A Survey of Evaluation in Music Genre Recognition

Bob L. Sturm
Audio Analysis Lab, AD:MT, Aalborg University Copenhagen
A.C. Meyers Vænge 15, DK-2450 Copenhagen SV, Denmark
bst@create.aau.dk

Abstract. Much work is focused upon music genre recognition (MGR) from audio recordings, symbolic data, and other modalities. While reviews have been written of some of this work before, no survey has been made of the approaches to evaluating approaches to MGR. This paper compiles a bibliography of work in MGR, and analyzes three aspects of evaluation: experimental designs, datasets, and figures of merit.

1 Introduction

Despite much work [1–467], music genre recognition (MGR) remains a compelling problem to solve by a machine. In addition to many background chapters of master’s theses [39,79,113,132,153,154,188,193,239,361,367,371,418] and doctoral dissertations [9,141,146,280,284,290,320,341,342,381,427,447] at least five reviews are devoted specifically to MGR [23,85,123,241,373], and 19 other reviews discuss related aspects [24,25,51,71,84,100,101,152,181,198,224,233,270,282,315,398,423,441,442]. Many of these reviews compile the variety of feature extraction methods and classification algorithms that have been applied to MGR, and some compare system performance using specific figures of merit (FoM) on particular benchmark datasets. There have also been no fewer than 10 campaigns to formally evaluate and compare state-of-the-art algorithms for MGR [170,171,293–299,316]. However, the variety of approaches used for evaluating performance in MGR has yet to be surveyed. How does one measure the capacity of a system — living or not — to recognize and discriminate between abstract characteristics of the human phenomenon of music?

There currently exists at least eleven works [77,78,116,117,246,320,404,409,410,433,449] that address the difficult but clearly relevant question of how to evaluate the performance of MGR systems, not to mention how to properly create a dataset from which a machine is to learn an abstract and high-level concept such as genre [468,469]. A few works critically address evaluation in MGR. For instance, [77,78,409,410] argue for more realistic approaches than having a system apply a single label to music, and comparing against a “ground truth” — which itself can be quite wrong [404,408]. Furthermore, [77,78,246,449] argue for measuring performance in ways that take into account the natural ambiguity arising from genre.
In this paper, we take a different direction to answer the question we pose above. We review a significant portion of published research touching upon aspects of evaluation in MGR. We consider all work that is based upon recorded music, and/or symbolic representations of the music, e.g., MIDI, and/or other modalities, e.g., lyrics, album covers, user tags, movie scenes, etc. We do not, however, consider work addressing the more general problem of “tagging,” e.g., [470]. While we consider both “genre” and “style,” and make no attempt to differentiate them, we do not include “mood” or “emotion,” e.g., [471]. We are herein interested only in the ways systems for MGR are evaluated, be they algorithms, humans [79, 169, 201, 258, 261, 262, 278, 290, 366, 367, 370, 381, 383, 460], pigeons [347], sparrows [439, 440], koi [58], primates [278] or rats [317]. To facilitate this survey, we created a spreadsheet summarizing every relevant paper we found in terms of its experimental design, details of the datasets it uses, and the figures of merit it reports. This resource provides a simple means to delimit sets of references sharing particular aspects of evaluation.

Figure 1 shows how the number of the works we reference is distributed since the 1995 work of Matityaho and Furst [271] — before which we have only found the 1984 work of Porter and Neuringer [347]. Many papers allude to the 2002 article of Tzanetakis and Cook [426] as the beginning of research in automatic MGR. We find their manuscript (received Nov. 2001 and growing from [425]) is preceded by seventeen works [44, 53, 83, 91, 132, 148, 204, 270, 271, 346, 348, 350, 401, 443, 472, 473], and is contemporary with nine works [22, 70, 176, 193, 202, 218, 351, 385, 448]. The dataset created by Tzanetakis and Cook for [426], however, is the first “benchmark” MGR dataset to have been made publicly available, and as a result continues to be the most used public dataset for MGR.

In our analysis, we do not include [474–479] as they are written in Turkish, and [472] as it is written in German, and we can read neither. We could not obtain [473, 480, 481], and so do not include them in the analysis. Finally, we neither analyze nor cite seven published works because of plagiarism.

Fig. 1. Annual numbers of publications in this survey.

