Music genre recognition with risk and rejection

Sturm, Bob L.

Published in:
International Conference on Multimedia and Expo

Publication date:
2013

Document Version
Early version, also known as pre-print

MUSIC GENRE RECOGNITION WITH RISK AND REJECTION

Bob L. Sturm

Audio Analysis Lab, Dept. Architecture, Design and Media Technology
Aalborg University Copenhagen, A.C. Meyers Vænge 15, DK-2450
bst@create.aau.dk

ABSTRACT

We explore risk and rejection for music genre recognition (MGR) within the minimum risk framework of Bayesian classification. In this way, we attempt to give an MGR system knowledge that some misclassifications are worse than others, and that deferring classification to an expert may be a better option than forcing a label under high uncertainty. Our experiments show this approach to have some success with respect to reducing false positives and negatives.

Index Terms— Music genre recognition, machine learning, Bayesian classification

1. INTRODUCTION

The problem of making a machine recognize kinds of cultural content has been explored in various mediums, from identifying literary authors [1], to categorizing "tweets" [2], automatically writing music reviews [3], characterizing painting styles [4], recognizing spoken dialects [5], and classifying the genre and mood of music [6]. A large amount of work addresses the specific problem of music genre recognition (MGR) [7]; and a recent review [6] appears to show significant progress has been made.

Our recent work [8, 9], however, challenges the claim that this progress is due to any real increase of the capacity of MGR systems to recognize genre. In [8], we take two systems exhibiting some of the highest classification accuracies reported for a particular benchmark dataset, and submit them to three tests. In the first, we show each system persistently makes misclassifications that are very poor with respect to musicological principles — explored further in [9]. For instance, one system persistently misclassifies as Metal, "Mamma Mia" by ABBA. In the second test, we show how each system can be tricked into classifying as different genres the same piece of music just by minor filtering of the signal. Finally, we show humans are unable to recognize the genres used by music excerpts composed by each system to be highly representative of the genres in which they have been trained. Whether or not an MGR system is really recognizing genre, these findings motivate the question: *can we modify an MGR system such that it avoids making very poor misclassifications?* In this paper, we explore this idea within the minimum risk framework offered by Bayesian classification [10].

Surprisingly little research in MGR explores the idea that some kinds of misclassification are worse than others. A few works, e.g., [11–19], argue that MGR systems should be evaluated in light of the specific confusions humans make. For instance, confusing metal for rock music is "better" than confusing metal for classical music. Weighting more heavily errors of the latter kind might better reflect the usability of an MGR system. This "discounted" approach to evaluation appears, for instance, in the 2005, 2007, and 2009 editions of the MGR challenge of the Music Information Retrieval Evaluation eXchange (MIREX) [20]. Though we do not find work directly implementing such knowledge in an MGR system, rather than in its evaluation, there is some work in music autotagging employing risk, e.g., that of Lo et al. [21]. That work, however, estimates a cost from a dataset, whereas we define it according to what a user deems offensive. Other areas, such as automatic speech recognition, have also applied risk minimization, e.g., [22].

Another aspect little explored in MGR are systems able to defer classification. For instance, when asked whether a piece of music is "Blues" or "Disco," acceptable answers include "don't know," and "something other" [11,12,15,16,23]. Nearly all MGR systems so far proposed are designed to choose only from those genres in which they are trained, but we find four exceptions. Dannenberg et al. [24] describe how they can reduce the false positives committed by their style-recognition system with a threshold on the minimum difference in distances between class means to observations. Pye [25] describes, but provides no details about, forming a "garbage model" by augmenting his dataset with songs outside of his six selected genres. Akin to this, Harb and Chen [26] also state they use a "garbage model," but do not discuss what music they use to define that class, and how it affects their system. Finally, the system of McKay [13] labels an excerpt "unknown" if the largest weighted sum of class-specific "scores" from component classifiers (neural networks) is too low, or not high enough with respect to other classes.

BLS is supported in part by Independent Postdoc Grant 11-105218 from Det Frie Forskningsråd; and the Danish Council for Strategic Research of the Danish Agency for Science Technology and Innovation in the CoSound project, case number 11-115328.

