To appear in *Proceedings of 19th IEEE International Conference on Tools with Artificial Intelligence* (ICTAI-07), Patras, Greece, October 29-31, 2007.

A Music Information Retrieval Approach Based on Power Laws

Patrick Roos and Bill Manaris

Computer Science Department, College of Charleston, 66 George Street, Charleston, SC 29424, USA
{patrick.roos, manaris}@cs.cofc.edu

Abstract

We present a music information retrieval approach based on power laws. Research in cognitive science and neuroscience reveals connections between power laws, human cognition, and human physiology. Empirical studies also demonstrate connections between power laws and human aesthetics. We utilize 250+ power-law metrics to extract statistical proportions of music-theoretic and other attributes of music pieces. We discuss an experiment where artificial neural networks classify 2,000 music pieces, based on aesthetic preferences of human listeners, with 90.70% accuracy. Also, we present audio results from a music information retrieval experiment, in which a music search engine prototype retrieves music based on "aesthetic" similarity from a corpus of 15,200+ pieces. These results suggest that power-law metrics are a promising model of music aesthetics, as they may be capturing statistical properties of the human hearing apparatus.

1. Introduction

The field of music information retrieval (MIR) focuses on retrieving information from large, on-line repositories of music content, using various forms of query-based or navigation-based approaches [1, 2, 3, 4]. MIR techniques can be applied in a wide variety of contexts, ranging from searches in music libraries (e.g., [5]), to consumer-oriented music e-commerce environments [6].

Given today's commercial music libraries with millions of music pieces (and with hundreds of new pieces added monthly), MIR approaches that utilize (even partial) models of human aesthetics are of great importance. This paper describes such an MIR approach, which utilizes power-law metrics.

In earlier work, we have shown that metrics based on power laws (e.g., Zipf's law) comprise a promising approach to modeling music aesthetics [7, 8, 9]. Herein, we present new results on the relationship between power-law metrics and aesthetics for music classification and MIR. First, we discuss a large-scale experiment where an artificial neural network (ANN) classifies 2,000 music pieces into two categories, based on data related to human aesthetic preferences, with 90.70% accuracy. Then, we present audio results from a MIR experiment, where a search engine utilizing power-law metrics automatically retrieves "aesthetically" similar music pieces from a 15,200+ corpus.

2. Background

Content-based MIR approaches focus (a) on extracting features from music pieces, and (b) on using these features, in conjunction with machine learning techniques, to automatically classify pieces, e.g., by composer, genre, mood, etc.

Tzanetakis and Cook (2002) work at the audio level with three types of features, i.e., timbral texture features, rhythmic content features, and pitch content features [10]. They classify 1000 music pieces distributed equally across 10 musical genres (i.e., Blues, Classical, Country, Disco, Hip-Hop, Jazz, Metal, Pop, Regae, and Rock) with an accuracy of 61%. (This is one of the most referenced studies in the music classification literature.)

Basili et al. (2004) work at the MIDI level with features based on melodic intervals, instruments, instrument classes and drum kits, meter/time changes, and pitch range [11]. They classify approx. 300 MIDI files from six genres (i.e., Blues, Classical, Disco, Jazz, Pop, Rock) with accuracy near 70%.

Dixon et al. (2004) work at the audio level with features based strictly on rhythm, including various features derived from histogram calculations [12]. They classify 698 pieces from the 8 ballroom dance subgenres from the ISMIR 2004 Rhythm classification contest (i.e., Cha Cha, Jive, Quickstep, Rumba, Samba, Tango, Viennese Waltz, and Waltz) with an accuracy of 96%. It should be noted that rhythm classification is much easier than general genre classification.

Lidy and Rauber (2005) work at the audio level with features similar to [12] including various psycho-acoustic transformations [13]. They use three different corpora, namely the set from the ISMIR 2004 Rhythm classification contest (698 pieces across 8 genres); the set from the ISMIR 2004 Genre classification contest (1458 pieces across 10 genres); and the set used by [10] (1000 pieces across 10 genres). Classification experiments reach accuracies of 70.4%, 84.2%, and 74.9%, for each corpus, respectively.

McKay and Fujinaga (2004) work at the MIDI level with 109 features based on instrumentation, texture, rhythm, dynamics, pitch statistics, melody and chords [14]. They classify 950 pieces from three broad genres (Classical, Jazz, Popular) with an accuracy of 98%. However, according to Karydis et al. (2006), "the system requires training for the 'fittest' set of features, a cost that trades-off the generality of the approach with the overhead of feature selection." [15]

Li et al. (2003) work at the audio level with statistical features, which capture amplitude variations [16]. On the same set of 1000 music pieces used by [10], they classify across the 10 genres with an accuracy of 78.5%, a significant improvement over [10].