1 Upon request we can make available this spreadsheet and bibliography.
Fig. 2. Annual numbers of publications in this survey having an experimental component (top number), and which use any form of statistical testing for making comparisons (bottom number).

2 Evaluation Approaches in Music Genre Recognition

We now catalogue approaches to evaluation in MGR along three dimensions — experimental design, datasets, and figures of merit (FoM) — and present summary statistics of each. Experimental design is the method employed to answer a specific hypothesis, e.g., in the case of MGR, “System A recognizes ‘Blues’.” The dataset is simply the collection of data used in the experiment. A FoM quantifies the performance of a system in the experiment, e.g., accuracy. Figure 2 shows how the number of works having an experimental component is distributed over the years. Compared to Fig. 1, the remaining works are reviews, or primarily concerning evaluation.

2.1 Experimental Design


2 Numbers in parentheses are the number of works in the references.
Table 1. Ten experimental designs of MGR, and the percentage of references having an experimental component (435) in which they appear

<table>
<thead>
<tr>
<th>Design</th>
<th>Description</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classify</td>
<td>To answer the question, “How well does the system predict the genres used by music?” The system applies genre labels to music, which researcher then compares to a “ground truth”</td>
<td>91.3</td>
</tr>
<tr>
<td>Features</td>
<td>To answer the question, “At what is the system looking to identify the genres used by music?” The system ranks and/or selects features, which researcher then inspects</td>
<td>32.6</td>
</tr>
<tr>
<td>Generalize</td>
<td>To answer the question, “How well does the system identify genre in varied datasets?” Classify with two or more datasets having different genres, and/or various amounts of training data</td>
<td>15.9</td>
</tr>
<tr>
<td>Robust</td>
<td>To answer the question, “To what extent is the system invariant to aspects inconsequential for identifying genre?” The system classifies music that researcher modifies or transforms in ways that do not harm its genre identification by a human</td>
<td>6.9</td>
</tr>
<tr>
<td>Eyeball</td>
<td>To answer the question, “How well do the parameters make sense with respect to identifying genre?” The system derives parameters from music; researcher visually compares</td>
<td>6.7</td>
</tr>
<tr>
<td>Cluster</td>
<td>To answer the question, “How well does the system group together music using the same genres?” The system creates clusters of dataset, which researcher then inspects</td>
<td>6.7</td>
</tr>
<tr>
<td>Scale</td>
<td>To answer the question, “How well does the system identify music genre with varying numbers of genres?” Classify with varying numbers of genres</td>
<td>6.7</td>
</tr>
<tr>
<td>Retrieve</td>
<td>To answer the question, “How well does the system identify music using the same genres used by the query?” The system retrieves music similar to query, which researcher then inspects</td>
<td>4.4</td>
</tr>
<tr>
<td>Rules</td>
<td>To answer the question, “What are the decisions the system is making to identify genres?” The researcher inspects rules used by a system to identify genres</td>
<td>3.7</td>
</tr>
<tr>
<td>Compose</td>
<td>To answer the question, “What are the internal genre models of the system?” The system creates music in specific genres, which the researcher then inspects</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Almost all of the experimental work that applies Classify employ a single-label approach, but at least ten employ a multi-labeling approach [31, 255, 258, 280, 367, 369, 373, 415, 437]. For instance, McKay [280] looks at how genres at both root and leaf nodes are applied by his hierarchical approach to classification.

One extension of the Classify experimental design is Generalize [1, 2, 9, 13, 15, 16, 19, 30, 31, 38, 45, 55, 58, 62, 66, 75, 82, 94, 97, 128, 137, 146, 153, 161, 164, 199].
For instance, Porter and Neuringer [347] test whether pigeons trained to discriminate between music by J. S. Bach and Stravinsky are able to discriminate between music by composers contemporary with J. S. Bach (Buxtehude and Scarlatti) and Stravinsky (Carter and Piston). Another extension of Classify is Scale [9, 13, 14, 19, 45, 48, 49, 53, 62, 68, 83, 94, 132, 144, 199]. For instance, Chai and Vercoe [53] test their system on various class pairs from their dataset of three folk music genres, as well as on all three classes together.

