen. Personal communication to Alastair Craft, June 2006.

[3] Stefaan Lippens, Jean-Pierre Martens, Tom De Mulder, and George Tzanetakis. A comparison of human and automatic musical genre classification. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICAASSP)*, volume 4, pages 223–236, 2004.

[4] Anders Meng, Peter Ahrendt, and Jan Larsen. Improving music genre classification by short-time feature integration. In *IEEE International Conference on Acoustics, Speech, and Signal Processing—Proceedings*, volume V, pages 497–500, March 2005.

[5] Anders Meng and John Shawe-Taylor. An investigation of feature models for music genre classification using the support vector classifier. In Joshua D. Reiss and Geraint A. Wiggins, editors, *Proceedings of the Sixth International Conference on Music Information Retrieval*, pages 604–609, London, September 2005. Queen Mary, University of London.

[6] François Pachet and Daniel Cazaly. A taxonomy of musical genres. In *Proceedings of the Content-Based Multimedia Information Access Conference (RIAO)*, April 2000.

[7] D. Perrot and R. O. Gjerdingen. Scanning the dial: An exploration of factors in the identification of musical style. Research notes, Department of Music, Northwestern University, 1999.