Musical genre classification: Is it worth pursuing and how can it be improved?

Cory McKay and Ichiro Fujinaga
Music Technology, Schulich School of Music, McGill University
Montreal, Quebec, Canada
cory.mckay@mail.mcgill.ca, ich@music.mcgill.ca

Abstract

Research in automatic genre classification has been producing increasingly small performance gains in recent years, with the result that some have suggested that such research should be abandoned in favor of more general similarity research. It has been further argued that genre classification is of limited utility as a goal in itself because of the ambiguities and subjectivity inherent to genre.

This paper presents a number of counterarguments that emphasize the importance of continuing research in automatic genre classification. Specific strategies for overcoming current performance limitations are discussed, and a brief review of background research in musicology and psychology relating to genre is presented. Insights from these highly relevant fields are generally absent from discourse within the MIR community, and it is hoped that this will help to encourage a more multi-disciplinary approach to automatic genre classification in the future.

Keywords: Genre, classification, music, improvements.

1. Introduction

Automatic genre and style classification have been popular topics in MIR research in the past. The ground-breaking work of Dannenberg, Thom and Watson [1] and of Tzanetakis and Cook [2] is particularly well-known, and more recent work includes the MIREX 2005 winning work of both Bergstra et al. [3] and of McKay and Fujinaga [4]. The many other exciting approaches applied to these problems are too numerous to include here, but a survey of the ISMIR proceedings from the past several years and their references will reveal the many different approaches used.

Despite this popularity, the view that further research in automatic genre classification will offer little of value has increasingly been expressed in informal discussions among researchers and on the MIREX and Music-IR mailing lists. It has been suggested that it would be more profitable to pursue research on more general music similarity instead, such as playlist generation, similarity-based browsing interfaces and recommendation systems.

This paper seeks to consider the arguments for and against further research in automatic genre classification, and a variety of major changes are proposed regarding how genre classification should be approached. This paper additionally includes a brief review of musicological and psychological research on genre and human classification.

Before proceeding, it is useful to briefly discuss the difference between "style" and "genre," as some disagreement has been expressed regarding these terms. Although there are no universally accepted definitions, Franco Fabbri has usefully defined genre as "a kind of music, as it is acknowledged by a community for any reason or purpose or criteria, i.e., a set of musical events whose course is governed by rules (of any kind) accepted by a community" and style as "a recurring arrangement of features in musical events which is typical of an individual (composer, performer), a group of musicians, a genre, a place, a period of time" [5]. It might be added that style is primarily related to individuals or groups of people involved in music production and that genre is related to more general groups of music and their audiences. Genre can thus be broader and more nebulous than style from a content-based perspective, and may be more strongly characterized by cultural features. The differences between genre and style are discussed in more detail in many of the references referred to in Section 2.

Despite these differences, many systems designed for genre classification could easily be applied to style classification, and vice versa, so strictly differentiating between the two is not necessarily an issue of primary importance from an MIR perspective. Although this paper targets issues relating to genre specifically, many of the points made here apply to style classification as well.

2. Insights From Other Disciplines

Genre is an area of inquiry that has been given significant attention in a variety of academic fields, with a particular emphasis found in literary (e.g., [6]) and film (e.g., [7]) studies. This research attempts to address issues such as how genres are created, how they can be defined, how they are perceived and identified, how they are disseminated, how they change, how they are interrelated and how we make use of them.

A number of musicologists have adapted this work to music and expanded upon it. Much of their research has emphasized the role of cultural factors in genre. Fabbri, for example, suggests that musical genres can be characterized using the following types of rules, of which only the first is related strictly to musical content [8]:

- *Formal and technical:* Content-based practices.
- *Semiotic:* Abstract concepts that are communicated (e.g., emotions or political messages).
- *Behavior:* How composers, performers and audiences appear and behave.
- *Social and ideological:* The links between genres and demographics such as age, race, sex and political viewpoints.
- *Economical and juridical:* The laws and economic systems supporting a genre, such as record contracts or performance locales (e.g., cafés or auditoriums).

Fabbri has also contributed many other ideas on diverse aspects of musical genre [5, 8, 9]. Frith offers insights on how musical genres are formed and what they mean [10]. Toynbee discusses how genres inform musicians and how they are influenced by identification with different communities and by the music industry [11]. Brackett has provided useful ideas on how genres can be characterized and on how genres are constructed, how they can be grouped and how they change [12, 13]. Important research has also been published on how genres can be organized from a technological perspective [14, 15].