The second most-used experimental design is Features [1, 7, 9, 16, 17, 26, 27, 33–35, 37, 43, 48, 49, 53, 68, 69, 72, 93, 95, 102, 103, 105, 109, 115, 122, 126, 127, 139, 141, 143, 146, 153, 157, 160, 179, 182, 184, 187, 189, 192, 196, 197, 199, 200, 211, 213, 219, 221, 226, 230, 232, 238, 240, 242, 244, 247, 249, 252, 277, 280, 281, 283–287, 289, 290, 303, 307, 309, 320, 327, 333, 335, 341, 345, 361, 362, 364, 365, 370, 371, 376, 377, 379, 385, 387, 390, 393, 396, 397, 401, 403, 408, 417, 420, 421, 425–427, 430, 432, 434, 436, 438, 443, 444, 447, 451, 454–456, 460, 462, 464, 465, 467]. We do not include in this experimental design work that performs feature selection without an interpretation of the results. For instance, Tzanetakis et al. [425] use Classify in comparing rhythmic features (statistics of an autocorrelation of wavelet decomposition) and timbral features (spectral centroid, rolloff, etc.). On the other hand, Yoon et al. [459] explore two different feature selection approaches using Classify, but do not discuss or list the selected features. Akin to Features is a fifth design, Rules, which appears in at least sixteen works [3, 5, 13, 14, 26, 42, 43, 70, 94, 98, 130, 303, 308, 340, 341, 344, 139]. For instance, Bickerstaffe and Makalic [43] look at a decision stump that discriminates “rock” and “classical” music. As another example, Abesser et al. [5] provide the details of a decision tree algorithmically built for discriminating between 13 genres.

Another experimental design is Cluster [22, 33, 69, 67, 72, 107, 126, 136, 189, 196, 218, 230, 235, 237, 242, 255, 261, 301, 302, 304, 315, 320, 326, 344, 350, 351, 360, 369, 374, 113, 417, 430, 438]. For instance, Rauber and Frühwirth [350] employ the self-organizing map method with features extracted from 230 music excerpts, and analyze the contents of the resulting groupings. We find that both Classify and Cluster are used in about 2.6% (12) of the experimental work [33, 72, 126, 189, 196, 242, 253, 261, 301, 320, 430, 438]. A seventh experimental design is Retrieve, which appears in at least 19 works [10, 46, 57, 61, 86, 118, 119, 121, 203, 232, 262, 320, 348, 384, 388, 446, 447, 466]. For instance, Kuo and Shan [203] incorporate style recognition into their music retrieval system.

An eighth experimental design is Eyeball, which appears in at least 29 works [26, 29, 44, 83, 91, 105, 146, 149, 155, 166, 173, 189, 218, 242, 259, 261, 288, 300, 302, 304, 310, 314, 320, 358, 360, 402, 403, 413, 463]. For instance, Dannenberg et al. [83] visually inspect class separability for a few pairs of features to explore the reason for a discrepancy in performance in identifying style between an expert approach and machine learning approach. Bigerelle and Iost [44] visually compare means of fractal dimensions computed from several musical excerpts of various genres.
A ninth experimental design is Robust. For instance, Porter and Neuringer [347] test whether pigeons that have been taught to discriminate between music by J. S. Bach and Hindemith demonstrate their ability regardless of excerpt content and loudness. Soltau et al. [401] investigate the variability of their system using Classify by using features computed from excerpts of several durations. Burred and Lerch [48] consider the effect of noise and filtering in feature extraction using Classify.

The final experimental design we consider is Compose, which appears in only three works [80, 82, 409]. For instance, Cruz and Vidal [80, 82] invert their music style identification system to compose music in the styles it has learned, which the authors then qualitatively evaluate. While Cruz and Vidal do not directly use this as a means to assess the extent to which their system has learned a style, [409] shows by a formal listening test that excerpts composed to be genre-representative by two high-accuracy MGR systems embody little in common with what is commonly held to be characteristic of those genres.

The bias that results from training and testing MGR systems using music data from the same artist and/or excerpted from the same album are well-documented, e.g., [117–119, 319]. Among the 435 works that include experimental work, we find that only 8.3% (36) explicitly mention the use of an artist or album filter [30, 57, 73, 76, 117, 119, 131, 142, 200, 235, 247, 267, 268, 290, 313, 320, 333, 347, 357, 381, 383, 384, 401, 409, 428, 439]. For instance, Porter and Neuringer [347] test whether pigeons that have been taught to discriminate between music by J. S. Bach and Hindemith demonstrate their ability regardless of excerpt content and loudness. Soltau et al. [401] investigate the variability of their system using Classify by using features computed from excerpts of several durations. Burred and Lerch [48] consider the effect of noise and filtering in feature extraction using Classify.