When an MGR system works by virtue of confounds and not musicological factors [8,9], it is certain to make unacceptable misclassifications. Since we find no work adequately addressing this problem, the opportunity arises to explore whether incorporating the concept of misclassification risk into an MGR system, and the option to defer classification, can reduce the number of "bad" decisions. In the next section, we review classification with loss and rejection from the point of view of MGR. The third section presents our experimental results.

2. CLASSIFICATION WITH RISK AND REJECTION

Consider an M-dimensional real observation $\mathbf{x}$ from a sample space of K classes. Bayesian classification [10] seeks the class with the minimum expected risk $R(k|\mathbf{x})$:

$$\hat{k} = \arg\min_{k\in\mathcal{K}} R(k|\mathbf{x}) = \arg\min_{k\in\mathcal{K}} \sum_{m=1}^{K} \ell_{km} P(m|\mathbf{x}) = \arg\min_{k\in\mathcal{K}} \mathbf{e}_k^T \mathbf{L}\mathbf{p}(\mathbf{x}) \quad (1)$$

where $\mathbf{e}_k$ is the kth standard basis vector, and we define the loss matrix and vector of posteriors

$$\mathbf{L} := \begin{bmatrix} \ell_{11} & \ell_{12} & \cdots & \ell_{1K} \\ \ell_{21} & \ell_{22} & \cdots & \ell_{2K} \\ \vdots & & \ddots & \\ \ell_{K1} & \ell_{K2} & \cdots & \ell_{KK} \end{bmatrix}, \mathbf{p}(\mathbf{x}) := \begin{bmatrix} P(1|\mathbf{x}) \\ P(2|\mathbf{x}) \\ \vdots \\ P(K|\mathbf{x}) \end{bmatrix} \quad (2)$$

where the set $\mathcal{K} := \{1, \ldots, K\}$ indexes the classes. The loss ℓ_{km} is user-defined to encapsulate the loss associated with choosing class k for $\mathbf{x}$ when it is actually from m. $P(m|\mathbf{x})$ is the posterior of class m given the observation $\mathbf{x}$. We assume that all priors $P(m)$ are non-zero, and all classes are disjoint. For MGR, the last assumption means our model restricts each piece of music to be of only one class. This is obviously artificial [27], but provides a starting point. Furthermore, out of 467 works in MGR, only ten do not use this model [7].

2.1. Uniform loss

If we believe all misclassifications are equally bad, then we can define $\ell_{km} := 1-\delta_{k-m}$ where $\delta_p = 1$ for $p = 0$, and zero otherwise. This is also called the "zero-one loss" function, which makes $\mathbf{L} = \mathbf{1}\mathbf{1}^T - \mathbf{I}$, where $\mathbf{1}$ is a vector of K ones, and $\mathbf{I}$ is the identity matrix of appropriate size. In this case, (1) becomes the maximum a posteriori (MAP) classifier [10]:

$$\hat{k} = \arg\min_{k\in\mathcal{K}}\ \mathbf{e}_k^T(\mathbf{1}\mathbf{1}^T - \mathbf{I})\mathbf{p}(\mathbf{x}) = \arg\min_{k\in\mathcal{K}}\ \mathbf{e}_k^T[\mathbf{1} - \mathbf{p}(\mathbf{x})] = \arg\max_{k\in\mathcal{K}} P(k|\mathbf{x}). \quad (3)$$

MAP classification in MGR has been used in, e.g., [28–32]. If all classes are equally likely, the MAP classifier becomes the maximum likelihood (ML) classifier [10]:

$$\hat{k} = \arg\max_{k\in\mathcal{K}} P(\mathbf{x}|k). \quad (4)$$

ML classification in MGR has been used in [24, 33–43].

2.2. Non-uniform loss

Consider in the sample space a $c \in \mathcal{K}$ for which it is imperative the system has high precision (most observations it labels c are from c) and recall (it mislabels few observations from c). We are not concerned with other misclassifications. Thus, define the "persnickety-apathetic" loss function with loss l

$$\ell_{km}\ :=\ (1-\delta_{m-k})\Big[1 + (l-1)(\delta_{k-c} + \delta_{m-c})\Big]. \quad (5)$$

The loss matrix $\mathbf{L}$ has zeros on its diagonal, $K-1$ elements in row c of value l, and $K-1$ elements in column c of value l, and ones everywhere else. We now analyze the role of l.