A better understanding of the psychological processes involved in human music classification can also prove useful to MIR genre researchers. Not only does it help one model human classification behavior, but it can also be useful in designing interfaces that better meet human needs.

Early psychological models assumed that humans form categories by specifying necessary and sufficient conditions for each category. The diverse work of Eleanor Rosch has been influential in experimentally demonstrating the shortcomings of this approach and in promoting alternative models where categories are hierarchically organized and are defined by prototypical exemplars. Lakoff has written an excellent overview of these developments in the psychology of classification [16], and many papers have since been published proposing variations of exemplar theory.

A number of psychologists have applied these ideas to music. Deliege, for example, has suggested that humans abstract useful features contained in music into "cues," and use these to segment music, judge musical similarity and form "imprints" that help us to perceive musical structure, evaluate similarity and perform classifications [17]. There is also an extensive literature on the perception and cognition of musical similarity that is too extensive to cite here, but is certainly relevant to MIR-related similarity research.

Important research has also been performed by popular and ethnomusicologists on developing features that can be applicable to categorizing a diverse range of musics. Although there is insufficient room to present details here, a review has been previously written [18].

3. Problematic Aspects of Genre

The acquisition of reliable ground truth is a key requirement of training effective genre classifiers. It has been suggested that only limited agreement can be achieved among human annotators when classifying music by genre, and that such limits impose an unavoidable ceiling on automatic genre classification performance. Not only can individuals differ on how they classify a given recording, but they can also differ in terms of the pool of genre labels from which they choose. Very few genres have clear definitions, and what information is available is often ambiguous and inconsistent from source to source. There is often significant overlap between genres, and individual recordings can belong to multiple genres to varying degrees. There are often complex relationships between genres, and some genres are broad while others are narrow. Furthermore, genres often encapsulate multiple discrete clusters (e.g., Baroque music could include both a Monteverdi opera and a Scarlatti harpsichord sonata).

Only a small amount of experimental psychological research has been performed on human genre classification. An often-cited preliminary study found that a group of undergraduate students made classifications agreeing with those of record companies only 72% of the time when classifying among ten genres [19]. Listeners in these experiments were only exposed to 300 ms of audio per recording, however, and higher agreement rates could quite possibly have been attained had longer listening intervals been used. Another study involving longer thirty second listening intervals found inter-participant genre agreement rates of only 76% [20]. However, one of the six categories used in this study was "Other," an ambiguity that could lead to substantial disagreement due to degree of membership and category coarseness rather than entirely different classifications. So, although these two studies do provide useful insights, there is clearly a need for more experimental evidence before definitive conclusions can be drawn regarding upper bounds on software performance due to limits in human genre classification.

In any case, although some good sources of ground truth do exist, such as the AllMusic Guide [21], they are few in number and often contain too much or too little information. Furthermore, genre classifications tend to be by artist or album rather than by individual recording. Those sources that do provide classifications of individual recordings—such as Gracenote CDDB [22] or the metadata found in MP3 ID3 tags—tend to have unreliable annotations. There is also usually little or no documentation on how classifications were performed, and it is often doubtful whether serious effort was put into thoughtful, methodical and consistent annotations.

The expertise and time needed to manually clasify recordings pose serious obstacles to the production of quality ground truth. This is particularly true when large datasets are needed to avoid overtraining and to effectively learn models that incorporate the ambiguities and inconsistencies that one finds with genre.

To further complicate matters, not only are new genres introduced regularly, but the understanding of existing genres can also change with time, which can necessitate retraining of systems and re-annotation of ground truth. The need for a large training set also has implications in terms of machine learning. Powerful learning algorithms such as support vector machines or AdaBoost are needed to model highly complex genre spaces effectively, but many of the most powerful algorithms do not scale well.

In terms of actual software performance to date, no system has yet achieved sufficiently high success rates to make it usable in realistic situations. For example, the highest success rates in the MIREX 2005 audio genre classification contest were 75% and 87% when classifying among ten and six genres, respectively, and the highest rates were 46% and 86% in the symbolic classification contest for 38 and 9 genres, respectively [23]. Furthermore, it has been observed that recent systems that assess audio similarity in general using timbre-based features have failed to achieve major performance gains over earlier systems [24]. It is clear that fundamentally new approaches are needed if automatic genre classification is to become practically viable.