Karydis et al. (2006) work at a MIDI-like level with features based on repeating patterns of pitches, and selected properties of pitch and duration histograms. On a corpus of 250 music pieces spanning 5 classical subgenres (i.e., ballads, chorales, fugues, mazurkas, sonatas), they reach an accuracy of approximately 90% [15].

In the next section, we discuss features based on power laws as a promising new approach for MIR applications.

3. Power Laws and Music

A power law denotes a relationship between two variables where one is proportional to a power of the other. One of the most well-known power laws is *Zipf's law*:

$$P(f) \sim 1 / f^{n} \qquad (1)$$

where $P(f)$ denotes the probability of an event of rank f, and n is close to 1. It is named after the Harvard linguist, George Kingsley Zipf, who studied it extensively in natural and social phenomena [23]. The generalized form is:

$$P(f) \sim a / f^{b} \qquad (2)$$

where a and b are real constants.

Theories of aesthetics suggest that artists may subconsciously introduce power-law proportions into their artifacts by trying to strike a balance between chaos and order [17, 18]. Empirical studies demonstrate connections between power laws and human aesthetics [8, 19, 20, 21]. For instance, "socially-sanctioned" (popular) music exhibits power laws across various attributes [7, 21, 22, 24]. Finally, power laws have been used to automatically generate aesthetically pleasing music, further validating the connection between power laws and aesthetics [9, 23].

In earlier work, we developed a large set of power law metrics (currently more than 250), which we use to measure statistical proportions of a variety of music theoretic and other attributes. These attributes include pitch, duration, melodic intervals, harmonic intervals, as well as higher-order and local-variability variants of these metrics [9]. Each of these metrics, creates a log-log plot of $P(f)$ and f, computes the linear regression of the data points, and returns two values: the slope of the trendline, b, and the strength of the linear relation, r^2 [8]. These values are used as features in classification experiments.

These features have been validated through ANN classification experiments, including composer identification with 93.6% to 95% accuracy [25, 26]; and pleasantness prediction using emotional responses from humans with 97.22% accuracy [8].

Currently, we are conducting various style classification experiments. Our corpus consists of 1566 pieces from various genres, including Renaissance, Baroque, Classical, Romantic, Impressionist, Modern, Jazz, Country, and Rock. Our results range from 71.52% to 96.66% accuracy (pending publication).

In addition to genre classification, we are exploring the applicability of power-law metrics for modeling aesthetic preferences of listeners. This type of validation goes beyond traditional style classification experiments (e.g., see Section 2).

In an earlier experiment, we trained ANNs to classify 210 music excerpts according emotional responses from human listeners. Using a 12-fold, cross-validation study, ANNs achieved an average success rate of 97.22% in predicting (within one standard deviation) human emotional responses to those pieces [8].

The following section presents a large-scale experiment exploring the connection between power laws and music aesthetics.

4. A Classification Experiment Based on Aesthetic Preferences

The problem with assessing aesthetics is that (similarly to assessing intelligence) there seems to be no objective way of doing so. One possibility is to use a variant of the Turing Test, where we ask humans to rate the aesthetics of music pieces, and then check for correlations between those ratings and features extracted using our power-law metrics. In this section, we explore this approach.

For this experiment, we trained ANNs to classify 2,000 pieces into two categories using aesthetic preferences provided by humans. We used the Classical Music Archives (CMA) corpus, which consists of 14,695 classical MIDI encoded pieces. A download log for November 2003 (1,034,355 downloads) served to identify the 1000 most downloaded vs. the 1000 least downloaded pieces.[1] Given this configuration, the *most-preferred* vs. *least-preferred* classes were separated by over 12,000 pieces. Although there may exist other possibilities for a piece's preference among CMA listeners (e.g., how famous it is), given the size of the corpus and the large

[1] A pilot study appears in [9].

separation between the two classes, we believe that these possibilities are for the most part subsumed by aesthetic preference.[2]

First, we conducted a classification task using 156 features per piece to train an ANN. These features consisted of the 13 regular metrics, two higher-order metrics for each regular metric, and a local-variability metric for each regular and higher-order metric.

For control purposes, we conducted a classification task identical to the first, but with classes assigned randomly for each piece.