The final experimental design we consider is Compose, which appears in only three works [80, 82, 409]. For instance, Cruz and Vidal [80, 82] invert their music style identification system to compose music in the styles it has learned, which the authors then qualitatively evaluate. While Cruz and Vidal do not directly use this as a means to assess the extent to which their system has learned a style, [409] shows by a formal listening test that excerpts composed to be genre-representative by two high-accuracy MGR systems embody little in common with what is commonly held to be characteristic of those genres.

The bias that results from training and testing MGR systems using music data from the same artist and/or excerpted from the same album are well-documented, e.g., [117–119, 319]. Among the 435 works that include experimental work, we find that only 8.3% (36) explicitly mention the use of an artist or album filter [30, 57, 73, 76, 117, 119, 131, 142, 200, 235, 247, 267, 268, 290, 313, 320, 333, 347, 357, 381, 383, 384, 401, 409, 428, 439]. For instance, Porter and Neuringer [347] test whether pigeons that have been taught to discriminate between music by J. S. Bach and Hindemith demonstrate their ability regardless of excerpt content and loudness. Soltau et al. [401] investigate the variability of their system using Classify by using features computed from excerpts of several durations. Burred and Lerch [48] consider the effect of noise and filtering in feature extraction using Classify.

The final experimental design we consider is Compose, which appears in only three works [80, 82, 409]. For instance, Cruz and Vidal [80, 82] invert their music style identification system to compose music in the styles it has learned, which the authors then qualitatively evaluate. While Cruz and Vidal do not directly use this as a means to assess the extent to which their system has learned a style, [409] shows by a formal listening test that excerpts composed to be genre-representative by two high-accuracy MGR systems embody little in common with what is commonly held to be characteristic of those genres.

The bias that results from training and testing MGR systems using music data from the same artist and/or excerpted from the same album are well-documented, e.g., [117–119, 319]. Among the 435 works that include experimental work, we find that only 8.3% (36) explicitly mention the use of an artist or album filter [30, 57, 73, 76, 117, 119, 131, 142, 200, 235, 247, 267, 268, 290, 313, 320, 333, 347, 357, 381, 383, 384, 401, 409, 428, 439]. For instance, Porter and Neuringer [347] test whether pigeons that have been taught to discriminate between music by J. S. Bach and Hindemith demonstrate their ability regardless of excerpt content and loudness. Soltau et al. [401] investigate the variability of their system using Classify by using features computed from excerpts of several durations. Burred and Lerch [48] consider the effect of noise and filtering in feature extraction using Classify.

The final experimental design we consider is Compose, which appears in only three works [80, 82, 409]. For instance, Cruz and Vidal [80, 82] invert their music style identification system to compose music in the styles it has learned, which the authors then qualitatively evaluate. While Cruz and Vidal do not directly use this as a means to assess the extent to which their system has learned a style, [409] shows by a formal listening test that excerpts composed to be genre-representative by two high-accuracy MGR systems embody little in common with what is commonly held to be characteristic of those genres.

The bias that results from training and testing MGR systems using music data from the same artist and/or excerpted from the same album are well-documented, e.g., [117–119, 319]. Among the 435 works that include experimental work, we find that only 8.3% (36) explicitly mention the use of an artist or album filter [30, 57, 73, 76, 117, 119, 131, 142, 200, 235, 247, 267, 268, 290, 313, 320, 333, 347, 357, 381, 383, 384, 401, 409, 428, 439]. For instance, Porter and Neuringer [347] test whether pigeons that have been taught to discriminate between music by J. S. Bach and Hindemith demonstrate their ability regardless of excerpt content and loudness. Soltau et al. [401] investigate the variability of their system using Classify by using features computed from excerpts of several durations. Burred and Lerch [48] consider the effect of noise and filtering in feature extraction using Classify.

The final experimental design we consider is Compose, which appears in only three works [80, 82, 409]. For instance, Cruz and Vidal [80, 82] invert their music style identification system to compose music in the styles it has learned, which the authors then qualitatively evaluate. While Cruz and Vidal do not directly use this as a means to assess the extent to which their system has learned a style, [409] shows by a formal listening test that excerpts composed to be genre-representative by two high-accuracy MGR systems embody little in common with what is commonly held to be characteristic of those genres.