The risk in (1) with this loss function becomes

$$R(k|\mathbf{x}) = \mathbf{e}_k^T\Big[(\mathbf{1}\mathbf{1}^T - \mathbf{I})\mathbf{p}(\mathbf{x}) + (l-1)P(c|\mathbf{x})\mathbf{1} - (l-1)\big[2P(c|\mathbf{x}) - 1\big]\mathbf{e}_c\Big] = 1 - P(k|\mathbf{x}) + (l-1)P(c|\mathbf{x}) - (l-1)[2P(c|\mathbf{x}) - 1]\delta_{k-c} \quad (6)$$

which makes the selection criterion become

$$\hat{k} = \arg\max_{k\in\mathcal{K}}\ P(k|\mathbf{x}) - (l-1)[P(\neg c|\mathbf{x}) - P(c|\mathbf{x})]\delta_{k-c} \quad (7)$$

with $P(\neg c|\mathbf{x}) := 1 - P(c|\mathbf{x})$. The classifier selects c when

$$P(c|\mathbf{x}) - (l-1)[P(\neg c|\mathbf{x}) - P(c|\mathbf{x})] > \max_{k\in\mathcal{K}\setminus c} P(k|\mathbf{x}). \quad (8)$$

If $l = 1$, this reduces to (3). For $l > 1/2$, if

$$P(c|\mathbf{x}) \le \frac{l-1}{2l-1} \quad (9)$$

then the classifier will never select c since the left hand side of (8) is then negative. The same thing occurs for $l < 1/2$ if

$$P(c|\mathbf{x}) \ge \frac{l-1}{2l-1}. \quad (10)$$

In the limit as $l \to \infty$, the classifier will not select c as long as $P(c|\mathbf{x}) \le 1/2$; and in the limit as $l \to -\infty$, the classifier will not select c as long as $P(c|\mathbf{x}) \ge 1/2$. We see by substitution that for $l = 1/2$, if $\max_{k\in\mathcal{K}\setminus c} P(k|\mathbf{x}) < 0.5$, then the classifier selects c.

2.3. Rejection

We can enable this system to reject classification when

$$\min_{k\in\mathcal{K}} R(k|\mathbf{x}) > R_{\max} \quad (11)$$

for a maximum risk $R_{\max}$. For the "persnickety-apathetic" loss function (5), the system rejects classification if

$$\min_{k\in\mathcal{K}} \Big[1 - P(k|\mathbf{x}) + (l-1)P(c|\mathbf{x}) - (l-1)[2P(c|\mathbf{x}) - 1]\delta_{k-c}\Big] > R_{\max} \quad (12)$$

If, for a given l, the class with smallest risk in (6) is c, and $lP(\neg c|\mathbf{x}) > R_{\max}$, then the classifier rejects classification. If the least-risk class in (6) is $k^* \neq c$, and $P(\neg k^*|\mathbf{x}) > R_{\max} - (l-1)P(c|\mathbf{x})$, then the classifier will reject classification. Hence, we see that in order to make the system capable of rejecting classification, we must define $R_{\max} < l$.

3. SIMULATIONS

We now test the impact of loss and rejection for an MGR system. As an example scenario, consider we have many hours of radio recordings, and wish to estimate the extent to which music that sounds "classical" appears. The amount of data we have is such that it prohibits manual listening and labeling. The success criteria of a useful MGR system include: high precision, i.e., that it is correct in most of what it identifies as "classical-sounding"; high recall, i.e., that very few "classical-sounding" excerpts are mislabeled; and that it produces a set of rejected classifications that is either mostly "classical-sounding" or mostly "not classical-sound", i.e., manual listening and labeling of this set is not prohibitive.