Finally, we conducted a classification task identical to the first, but using only 12 most relevant slope values to train the ANN. These attributes were selected to be most correlated with a class, but least correlated with each other, by searching a space of attribute subsets through greedy hill-climbing augmented with a backtracking facility.

All classification tasks involved feed-forward ANNs trained via backpropagation. Training ran for 500 epochs, with a value of 0.2 for momentum and 0.3 for learning rate. The ANNs contained a number of nodes in the input layer equal to the features used for training, 2 nodes in the output layer and *(input nodes + output nodes)/2* nodes in the hidden layer. For evaluation, we used 10-fold cross validation. The corpus of 2,000 songs was separated randomly into 10 unique parts; the ANN was trained on 9 out of the 10 parts (90% training), and evaluated on the 1 remaining part (10% testing). This process was repeated 10 times, each time choosing a different testing part. The average success rate was reported.

4.1. Results and Discussion

For the first classification task, the ANN classified 1,814 of the 2,000 pieces correctly, achieving a success rate of 90.70%. Table 1 shows the confusion matrix. In the control run, with classes assigned randomly, the ANN classified 1,029 pieces correctly, a success rate of 51.45%. This suggests that the high success rates of the first classification task are largely due to the effectiveness of the extracted music features.

In the final classification task, using only the 12 most relevant slope values for training, the ANN still achieved a success rate of 83.29% (see Table 2). This and other results suggest that many of the original 156 features are highly correlated. Tables 3 and 4 provide basic statistics for the 156 features and 12 selected features, respectively, for the two classes. It should be noted that the 12 selected slopes for most preferred pieces (Table 4) approximate an ideal Zipfian slope of –1 (average of –1.0621), whereas the slopes for least preferred pieces indicate more chaotic proportions (average of –0.8846). This is consistent with slopes seen in earlier studies [7, 8, 24].

Table 1. Confusion matrix for ANN classification with all 156 features (**bold** denotes correct).

		ANN Output	
		Most Preferred	*Least Preferred*
Actual	*Most Preferred*	**917**	83
	Least Preferred	103	**897**

Table 2. Success rates of different ANN classification experiments.

Classification Experiment	*Success (%)*
ANN with 156 features	90.70%
ANN with 12 selected features	83.29%
ANN with 156 features and randomly assigned classes (control)	51.45%

Table 3. Average and standard deviation (*Std*) of *slope* and r^2 values across all 156 features for most and least preferred music pieces.

Class	*Value*	*Average*	*Std*
Most Preferred	*slope*	- 0.8834	0.5213
	r^2	0.7663	0.2150
Least Preferred	*slope*	- 0.7930	0.4723
	r^2	0.7526	0.2252

Table 4. Average and standard deviation (*Std*) of 12 most relevant *slopes* and the corresponding r^2 values for most and least preferred music pieces.

Class	*Value*	*Average*	*Std*
Most Preferred	*slope*	- 1.0621	0.3883
	r^2	0.7366	0.2089
Least Preferred	*slope*	- 0.8846	0.3457
	r^2	0.7325	0.2548

The 12 most relevant features (slope values) were related to chromatic tone and harmonic/melodic consonance. Interestingly, similar metrics were also found to be most relevant in our previous classification experiment involving emotional responses of listeners [8]. These results are consistent with music theory, and suggest that our metrics are capturing aspects of music aesthetics.

5. A Music Search Experiment

Motivated by the high success rates of classification experiments validating power-law metrics, we created a

[2] These pieces have been around for more than 100 years. Both groups share composers, genres, and form (e.g., fugue). The only difference between them is that listeners have considerably more preference for one group than the other; otherwise the two groups are hard to differentiate.

prototype of a music search-engine that utilizes such metrics for music retrieval based on aesthetic similarity. In this section, we report empirical results from this effort.

As far as the search engine is concerned, each music piece is represented as a vector of 250+ power-law slope and r^2 values. As input, the engine is presented with a single music piece. The engine searches the corpus for pieces "aesthetically" similar to the input, computing the mean squared error (MSE) of the vectors. The pieces with the lowest MSE (relative to the input) are returned as best matches.

For this experiment, we used the CMA corpus (14,695 MIDI pieces) augmented with 500+ MIDI pieces from other music genres including Jazz, Rock, Country, and Pop (a total of 15,200+ music pieces). As input, the music search engine was given random pieces from the corpus, and returned the three best matches for each of the inputs.

5.1. Results and Discussion

Table 5 shows the output from a typical query. This and other examples (with audio) may be found at http://www.cs.cofc.edu/~manaris/music-search. Readers may assess for themselves the aesthetic similarity between the input and the retrieved pieces.