We find that at least twelve works use human evaluation in the analysis of the experiment [22, 36, 80, 82, 83, 260, 320, 347, 409, 410, 434, 447]. For instance, Dannenberg et al. [83] discuss the performance of their system in a live-performance context. Cruz and Vidal [80] rate the quality of the melodies composed by their style recognition system. And Pampalk [320] uses a formal listening test to show genre labels are strongly correlated with perceptual similarity.

Figure 1 shows the number of experimental works employing formal statistics over each year. Only 16.5% (72) of the experimental work we survey contains formal statistical testing [9, 15, 25, 27, 37, 44, 46, 58, 68, 73, 79, 114, 117, 121, 122, 124, 131, 132, 135, 146, 169, 174, 201, 221, 252, 258, 272, 273, 275, 277, 278, 283, 285, 289, 291, 295, 299, 304, 308, 310, 317, 320, 333, 337, 341, 349, 357, 377, 384, 395, 397, 409, 422, 434, 439, 440, 444, 447, 457, 466]. For instance, Flexer [116] provides excellent argumentation for the need for statistical testing in music information research, and provides a demonstration of its use in comparing the performance between two MGR systems. We find half of the work using living subjects (11 of 22) employ formal statistical tests [58, 79, 131, 169, 201, 258, 278, 290, 317, 439, 440]. For instance, Chase [58] uses a one-tailed paired t-test of percentages of non-responses of koi fish to test the null hypothesis that the koi are unable to discriminate between music that uses Blues or Classical genres.

Nearly half (213) of the experimental work we survey employs only one experimental design from Table 1. For instance, in several formal MGR chal-
lenges 170 [171,293,299], performance is evaluated only by Classify. We find about 32% (142) of the work we survey employ two experimental designs. For instance, Golub [132] uses Classify to test his MGR system for a three-genre problem, and then uses Scale to observe how its behavior changes when he augments the dataset with four other genres. More than 18% (80) employ more than two experimental designs. For instance, the only two experimental designs not used by Pampalk [220] are Rules and Compose.

2.2 Datasets

We find that of the works we survey having experimental components (435) over 58% (253) use private data [1–5, 7–11, 13–16, 18, 19, 22, 26, 28–31, 34, 40, 43–49, 53, 56–58, 62–70, 72, 73, 79–83, 87–89, 91–99, 104, 105, 109–111, 118–120, 125–128, 130–138, 142–146, 148–151, 154, 156–160, 163–164, 166–169, 172–175, 176–178, 180–184, 188, 190, 191, 196, 199, 205, 207, 209, 211, 217, 221, 225, 226, 228, 229, 231, 232, 242, 243, 245–253, 255, 257–261, 266, 271–275, 277, 281, 287–292, 300–305, 308, 312, 313, 317–320, 327–331, 334–342, 344, 346–348, 350, 358, 360, 361, 363–365, 372, 374, 377, 379–382, 384, 387, 389, 390, 401, 408, 419–422, 425–430, 437–440, 443–448, 450, 451, 453–456, 458, 463, 467]. Of those works that use private data, we find over 75% (191) exclusively use private data. Some work provides a detailed description of the composition of the data such that one can recreate it. For instance, Tsatsishvili [418] lists the 210 names of the albums, artists, and songs in his dataset. Schedl et al. [374] provide a URL for obtaining the list of the artists in their dataset, but the resource no longer exists. Mace et al. [258] also provide a list, but since they only list the song and artist name uncertainty arises, e.g., which recording of “The Unanswered Question” by Ives do they use? It is impossible to recreate the dataset used in [48,49] since they only state that they assemble 850 audio examples in 17 different genres. We find that about 51% (224) of the works we survey having experimental components use datasets that are publicly available. Of these, over 79% (177) only use public data.