4. Arguments in Favor of Using Genre

Before proposing ways of overcoming the serious issues described in the previous section, it is appropriate to first emphasize the usefulness of automatic genre classification. It has been suggested in a number of discussions that genre is a hopelessly ambiguous and inconsistent way to organize and explore music, and that users' needs would be better addressed by abandoning it in favor of more general similarity-based approaches. Those adhering to this perspective generally hold that genre is only a subset of broader similarity research and has only been worth pursuing as an initial limited stage of research where features and learning algorithms can be developed, compared and refined.

Although it is true that genre is in some ways a subset of the more general similarity problem, genre involves a special emphasis on culturally predetermined classes that makes it worthy of separate attention. Even similarity measurements that involve cultural features such as playlist co-occurrence tend to be based on individual preferences rather than genre's more formal sociocultural agreements (see Section 2). In essence, the query "find me something like this (relatively small) set of recordings" is intrinsically different from "find me something in this generally understood genre category," which could encompass a potentially huge set of recordings and which is based on culturally determined categories rather than more content-oriented or individually defined similarity.

This highlights the importance of cultural features and the ever increasing variety and scale of metadata that can be mined from the web. Relatively little attention has been given to these types of features to date, yet they could well hold the potential to surpass current limitations on classification performance. The potential of such features will continue to increase as more metadata becomes available on the web and in recordings themselves.

The question remains whether genre classification is useful to end users, or simply an awkward and obsolete labeling system. Although browsing and searching by genre is certainly not perfect—and alternatives are always worth researching—end users are nonetheless already accustomed to browsing both physical and on-line music collections by genre, and this approach is proven to be at least reasonably effective. A recent survey, for example, found that end users are more likely to browse and search by genre than by recommendation, artist similarity or music similarity, although these alternatives were each popular as well [25]. Resources such as the AllMusic Guide [21], which use labeled fields such as genre, mood and style are commonly used, while alternative similarity-based interfaces have yet to be widely adopted by the public, despite the increasing media attention that they have been receiving.

Labels such as genre and mood have the important advantage that they provide one with a vocabulary that can be used to discuss musical categories. Conversations concerning more general notions of similarity quickly become bogged down due to the necessity of making frequent references to musical examples. Moreover, such discussions can be unclear in terms of which dimensions of similarity are being considered.

MIR researchers should avoid adopting a patronizing approach where they insist that end users abandon a form of music retrieval for which they have a demonstrated attachment and which they find useful. A better approach is to recognize and utilize genre in MIR systems while at the same time also presenting alternatives utilizing more general similarity that can also be useful, potentially in entirely different user scenarios.

Once one accepts the usefulness of genre for end users, the utility of automatically classifying the genres of recordings stored in music databases becomes clear. Although the time needed to label training data and the noisiness of existing annotations have already been presented as serious problems when training automatic genre classifiers, the difficulty of manually labeling the entirety of huge and rapidly growing databases is much greater.

Also, similarity research has many of its own problems related to ground truth ambiguity and subjectivity, particularly when it comes to evaluating systems and comparing their performance. It is therefore inconsistent to suggest similarity as an alternative to genre classification specifically because of problems relating to ground truth.

Musical genre also has significant importance beyond simply its utility in organizing and exploring music, and should not be evaluated solely in terms of commercial applicability. Many individuals actively identify culturally with certain genres of music, as can easily be observed in the differences in the ways that many fans of death metal or rap dress and speak, for example. Genre is so important to listeners, in fact, that psychological research has found that the style of a piece can influence listeners' liking for it more than the piece itself [26]. Additional psychological research has found that categorization in general plays an essential role in music appreciation and cognition [27].

Research in automatic genre classification can also provide valuable empirical contributions to the fields of musicology and music theory (e.g., [28]). Genre research that forms correlations between particular cultural and content-based characteristics or that involves ontological structures that can successfully map genre interrelationships can also have important musicological significance.

5. Improving Automatic Genre Classification

There is truth to the criticisms that genre classifiers appear to have reached a maximum in performance, but there is no evidence that this is a ceiling that cannot be surpassed. Although continuing minor refinements are not likely to accomplish much, there are a number of major changes to how automatic genre classification is approached that could result in significant improvements.