An intriguing observation is that the search engine discovers similarities across established genres. For instance, searching for music similar to Miles Davis' "Blue in Green" (Jazz), identifies a very similar (MSE 0.0043), yet obscure cross-genre match: Sir Edward Elgar's "Chanson de Matin" (Romantic). Such matches can be easily missed even by expert musicologists. We think the ability to find such matches is empowering, given today's commercial music libraries with millions of pieces. This preliminary experiment demonstrates the potential of a music search engine based on aesthetic similarity captured via power-law metrics.

6. Conclusion

In this paper, we have described a MIR approach based on power-law metrics. We presented two experiments applying this approach: (a) a classification experiment based on aesthetic preferences of human listeners, and (b) a music retrieval experiment, along with audio results, on searching a music collection by aesthetic similarity.

The results of the first experiment are intriguing. Have we discovered a "black box" that can predict the popularity of music? Or have we discovered a model of music aesthetics, i.e., a model that captures relevant statistical properties of the human hearing apparatus (i.e., proportions of sounds that are pleasing to the ear)? Earlier work (e.g., [17, 18, 19, 21, 22, 23]) supports the second interpretation. To verify, we are exploring new techniques for assessing MIR technology based on measuring human emotional responses. The experimental methodology is partially described in [8]. Early results are supportive of the aesthetics claim [9].

Table 5. Sample input pieces (*in italics*) and results (pieces with lowest MSE) from the music search engine.

	Music Piece	***MSE***
Input	Classical, BEETHOVEN, Ludwig van: 8 Lieder, Op.52, 6.Das Blümchen Wunderhold	0.0000
Output	**1)** Classical, BURGMÜLLER, Johann Friedrich: Etudes, Op.100, No.1, La Candeur	0.0182
	2) Classical, BEETHOVEN, Ludwig van Bagatelles, Op.126, 5.Quasi allegretto in G	0.0191
	3) Classical, BEETHOVEN, Ludwig van: 8 Lieder, Op.52, 3.Das Liedchen von der Ruhe	0.0193

Finally, we are adapting our metrics for use with audio formats (as opposed to only MIDI). Preliminary results are encouraging. A music search engine with the ability to identify aesthetically similar music may have significant implications for music retrieval on the Web (e.g. Google), the music industry (e.g. iTunes), and digital libraries (e.g. the US National Science Digital Library). Since music permeates society, the proposed MIR approach may have significant societal implications, as it may drastically enhance the way people access and enjoy music.

Acknowledgements

We are thankful to Juan Romero, Penousal Machado and their students for their contributions to power-law metrics validation using ANNs, and for discussions on modeling art aesthetics. Dwight Krehbiel and his students have contributed, through discussions and music psychology experiments with human subjects, to the assessment of the connection between power-law metrics and aesthetics. Luca Pellicoro and Thomas Zalonis contributed through coding, testing, and discussions. Walter Pharr has contributed through discussions on classical music genres and time periods. This project has been supported by the College of Charleston, and the Classical Music Archives (http://www.classicalarchives.com).

References

[1] P. Cano, M. Koppenberger, and N. Wack, "Content-Based Music Audio Recommendation", in *Proceedings of the 13th*

Annual ACM International Conference on Multimedia (MULTIMEDIA '05), Hilton, Singapore, Nov. 2005, pp. 211–212.

[2] H.H. Hoos, and D. Bainbridge, “Editors’ Note”, special issue on Music Information Retrieval, *Computer Music Journal*, 2004, 28(2): 4-5.

[3] B. Pardo, “Music Information Retrieval”, *Communications of the ACM*, 2006, 49(8): 28-31.

[4] P.-Y. Rolland, “Music Information Retrieval: A Brief Overview of Current and Forthcoming Research”, in *Proceedings of 1st International Workshop on Human Supervision and Control in Engineering and Music*, Stadthalle Kassel, Germany, Sep. 2001.

[5] J.W. Dunn, D. Byrd, M. Notess, J. Riley, and R. Scherle, “Variations2: etrieving and using Music in an Academic Setting”, *Communications of the ACM*, 2006, 49(8): 53-58.

[6] D. Byrd, “Music-Notation Searching and Digital Libraries”, in *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries* (JCDL '01), Roanoke, Virginia, 2001, pp. 239-246.

[7] B. Manaris, T. Purewal, and C. McCormick, “Progress Towards Recognizing and Classifying Beautiful Music with Computers - MIDI-Encoded Music and the Zipf-Mandelbrot Law”, in *Proceedings of the IEEE SoutheastCon 2002 Conference*, Columbia, SC, Apr. 2002, pp. 52-57.