Table 2. Datasets used in MGR, the type of data they contain, the references in which they are used, and the percentage of experimental work (435) that use them. All datasets listed after Private are public.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Description</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Private</td>
<td>Constructed for research but not made available; used in: see text</td>
<td>58.2</td>
</tr>
<tr>
<td>GTZAN</td>
<td>Audio; <a href="http://marsyas.info/download/data_sets">http://marsyas.info/download/data_sets</a> used in: see text</td>
<td>23.0</td>
</tr>
<tr>
<td>ISMIR2004</td>
<td>Audio; <a href="http://ismir2004.ismir.net/genre_contest">http://ismir2004.ismir.net/genre_contest</a> used in: see text</td>
<td>17.4</td>
</tr>
<tr>
<td>Latin 394</td>
<td>Features; <a href="http://www.pggia.pucpr.br/~silla/lmd/">http://www.pggia.pucpr.br/~silla/lmd/</a> used in: see text</td>
<td>5.1</td>
</tr>
<tr>
<td>Ballroom 165</td>
<td>Audio; <a href="http://mtg.upf.edu/ismir2004/contest/tempoContest/">http://mtg.upf.edu/ismir2004/contest/tempoContest/</a> used in: see text</td>
<td>3.4</td>
</tr>
<tr>
<td>Homburg 165</td>
<td>Audio; <a href="http://www-si.cs.uni-dortmund.de/audio.html">http://www-si.cs.uni-dortmund.de/audio.html</a> used in: see text</td>
<td>3.4</td>
</tr>
<tr>
<td>Bodhidharma</td>
<td>Symbolic; used in: <a href="http://jmir.sourceforge.net/Codaich.html">http://jmir.sourceforge.net/Codaich.html</a></td>
<td>2.5</td>
</tr>
<tr>
<td>1517-artists</td>
<td>Audio; <a href="http://www.seyerlehner.info">http://www.seyerlehner.info</a> used in: see text</td>
<td>1.1</td>
</tr>
<tr>
<td>RWC 483</td>
<td>Audio; <a href="http://staff.aist.go.jp/m.goto/RWC-MDB/">http://staff.aist.go.jp/m.goto/RWC-MDB/</a> used in: see text</td>
<td>0.9</td>
</tr>
<tr>
<td>SOMeJB 482</td>
<td>Features; <a href="http://www">http://www</a> ifs.tuwien.ac.at/~andi/somejb/ used in: see text</td>
<td>0.9</td>
</tr>
<tr>
<td>SLAC</td>
<td>Audio &amp; symbols; <a href="http://jmir.sourceforge.net/Codaich.html">http://jmir.sourceforge.net/Codaich.html</a> used in: see text</td>
<td>0.9</td>
</tr>
<tr>
<td>SALAMI 400</td>
<td>Features; <a href="http://ddmal.music.mcgill.ca/research/salami">http://ddmal.music.mcgill.ca/research/salami</a> used in: see text</td>
<td>0.7</td>
</tr>
<tr>
<td>Unique 481</td>
<td>Features; <a href="http://www.seyerlehner.info">http://www.seyerlehner.info</a> used in: see text</td>
<td>0.7</td>
</tr>
<tr>
<td>Million Song</td>
<td>Features; <a href="http://labrosa.ee.columbia.edu/millionsong/">http://labrosa.ee.columbia.edu/millionsong/</a> used in: see text</td>
<td>0.7</td>
</tr>
<tr>
<td>ISMIS2011</td>
<td>Features; <a href="http://tunedit.org/challenge/music-retrieval">http://tunedit.org/challenge/music-retrieval</a> used in: see text</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Over 79% (344) of the experimental work we survey approaches MGR using audio data or features of audio.
Figure 3 shows the number of experiments in the evaluative work we survey using datasets with specific numbers of labels. We can clearly see the influence of the GTZAN (10 genre labels) and ISMIR2004 (6 genre labels) datasets. We find 16 works using datasets having 25 or more labels [25, 30, 31, 40, 106, 107, 153, 199, 228, 232, 280, 309, 333, 376, 434, 437], and only two using datasets having more than 100 labels [40, 437]. Over 72% (316) of the papers with an experimental component uses only a single dataset, at least 20% (90) use two datasets, and 6.2% (27) use more than two datasets. Three references provide no details about the dataset used [54, 145, 331].