Most genre classification systems to date have utilized primarily low-level features relating to timbre. It is not surprising that the performance of such systems has been limited, as timbre represents only a relatively small part of what humans consider when they classify music. High-level features based on musical abstractions are central to composers and performers and, as discussed in Section 2, many musicologists hold that cultural information beyond the scope of musical content is of paramount importance.

Each of these three types of features can encompass significantly different information, and combining features from the different groups could significantly improve success rates. Promising results have already been attained by combining high-level features with automatic feature selection [4, 18], and it has been experimentally demonstrated that combining cultural features mined from the web with low-level features can significantly improve performance over low-level features alone [29]. Although cultural and high-level features can be more difficult to extract than low-level features, extensive existing research in text mining can be taken advantage of to extract cultural features from the web, and improvements in transcription technology are making high-level features increasingly accessible in audio recordings as well as symbolic recordings.

An additional important issue that has rarely been addressed by published systems is that it should be possible to assign multiple genres to individual recordings, both in terms of classifier output and ground truth. Research in fuzzy logic should also be considered, and class memberships should be weighted, even if only casually. Although it can be argued that this puts an even greater load on annotators, weights do not have to be perfectly precise, as the point is simply to allow one to express some general sense of the relative importance of different genres. Allowing annotators to assign multiple labels could actually make their work easier, as they could express their real views without being confined to awkward artificial classification schemes. Most importantly, this approach would significantly improve the quality of ground truth, and would make the evaluation of systems more realistic.

Ground truth collection and labeling should be considered high priority goals in and of themselves. The construction of custom training and testing databases have traditionally comprised only a small part of larger projects, and as such have not received the attention that they warrant. Although some researchers have used existing collections such as Magnatune or Epitonic rather than constructing their own databases, they have generally not made serious efforts to refine and correct the provided metadata, which can be inconsistent or even incorrect, and have often failed to ensure that the music in such collections is representative of the commercial music that most users are actually interested in.

In general, training and testing databases tend to be too small to be sufficiently diverse or to average out annotation noise. Annotations also tend to be error-prone due to the limited expertise of individual annotators, insufficient time for thoughtful annotations and biases due to the needs of particular systems. Trained models and evaluation metrics can ultimately only be as good as the ground truth that they are built upon, and the need for high-quality ground truth must be addressed before truly successful systems can be produced. Research should therefore be performed on different ways of constructing and maintaining research databases, including comparisons of methodologies such as using panels of experts, general surveys and automated web-based mining of ground truth labels.

Another issue is that recordings are often annotated as groups based on artist or album. Although this can be effective in some limited cases, it is almost always problematic if sufficiently fine genres are being considered. For example, although one might label Christina Aguilera in general as pop, a thoughtful annotator might label some songs as pop ballads, some as dance pop and some as R&B. Furthermore, some artists (e.g., Neil Young or Miles Davis) have had such musically diverse careers that attempting to label all of their work with even a fairly broad genre is unrealistic. Serious efforts must therefore be made to annotate databases on a song by song basis.

It is also important to pay careful attention to the palette of genre labels that can be assigned to recordings. Expert opinion, general surveys and web mining could once again prove useful. Genre labels must be chosen that actually represent categories that users are interested in and that do not force annotators to make artificial decisions. Also, there should be many different candidate genres, including both coarse and broad categories. The common practice of using only ten or so categories is very unrealistic. Electronic dance music alone can easily be broken into twenty or thirty sub-genres, for example.

Incorporating some sort of ontological structure that maps the interrelationships and intersections between genre categories could also be highly beneficial. This would not only provide annotators with a helpful framework that could aid in synchronizing differences of opinion, but would also allow the use of structured classification strategies. Research has already found that even a simple hierarchical classification strategy that allows individual categories to appear in multiple branches can result in improved success rates [18]. The use of ontological genre structures would also provide end users with a structure to use when browsing music collections, and would allow exploration at different levels of granularity. A user only mildly interested in electronic dance music might be happy with an overall classification such as techno, for example, while other users would require many finer categories that might confuse the first user. The "emergent" approach proposed by Aucouturier and Pachet [14] could prove useful in constructing such ontologies.

An additional issue is that some forms of misclassification are far more significant than others. Misclassifying hard rock as heavy metal, for example, is less serious than labeling it ragtime. Failure to consider this during training and evaluation could limit the quality of a learned model. An ontological structuring as discussed above has the additional advantage that it could help to implement realistic penalization schemes during training and evaluation.