3.1. Method

We use the following experimental design. We train a classifier in six genre categories using features extracted from all disjoint 27.9 s ($5 \cdot 2^{17}$ samples at 22.05 kHz sampling rate) excerpts from the 729 audio files of the ISMIR2004 training dataset [44]. Table 1 lists the six categories, the number of files, and the number of excerpts from each. Then, using the validation set of ISMIR2004 [44], we estimate the best loss l and rejection threshold $R_{\max}$ with respect to minimizing

$$f(l,R) = [1 - \text{tpr}(l,R)]^2 + [\text{fpr}(l,R)]^2 + \beta\frac{X}{N}\left(\frac{1}{2} - \frac{1}{2}\cos[2\pi \text{pur}(l,R)]\right) \quad (13)$$

where $\text{tpr}(l,R)$ is the true positive rate, $\text{fpr}(l,R)$ is the false positive rate, $\beta X/N$ is the weighted ratio of rejections X to the total number of excerpts N, and $\text{pur}(l,R)$ is the proportion of the rejected classifications that are labeled "classical." For our problem, we want $\text{tpr}(l,R) \approx 1$, $\text{fpr}(l,R) \approx 0$, and, if $\beta > 0$, the number of rejected classifications either to be small, or to be large *and* consist mostly of excerpts that sound "non-classical" (in which case we ignore them), or "classical" (in which case we add them all to the positives). In our implementation, we compute $f(l,R)$ at several $\{(l,R)\}$, and find where it is the smallest for each classifier. Finally, we test the tuned system using features extracted from the 1000 excerpts of the GTZAN dataset [45]. Since GTZAN has recently been analyzed and shown to have several faults [46], we remove from the analysis 49 exact replicas, and relabel two "jazz" excerpts as "classical." The GTZAN column in Table 1 reflects these changes.

Few papers in MGR apply training on one dataset and testing on another, e.g., [40, 47]. The danger of this is that the concepts between two datasets may not be the same, even though names of some of their classes are identical. In our simulations, we thus assume the concepts of "Classical" in ISMIR2004 and GTZAN are similar enough that this will not be a serious problem for the validity of our experiments.

To create features for each 29.7 s excerpt, we compute "scattering coefficients" — recently proposed by Anden and Mallet [48] — using the implementation in [49]. The specific settings we define are: second-order decompositions, filter q-factor 16, and maximum scale 160. Since each excerpt produces 40 feature vectors of dimension 469, we classify an excerpt by selecting the class with the least *total* risk, i.e.,

$$\hat{k} = \arg\min_{k\in\mathcal{K}} \mathbf{e}_k^T \mathbf{L} \sum_{i=1}^{40} \mathbf{p}(\mathbf{x}_i) \quad (14)$$

where $\mathbf{p}(\mathbf{x}_i)$ are the posteriors for the ith feature vector. If the argument above is greater than $40R_{\max}$, then we make the system reject classification.

For each class, we define $P(\mathbf{x}|k) = \mathcal{N}(\mathbf{x}; \mu_k, \mathbf{C}_k)$, i.e., an observation of class k is distributed multivariate Gaussian with mean μ_k and covariance $\mathbf{C}_k$:

$$P(\mathbf{x}|k) \propto |\mathbf{C}_k|^{-1/2} e^{-\frac{1}{2}(\mathbf{x}-\mu_k)^T \mathbf{C}_k^{-1}(\mathbf{x}-\mu_k)}. \quad (15)$$

Label	*ISMIR2004* Training		*ISMIR2004* Validation		*GTZAN* Testing	
	files	excerpts	files	excerpts	files	excerpts
Classical	320	1864	318	2016	102	102
Country	-	-	-	-	100	100
Disco	-	-	-	-	94	94
Electronic	114	1150	115	1216	-	-
Hip hop	-	-	-	-	98	98
Jazz/Blues	26	206	26	190	185	185
Metal/Punk	45	334	45	357	92	92
Reggae	-	-	-	-	89	89
Rock/Pop	102	776	101	718	191	191
World	122	1256	122	1384	-	-
Total	729	5586	727	5881	951	951