We find a majority of the works with experimental components involves datasets that consist primarily of Western music. For instance, the label “classical” is part of GTZAN, ISMIR2004, and RWC, and exists in the private datasets used in [22, 125, 143, 146, 173, 209, 225, 246, 266, 313, 341, 342, 344]. The label “blues” is in GTZAN, ISMIR2004, and Homburg, and exists in the private datasets used by [19, 22, 57, 125, 144, 146, 173, 209, 225, 299, 369, 393]. And the label “jazz” is in GTZAN, ISMIR2004, Homburg, and RWC, and exists in the private datasets used by [19, 22, 57, 125, 144, 146, 173, 209, 225, 299, 369, 393]. However, we find that only about 10% (48) of the private
datasets used include music from around the world, such as Asia, Africa, and South America [3–5,11,19,22,57,96–98,125,133–135,137,163,164,173,179,202, 205, 209, 225, 229, 242, 243, 247–250, 255, 257, 266, 303, 308, 312, 313, 342, 344, 358, 360, 363, 385, 389, 430, 462, 463]. Finally, we find only 5% (22) of the work with experimental components perform human validation of the “ground truth” labels in the public and/or private datasets used [8, 9, 34, 45, 79, 246, 247, 287, 289, 301, 346, 366, 367, 370, 383, 387, 394, 401, 408, 434]. For instance, Soltau et al. [401] validate the labels in their private four-class dataset with a human listening experiment.

2.3 Figures of Merit (FoMs)

Table 3 defines several FoMs we find in the work we surveyed. The FoMs most often reported in the work we survey here are those that accompany the Classify experimental design: Mean accuracy, Recall, Precision, F-measure, Receiver Operating Characteristic (ROC), and the Contingency table. We find Mean accuracy in over 82% (385) of the references. For instance, Fu et al. [123] compare the reported mean accuracies of 16 MGR algorithms using Classify in GTZAN. This computation can also involve taking into consideration “partial credit” for labelings in the correct hierarchical branch, e.g., [293,294,296]. When it appears, Mean accuracy is accompanied by a standard deviation (SD), or standard error of the mean (SEM), about 25% (96) of the time. For instance, [116] uses these statistics to test the null hypothesis that the Mean accuracy of two MGR systems are not significantly different.

We find Recall in over 25% (119) of the references. For instance, this FoM appears in the MIREX evaluations of MGR algorithms [295,297,299]. When it appears, Recall is accompanied by the standard deviation or standard error of the mean in about 10% (12) of the references. Precision appears in over 10% (47) of the references. Together, Mean accuracy, Recall and Precision appear in over 6% (31) of the work we survey. The F-measure can be computed in “Micro form” and “Macro form” [437], but we make no distinction here. This FoM appears in at least 17 works. For instance, Burred and Peeters [50] cite the F-measure of their MGR system, as well as its Mean accuracy, Recall, and Precision. We find the ROC in only 7 references [105,121,245,349,432,440,466]. For instance, Watanabe and Sato [140] plot the ROC of their sparrow trained to discriminate Baroque and Modern music.

We find a Contingency table reported in over 32% (150) of the work we survey. For instance, Soltau et al. [401] show their MGR system often confuses the music in their private dataset having the labels “rock” or “pop,” and rarely confuses music labeled “classic” with music labeled “techno” or “rock.” Of those works that present contingency tables, only 52% (78) of them are accompanied by some musical reflection of the results. For instance, in the analysis of their Contingency table, Dixon et al. [93] reason that the high number of confusions produced between three of eight classes come from the fact that they are indistinguishable using meter- and tempo-sensitive features employed in their system. When they expand their feature set, the new Contingency table confirms this hypothesis.
Table 3. Figures of merit (FoMs) of MGR, their description, and the percentage of work (467) that cite them

<table>
<thead>
<tr>
<th>FoM</th>
<th>Description</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean accuracy</td>
<td>Proportion of the number of correct trials to the total number of trials of the system</td>
<td>82</td>
</tr>
<tr>
<td>Contingency table</td>
<td>Counts of labeling outcomes of the system for each labeled input</td>
<td>32</td>
</tr>
<tr>
<td>Recall</td>
<td>Proportion of the number of correct trials of the system to the total number of a specific input label</td>
<td>25</td>
</tr>
<tr>
<td>Confusions</td>
<td>Discussion of confusions of the system in general or with specifics</td>
<td>24</td>
</tr>
<tr>
<td>Precision</td>
<td>Proportion of the number of correct trials of the system to the total number of a specific output label</td>
<td>10</td>
</tr>
<tr>
<td>F-measure</td>
<td>Twice the product of Recall and Precision divided by their sum</td>
<td>4</td>
</tr>
<tr>
<td>Composition</td>
<td>Observations of the composition of clusters created by the system, and distances within and between</td>
<td>4</td>
</tr>
<tr>
<td>Precision@k</td>
<td>Proportion of the number of correct items of a specific label in the k items retrieved by the system</td>
<td>3</td>
</tr>
<tr>
<td>ROC</td>
<td>Precision vs. Recall for several systems, parameters, etc.</td>
<td>1</td>
</tr>
</tbody>
</table>