Yet another issue is that not only can different parts of a single recording belong to different genres, but different sections of a recording might be representative of the same genre in different ways. The verse and chorus of a pop song, for example, will be different from each other but will still both be characteristic of pop songs. Alternatively, different sections of a recording can belong to different genres, which might in itself be indicative of a broader genre (e.g., rap metal). In essence, different sections of a recording can correspond to separate clusters that may or may not belong to the same class, which means that individual recordings should ideally include segmented labels. This also means that averaging features over long windows or over an entire recording in order to make a single classification can be a limiting approach.

The structure of a piece and how it evolves over time can also be highly indicative of a genre. Examples include sonata form or twelve-bar blues form. This means that even classifying small windows of a recording independently can potentially ignore important information. Utilizing features that encapsulate changes over time and/or classifiers with memory (e.g., hidden Markov models or recurrent neural networks) is a potentially effective approach that has been largely neglected to date. A pre-processing system that segments recordings based on form could also help, not only by allowing musically meaningful window sizes to be set dynamically, but also by generating new features delineating form.

An additional issue, from a musicological perspective, is that researchers often use statistical techniques such as principal component analysis to reduce feature dimensionality. Although fine when considered purely in terms of success rates, this limits the quality of results from a theoretical perspective, as one loses potentially meaningful information about which musical qualities are most useful in different contexts. A more profitable approach, from a musicological perspective, might be to use feature selection techniques that reduce dimensionality while maintaining the original identity of the features, such as forward-backward selection or genetic algorithm-based selection.

There is also an important need to perform further psychological research on human genre classification. Studies should compare the classification differences between experts and non-experts, as well between individuals of different ages, cultures and musical backgrounds. This could prove beneficial not only in learning how to improve ground truth, but also in developing different systems that meet different user needs. A musicologist, a record industry scout, a teenaged consumer and a librarian, for example, will all have very different needs, and successful systems should be able to address such varied needs.

6. Conclusions

Automatic genre classification is a difficult and problematic task that nonetheless has important value in terms of both pure research and commercial application. Continuing research in automatic genre classification has much to offer, as does parallel research involving other aspects of musical similarity.

Automatic genre classification performance appears to have fallen into a local maximum recently, and serious modifications to the approaches used are needed in order to realize further improvements. The highlights of the suggestions offered in this paper are as follows:

- Information from low-level, high-level and cultural features should be combined.
- Each recording should be permitted to have more than one genre label, and labels should be weighted.
- A large number of realistic and diverse candidate genre labels of varying breadth should be used, and these genres should be organized into ontological structures.
- Misclassification penalties in training and evaluation should reflect the varying similarities between different genres.
- It should be possible to label different sections of a recording differently, and windows should be classified individually.

- Sequential classifiers and features that can encapsulate how recordings change over time should be experimented with.
- Dimensionality reduction techniques that preserve the original meaning of features should be used.
- Existing psychological, musicological and music theoretical knowledge should be taken advantage of by MIR researchers, and new empirical research in these domains should be performed to help fill the gaps in our understanding of musical genre and of how humans classify music.

Most importantly, all areas of MIR research could benefit from a concerted effort to develop carefully annotated music datasets that include varied metadata such as genre, mood, groove, composer, performer, lyrics, meter, chord progressions, instruments present, etc. Limited and/or poorly annotated ground truth is a problem that has an effect well beyond the scope of genre classification in limiting MIR research, and the benefits of a large-scale effort to construct high-quality ground truth would more than justify the extensive work needed to do so.

7. Acknowledgments

We would like to thank the researchers with whom we have informally discussed these issues, as well as those who have made insightful comments on the Music-IR and MIREX mailing lists. This includes J. Stephen Downie, Douglas Eck, Dan Ellis, Joe Futrelle, Paul Lamere, Elias Pampalk, George Tzanetakis, Kris West and the researchers at the McGill University Music Technology labs. We would also like to thank the *Social Sciences and Humanities Research Council of Canada* and the *Canada Foundation for Innovation* for their generous financial support.

References

[1] R. B. Dannenberg, B. Thom, and D. Watson. “A Machine Learning Approach to Musical Style Recognition,” in *Proceedings of the International Computer Music Conference,* 1997, pp. 344–7.

[2] G. Tzanetakis., and P. Cook. “Musical Genre Classification of Audio Signals,” *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[3] J. Bergstra, N. Casagrande, D. Erhan, D. Eck, and B. Kegl. “Aggregate Features and AdaBoost for Music Classification,” *Machine Learning*, 2006 (in press).