Table 1. Summary of training, validation, and testing datasets

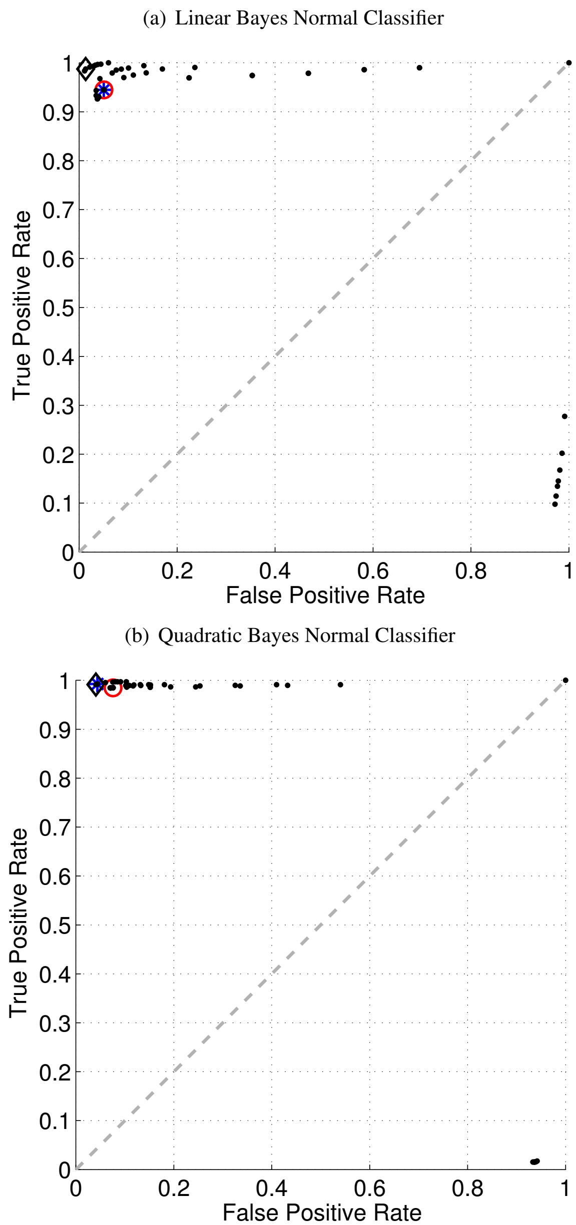

Fig. 1. Receiver operating characteristics (ROC) of our systems. Circle: system that does not use loss or rejection. Diamond: best system with respect to (13) for $\beta = 0$. Star: best system with respect to (13) for $\beta = 1$. The dots in the lower-right corner are all systems with $l < 0$.

To learn the parameters of each distribution, we use unbiased minimum mean-squared error estimators with the features we extract from the ISMIR2004 training data [50]. If we assume all classes are distributed with different means but the same covariance (estimated from the data of all classes) the classifier (1) is called "Linear Bayes Normal" (LBN); and if each class is distributed with different covariances (estimated from the data of each class), it is called "Quadratic Bayes Normal" (QBN) [10]. We define the priors of all classes to be the same.

Uniform loss, no rejection: $l = 1$, $R_{\max} = \infty$						
	Pred. (LBN)			Pred. (QBN)		
	+	−	r	+	−	r
+	1905	111	0	1985	31	0
−	195	3670	0	293	3572	0
Non-uniform loss, rejection, but no rejection purity ($\beta = 0$). LBN: $l = 2$, $R_{\max} = 0.5$; QBN: $l = 3$, $R_{\max} = 0.75$						
	Pred. (LBN)			Pred. (QBN)		
	+	−	r	+	−	r
+	1490	19	507	1971	17	28
−	36	2712	1117	140	3281	444
Non-uniform loss, rejection, and purity ($\beta = 1$). LBN: $l = 1$, $R_{\max} = 0.75$; QBN: $l = 2$, $R_{\max} = 0.5$						
	Pred. (LBN)			Pred. (QBN)		
	+	−	r	+	−	r
+	1905	111	0	1971	15	30
−	194	3670	1	140	3040	685

Table 2. Validation (ISMIR2004) confusion tables for LBN and QBN. Column "r" shows number of rejections in each.