General discussions about observed confusions without reference to a Contingency table are reported in over 8% (39) of the references. For instance, Matityaho and Furst [271] note that their MGR system trained to discriminate between music labeled “classic” and “pop” classifies as “classic” a signal of complete silence and a “complex tone,” and as “pop” a signal of white noise. Using Eyeball, Bigerelle and Iost [44] argue that “Music classification becomes very logical [by comparing the fractal dimension]. ... Progressive music has the same fractal dimension as the electronic one: we could explain this fact by the abundance of synthesizers used in progressive music.” Only 15 works mention confusions in detail, e.g., a specific piece of misclassified music [3, 68, 95, 100, 210, 228, 301, 342, 366, 367, 370, 407–410, 412]. For instance, [410] notices that one MGR system persistently misclassifies as Hip hop “Kung Fu Fighting” by Carl Douglas, and as Classical “Why?” by Bronski Beat.

We find over 44% (175) of the 397 works employing Classify report only one FoM and over 53% (214) report more than one FoM. Only 21 present four or more FoM [6, 34, 50, 53, 93, 156, 188, 200, 208, 238, 240, 333, 353, 355, 366, 367, 368, 410, 412, 418, 433]. For instance, Lidy [239] reports mean accuracies, recalls, precisions, F-measures, and contingency tables of the systems he tests.

The FoM most often reported in the case of the Retrieve experimental design is Precision@k. This FoM is reported in 12 of the 19 works using Retrieve [10, 57, 61, 56, 203, 222, 262, 320, 384, 446, 447, 466]; and [388] reports “normal-
ized precision” and “normalized recall,” which takes into account the ranking of retrieved elements. Of the references using Retrieval, the ROC is reported in [121, 166]. For instance, Fu et al. [121] plot the ROC of four systems to show their retrieval approach provides statistically significant improvement.

The FoMs most often reported in the case of Cluster experimental design are based on observations of the cluster compositions. The contents of clusters are analyzed in over 62% (18) of the Clustering experiments [33, 107, 136, 196, 236, 237, 253, 301, 302, 304, 318, 320, 334, 350, 351, 365, 415, 430]. For instance, Rauber and Frühwirth [350] show that one cluster created by the self-organizing map method consists mainly of music labeled “classical.” Comparisons of cluster distances, e.g., that within classes to that between classes, are reported in five works [22, 72, 302, 318, 334]. For instance, Aucouturier and Pachet [22] compare average distances between neighbors of the same class to those between neighbors of different classes. Visualizations of the clusters, e.g., using self-organizing maps, are presented in seven works [189, 218, 241, 242, 350, 351, 473]. Both [233, 417] report the “purity” of a collection of clusters, which measures the mean class homogeneity of the clusters.

Human-weighted ratings of classification and/or clustering results are reported in at least six works [22, 154, 203, 246, 366, 370]. Other FoMs include, “staying time” [278, 439] (measuring the time during which the subject stayed in the presence of musical stimuli for particular classes), “stability measure” [161, 162] (essentially inter-intra class distance), “Hamming loss” [367–369] (describing instance-label pair misclassifications in a multilabel scenario), and “persistent misclassifications” [65, 342, 409, 110] (noting instances that a system always mislabels).

3 Conclusion

While genre is an inevitable condition of human communication in general [469], a way to automatically identify it in music remains elusive. In this paper, we have attempted to present an exhaustive survey of evaluation in MGR, and to organize it along three dimensions: experimental design, datasets, and figures of merit. By the sheer size of this task, it is certain that we have missed some relevant work, misunderstood aspects of evaluation in some of the works we cite, and committed errors in the bibliography. We will thus continue to maintain this bibliography, and expand it when new work is published.

Acknowledgments

For Pepi [58]. Thank you to Carla Sturm for her bibliographic entry prowess. This work 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 project CoSound, case no. 11-115328.
Bibliographic Notes


References

95. Dor, O., Reich, Y.: An evaluation of musical score characteristics for automatic classification of composers. Computer Music J. 35(3) (Fall 2011) 86–97


24