[4] C. McKay, and I. Fujinaga. “Automatic Genre Classification Using Large High-Level Musical Feature Sets,” in *Proceedings of the International Conference on Music Information Retrieval*, 2004, pp. 525–30.

[5] F. Fabbri. “Browsing Music Spaces: Categories and the Musical Mind,” in *Proceedings of the IASPM Conference*, 1999.

[6] D. Duff, Ed., *Modern Genre Theory,* New York: Longman, 2000.

[7] B. K. Grant, Ed., *Film Genre Reader III,* Austin, TX: University of Texas Press, 2003.

[8] F. Fabbri. “A Theory of Musical Genres: Two Applications,” in *Popular Music Perspectives,* D. Horn and P. Tagg, Eds. Göteborg: IASPM, 1981.

[9] F. Fabbri. “What Kind of Music?,” *Popular Music,* no. 2, pp. 131–43, 1982.

[10] S. Frith, *Performing Rites: On the Value of Popular Music,* Cambridge, MA: Harvard University Press, 1996.

[11] J. Toynbee, *Making Popular Music: Musicians, Creativity and Institutions,* London: Arnold, 2000.

[12] D. Brackett, *Interpreting Popular Music,* New York: Cambridge University Press, 1995.

[13] D. Brackett, *(In Search Of) Musical Meaning: Genres, Categories and Crossover*, London: Arnold, 2002.

[14] J. J. Aucouturier, and F. Pachet. “Representing Musical Genre: A State of the Art,” *Journal of New Music Research,* vol. 32, no. 1, pp. 1–12, 2003.

[15] F. Pachet, and D. Cazaly. “A Taxonomy of Musical Genres,” in *Proceedings of the Content-Based Multimedia Information Access Conference,* 2000.

[16] G. Lakoff, *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind,* Chicago: University of Chicago Press, 1987.

[17] I. Deliege. “Prototype Effects in Music Listening: An Empirical Approach to the Notion of Imprint,” *Music Perception,* vol. 18, no. 33, pp. 371–407, 2001.

[18] C. McKay. “Automatic Genre Classification of MIDI Recordings,” *M.A. Thesis,* McGill University, Canada, 2004.

[19] D. Perrot, and R. Gjerdigen. “Scanning the Dial: An Exploration of Factors in the Identification of Musical Style,” in *Proceedings of the Society for Music Perception and Cognition,* 1999, p. 88 (abstract).

[20] S. Lippens, J. P. Martens, M. Leman, B. Baets, H. Meyer, and G. Tzanetakis. “A Comparison of Human and Automatic Musical Genre Classification,” in *Proceedings of the IEEE International Conference on Audio, Speech and Signal Processing,* 2004.

[21] “AllMusic Guide,” [Web site] 2006, [2006 March 29], Available: http://www.allmusic.com

[22] “Gracenote,” [Web site] 2006, [2006 March 29], Available: http://www.gracenote.com

[23] “MIREX 2005 Contest Results,” [Web site] 2005, [2006 April 18], Available: http://www.music-ir.org/evaluation/mirex-results.

[24] J. J. Aucouturier, and F. Pachet. “Improving Timbre Similarity: How High is the Sky?,” *Journal of Negative Results in Speech and Audio Sciences,* vol. 1, no. 1. 2004.

[25] J. H. Lee, and J. S. Downie. “Survey of Music Information Needs, Uses, and Seeking Behaviours: Preliminary Findings,” in *Proceedings of the International Conference on Music Information Retrieval,* 2004.

[26] A. C. North, and D. J. Hargreaves. “Liking for Musical Styles,” *Music Scientae,* vol. 1, no. 1, pp. 109–28, 1997.

[27] H. G. Tekman, and N. Hortacsu. “Aspects of Stylistic Knowledge: What are Different Styles Like and Why Do We Listen to Them?,” *Psychology of Music,* vol. 30, no. 1, pp. 28–47, 2002.

[28] C. McKay, and I. Fujinaga. “Automatic Music Classification and the Importance of Instrument Identification,” in *Proceedings of the Conference on Interdisciplinary Musicology,* 2005.

[29] B. Whitman, and P. Smaragdis. “Combining Musical and Cultural Features For Intelligent Style Detection,” in *Proceedings of the International Conference on Music Information Retrieval,* 2002, pp. 47–52.