3.2. Results of validation

From evaluation using the validation set of ISMIR2004, Fig. 1 shows the receiver operating characteristics (ROC) of LBN and QBN for several pairs of loss l and maximum risk $R_{\max}$; and Table 2 shows the confusion tables of three particular pairs. First, the open circles shows the performance of each system with uniform loss and no rejection, and Table 2(a) shows the confusion tables. We see LBN has about 100 fewer false positives than QBN, but about 80 more false negatives. The diamond in each ROC shows the best performance of each system with respect to (13) for $\beta = 0$ for non-uniform loss and rejection; and Table 2(b) shows the confusion tables. While the number of false positives and negatives for LBN decrease dramatically, it now finds 500 fewer true positives than before. QBN, however, now halves its number of false positives and negatives, and produces a set of rejections that is only 6% Classical. Finally, the star in each ROC shows the best performance of each system with respect to (13) for non-uniform loss and rejection, and rejection purity $\beta = 1$; Table 2(c) shows the confusion tables. Now we see that QBN produces 2 fewer false negatives, and has a set of rejections that is 4% Classical. The performance of LBN here essentially equals what it is for uniform loss and no rejection.

3.3. Results of classification

We now test each of these systems using the GTZAN dataset, the results of which are shown in Table 3. First, for the systems using uniform loss and no rejection we see in Table 3(a) that while LBN here has less than half as many false positives as QBN, it has a far higher number of false negatives. When these system consider non-uniform loss and rejection, but do not consider the purity of the rejections, Table 3(b) shows QBN and LBN both have a large decrease in the num-

	Pred. (LBN)			Pred. (QBN)		
Uniform loss, no rejection: $l=1$, $R_{\max}=\infty$						
	+	−	r	+	−	r
+	73	29	0	99	3	0
−	20	829	0	47	802	0
Non-uniform loss, rejection, but no purity ($\beta=0$). LBN: $l=2$, $R_{\max}=0.5$; for QBN: $l=3$, $R_{\max}=0.75$						
	+	−	r	+	−	r
+	31	0	71	94	0	8
−	2	617	230	19	760	70
Non-uniform loss and rejection, and rejection purity ($\beta=1$). LBN: $l=1$, $R_{\max}=0.75$; QBN: $l=2$, $R_{\max}=0.5$						
	+	−	r	+	−	r
+	73	29	0	94	0	8
−	20	828	1	19	711	119

Table 3. Test (GTZAN) confusion tables for LBN and QBN. Column "r" shows number of rejections in each.

ber of false positives, but LBN suffers significantly in its number of true positives. Finally, Table 3(c) shows, for our cases, considering the purity of the rejections produces little difference with the previous QBN system, or with LBN with uniform loss and no rejection. In summary, we see that with non-uniform risk and rejection we are able to make an MGR system, produce fewer false positives and false negatives for our given scenario then when it does not consider them.

4. CONCLUSION

That an artificial system misclassifies is no surprise; and to aim for one that does not misclassify certainly aims too high. When it comes to cultural content such as music genre, which escapes clear and definitive categories [27], and of which humans often disagree [11, 12, 23, 51, 52], misclassification is an inevitable part of an MGR system. One might see this as a selling point: "some people would be entertained by the predictions, especially when they were wrong" [53]. However, not all misclassifications are equal — some are worse (funnier?) than others — and little work in MGR explores this idea outside of evaluation. We have shown how such an idea can be naturally incorporated into an MGR system by using non-uniform loss and rejection and the minimum risk framework of Bayesian classification. We analyzed a particular form of the loss, and applied it within a scenario of detecting "classical-sounding" music. We could, of course, have trained the classifiers to discriminate between "classical" and "non-classical" — lumping together all excerpts in the ISMIR 2004 dataset that are not labeled "Classical" — but the point of this paper is not to solve that particular scenario. It is to investigate how loss and rejection can tune a multiclass MGR system to produce results that could be more useful in a scenario, regardless of whether or not the MGR system has any capacity to recognize the genre used by music [8, 9].

5. REFERENCES

[1] M. Gamon, "Linguistic correlates of style: authorship classification with deep linguistic analysis features," in *Proc. Int. Conf. Computational Linguistics*, Geneva, Switzerland, 2004.

[2] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "Short text classification in twitter to improve information filtering," in *Proc. ACM SIGIR*, 2010.

[3] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, 2008.

[4] T. Li and M. Ogihara, "Music artist style identification by semi-supervised learning from both lyrics and contents," in *Proc. ACM Multimedia*, 2004.