[8] B. Manaris, J. Romero, P. Machado, D. Krehbiel, T. Hirzel, W. Pharr, and R.B. Davis, “Zipf’s Law, Music Classification and Aesthetics.”, *Computer Music Journal*, 29(1), MIT Press, 2005, pp. 55-69.

[9] B. Manaris, P. Roos, P. Machado, D. Krehbiel, L. Pellicoro, and J. Romero, “A Corpus-Based Hybrid Approach to Music Analysis and Composition”, in *Proceedings of the 22nd Conference on Artificial Intelligence* (AAAI-07), Vancouver, BC, Jul. 2007, pp. 839-845.

[10] G. Tzanetakis, and P. Cook, “Musical Genre Classification of Audio Signals”, *IEEE Transactions on Speech and Audio Processing*, 2002, 10 (5): 293–302.

[11] R. Basili, A. Serafini, and A. Stellato, “Classification of Musical Genre: A Machine Learning Approach”, in *Proceedings of the 5th International Conference on Music Information Retrieval* (ISMIR-04), Barcelona, Spain, Oct. 2004.

[12] S. Dixon, F. Gouyon, and G. Widmer, “Towards Characterisation of Music via Rhythmic Patterns”, in *Proceedings of the 5th International Conference on Music Information Retrieval* (ISMIR-04), Barcelona, Spain, Oct. 2004.

[13] T. Lidy, and A. Rauber, “Evaluation of Feature Extractors and Psycho-acoustic Transformations for Music Genre Classification”, in *Proceedings of the 6th International Conference on Music Information Retrieval* (ISMIR-05), London, UK, Sep. 2005, pp. 34-41.

[14] C. McKay, and I. Fujinaga, “Automatic Genre Classification using Large High-level Musical Feature Sets”, in *Proceedings of the 5th International Conference on Music Information Retrieval* (ISMIR-04), Barcelona, Spain, Oct. 2004, pp. 525-530.

[15] I. Karydis, A. Nanopoulos, and Y. Manolopoulos, “Symbolic Musical Genre Classification based on Repeating Patterns”, in *Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia* (AMCMM '06), Santa Barbara, CA, Oct. 2006, pp. 53-58.

[16] T. Li, M. Ogihara, and Q. Li, “A Comparative Study on Content-Based Music Genre Classification”, in *Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada, Jul. 2003, pp. 282-289.

[17] M. Schroeder, *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*. New York: W. H. Freeman and Company, 1991.

[18] R. Arnheim, *Entropy and Art: an Essay on Disorder and Order*, Berkeley: University of California Press, 1971.

[19] N.A. Salingaros, and B.J. West, “A Universal Rule for the Distribution of Sizes”, *Environment and Planning B: Planning and Design*, 1999, 26: 909-923.

[20] B. Spehar, C.W.G. Clifford, B.R. Newell, and R.P. Taylor, “Universal Aesthetic of Fractals.” *Computers & Graphics*, 2003, 27: 813-820.

[21] R.F. Voss, and J, Clarke,“1/f Noise in Music and Speech”, *Nature*, 1975, 258: 317-318.

[22] R.F. Voss, and J. Clarke, “1/f Noise in Music: Music from 1/f Noise”, *Journal of Acoustical Society of America*, 1978, 63(1): 258–263.

[23] G.K. Zipf, Human Behavior and the Principle of Least Effort, Hafner Publishing Company, 1949.

[24] B. Manaris, D. Vaughan, C. Wagner, J. Romero, and R.B. Davis, “Evolutionary Music and the Zipf–Mandelbrot Law – Progress towards Developing Fitness Functions for Pleasant Music”, *Applications of Evolutionary Computing*, LNCS 2611, Springer-Verlag, 2003, pp. 522-534.

[25] P. Machado, J. Romero, B. Manaris, A. Santos, and A. Cardoso, “Power to the Critics - A Framework for the Development of Artificial Critics”, in *Proceedings of 3rd Workshop on Creative Systems, 18th International Joint Conference on Artificial Intelligence* (IJCAI 2003), Acapulco, Mexico, 2003, pp. 55-64.

[26] P. Machado, J. Romero, M.L. Santos, A. Cardoso, and B. Manaris, “Adaptive Critics for Evolutionary Artists”, *Applications of Evolutionary Computing*, LNCS 3005, Springer-Verlag, 2004, pp. 437-446.