[5] F. Biadsy, *Automatic Dialect and Accent Recognition and its Application to Speech Recognition*, Ph.D. thesis, Columbia University, New York, NY, USA, 2011.

[6] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation," *IEEE Trans. Multimedia*, vol. 13, no. 2, pp. 303–319, Apr. 2011.

[7] B. L. Sturm, "A survey of evaluation in music genre recognition," in *Proc. Adaptive Multimedia Retrieval*, Copenhagen, Denmark, Oct. 2012.

[8] B. L. Sturm, "Two systems for automatic music genre recognition: What are they really recognizing?," in *Proc. ACM MIRUM Workshop*, Nara, Japan, Nov. 2012.

[9] B. L. Sturm, "Classification accuracy is not enough: On the evaluation of music genre recognition systems," *J. Intell. Info. Systems (accepted)*, 2013.

[10] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Academic Press, Elsevier, Amsterdam, The Netherlands, 4 edition, 2009.

[11] A. Craft, G. A. Wiggins, and T. Crawford, "How many beans make five? The consensus problem in music-genre classification and a new evaluation method for single-genre categorisation systems," in *Proc. ISMIR*, 2007.

[12] A. Craft, "The role of culture in the music genre classification task: human behaviour and its effect on methodology and evaluation," Tech. Rep., Queen Mary University of London, Nov. 2007.

[13] C. McKay, *Automatic Genre Classification of MIDI Recordings*, Ph.D. thesis, McGill University, Montréal, Canada, June 2004.

[14] C. McKay and I. Fujinaga, "Music genre classification: Is it worth pursuing and how can it be improved?," in *Proc. ISMIR*, Victoria, Canada, Oct. 2006.

[15] K. Seyerlehner, G. Widmer, and T. Pohle, "Fusing block-level features for music similarity estimation," in *DAFx*, 2010.

[16] K. Seyerlehner, G. Widmer, and P. Knees, "A comparison of human, automatic and collaborative music genre classification and user centric evaluation of genre classification systems," in *Proc. Adaptive Multimedia Retrieval*, 2010, pp. 118–131.

[17] V. Tsatsishvili, "Automatic subgenre classification of heavy metal music," M.S. thesis, University of Jyväskylä, Nov. 2011.

[18] K. West and P. Lamere, "A model-based approach to constructing music similarity functions," *EURASIP J. Applied Signal Process.*, vol. 1, no. 1, pp. 149–149, Jan. 2007.

[19] K. West, *Novel techniques for Audio Music Classification and Search*, Ph.D. thesis, University of East Anglia, 2008.

[20] MIREX, "Genre results," `http://www.music-ir.org/mirex/wiki/2009:MIREX2009_Results`, 2009.

[21] H.-Y. Lo, J.-C. Wang, H.-M. Wang, and S.-D. Lin, "Cost-sensitive multi-label learning for audio tag annotation and retrieval," *IEEE Trans. Multimedia*, vol. 13, no. 3, pp. 518–529, June 2011.

[22] Q. Fu and B.-H. Juang, "Automatic speech recognition based on weighted minimum classification error (w-mce) training method," in *IEEE Workshop Auto. Speech Recog. Understanding*, 2007, pp. 278–283.

[23] S. Lippens, J.P. Martens, and T. De Mulder, "A comparison of human and automatic musical genre classification," in *Proc. ICASSP*, May 2004, pp. 233–236.

[24] R. B. Dannenberg, B. Thom, and D. Watson, "A machine learning approach to musical style recognition," in *Proc. ICMC*, 1997, pp. 344–347.

[25] D. Pye, "Content-based methods for the management of digital music," in *Proc. ICASSP*, 2000.

[26] H. Harb and L. Chen, "A general audio classifier based on human perception motivated model," *Multimedia Tools and Applications*, vol. 34, pp. 375–395, 2007.

[27] J. Frow, *Genre*, Routledge, New York, NY, USA, 2005.

[28] C. Pérez-Sancho, J. M. Iñesta, and J. Rubio, "Style recognition through statistical event models," *J. New Music Research*, vol. 34, no. 4, pp. 331–340, 2005.

[29] C. N. Silla, A. Koerich, and C. Kaestner, "Automatic music genre classification using ensembles of classifiers," in *Proc. IEEE Int. Conf. Systems, Man, Cybernetics*, 2007, pp. 1687–1692.

[30] C. Perez, D. Rizo, and J. M. Iñesta, "Genre classification using chords and stochastic language models," *Connection Science*, vol. 21, pp. 145–159, June 2009.

[31] C. Pérez, *Stochastic language models for music information retrieval*, Ph.D. thesis, Universidad de Alicante, Spain, June 2009.

[32] Y. M. G. Costa, L. S. Oliveira, A. L. Koerich, and F. Gouyon, "Comparing textural features for music genre classification," in *Proc. IEEE World Cong. Comp. Intell.*, June 2012.

[33] H. Deshpande, R. Singh, and U. Nam, "Classification of music signals in the visual domain," in *Proc. DAFx*, Limerick, Ireland, Dec. 2001.

[34] D.-N. Jiang, L.-Lu, H.-J. Zhang, J.-H. Tao, and L.-H. Cai, "Music type classification by spectral contrast features," in *Proc. ICME*, 2002.

[35] T. Heittola, "Automatic classification of music signals," M.S. thesis, Tampere University of Tech., Feb. 2003.

[36] R. Basili, A. Serafini, and A. Stellato, "Classification of musical genre: A machine learning approach," in *Proc. ISMIR*, 2004.

[37] J. J. Burred and A. Lerch, "Hierarchical automatic audio signal classification," *J. Audio Engineering Society*, vol. 52, no. 7, pp. 724–739, 2004.

[38] S. Sundaram and S. Narayanan, "Experiments in automatic genre classification of full-length music tracks using audio activity rate," in *Proc. IEEE Workshop Multimedia Signal Process.*, 2007.

[39] H. Lukashevich, J. Abeßer, C. Dittmar, and H. Großmann, "From multi-labeling to multi-domain-labeling: A novel two-dimensional approach to music genre classification," in *ISMIR*, 2009.

[40] E. Guaus, *Audio content processing for automatic music genre classification: descriptors, databases, and classifiers*, Ph.D. thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2009.

[41] T. Li and A. Chan, "Genre classification and the invariance of MFCC features to key and tempo," in *Proc. Int. Conf. MultiMedia Modeling*, Taipei, China, Jan. 2011.

[42] S. Jothilakshmi and N. Kathiresan, "Automatic music genre classification for indian music," in *Proc. Int. Conf. Software Computer App.*, 2012.

[43] Y. Ni, M. McVicar, R. Santos, and T. De Bie, "Using hyper-genre training to explore genre information for automatic chord estimation," in *Proc. ISMIR*, 2012.

[44] ISMIR, "Genre results," `http://ismir2004.ismir.net/genre_contest/index.htm`, 2004.

[45] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, July 2002.

[46] B. L. Sturm, "An analysis of the GTZAN music genre dataset," in *Proc. ACM MIRUM Workshop*, Nara, Japan, Nov. 2012.

[47] J. Schlüter and C. Osendorfer, "Music similarity estimation with the mean-covariance restricted boltzmann machine," in *Proc. ICMLA*, 2011.

[48] J. Andén and S. Mallat, "Multiscale scattering for audio classification," in *Proc. ISMIR*, 2011, pp. 657–662.

[49] J. Andén and S. Mallat, "Scatterbox v. 1.02," http://www.cmap.polytechnique.fr/scattering/, June 2012.

[50] R. P. W. Duin, P. Juszczak, D. de Ridder, P. Paclik, E. Pekalska, and D. M. J. Tax, "PR-Tools4.1, a matlab toolbox for pattern recognition," Delft University of Technology, 2007, `http://prtools.org`.

[51] G. Mitri, A. L. Uitdenbogerd, and V. Ciesielski, "Automatic music classification problems," in *Proc. Autralasian Comp. Science Conf.*, 2004.

[52] G. A. Wiggins, "Semantic gap?? Schemantic schmap!! Methodological considerations in the scientific study of music," in *Proc. IEEE Int. Symp. Mulitmedia*, Dec. 2009, pp. 477–482.

[53] S. Golub, "Classifying recorded music," M.S. thesis, University of Edinburgh, Edinburgh, Scotland, U.K., 2000.