Developing Fitness Functions for Pleasant Music: Zipf's Law and Interactive Evolution Systems

Bill Manaris[1], Penousal Machado[2], Clayton McCauley[3],
Juan Romero[4], and Dwight Krehbiel[5]

[1,3] Computer Science Department, College of Charleston, 66 George Street,
Charleston, SC 29424, USA
{manaris, mccauley}@cs.cofc.edu
[2] Instituto Superior de Engenharia de Coimbra, Qta. da Nora, 3030 Coimbra, Portugal
machado@dei.uc.pt
[4] Creative Computer Group - RNASA Lab - Faculty of Computer Science,
University of A Coruña, Coruña, Spain
jj@udc.es
[5] Psychology Department, Bethel College, North Newton KS, 67117, USA
krehbiel@bethelks.edu

Abstract. In domains such as music and visual art, where the quality of an individual often depends on subjective or hard to express concepts, the automating fitness assignment becomes a difficult problem. This paper discusses the application of Zipf's Law in evaluation of music pleasantness. Preliminary results indicate that a set of Zipf-based metrics can be effectively used to classify music according to pleasantness as reported by human subjects. These studies suggest that metrics based on Zipf's law may capture essential aspects of proportion in music as it relates to music aesthetics. We discuss the significance of these results for the automation of fitness assignment in evolutionary music systems.

1 Introduction

Interactive Evolution (IE) is one of the most popular approaches in current evolutionary music generation systems. In this paradigm the user assigns fitness to the generated pieces, guiding evolution according to his/hers aesthetic preferences. In the field of music, IE has been used for the evolution of rhythmic patterns, melodies, Jazz improvisations, composition systems, and many other applications (a comprehensive survey can be found in [1]).

In spite of its popularity, IE has several shortcomings that become particularly severe in time-based domains like music. Listening to all generated pieces is a tedious and demanding task; it leads to user fatigue and inconsistency in evaluation, and imposes severe limits on population size and number of generations. To overcome this shortcoming, some researchers (e.g. [2, 3, 4]) resort to Artificial Neural Networks (ANNs). The ANNs can be trained using a set of user-evaluated pieces created by an IE system [3]; scores of well-known musicians [2]; rhythmic boxes [4]; etc.

F. Rothlauf et al. (Eds.): EvoWorkshops 2005, LNCS 3449, pp. 498–507, 2005.

Although appealing, this approach has several shortcomings (see e.g. [2, 3]), most notably the difficulty of identifying a representative training set and, consequentially, of avoiding shortcuts – ways of creating false maximums.

Our research explores the connection between Zipf's law and music in the context of developing fitness functions for evolutionary music systems. We begin by performing an analysis of the music by extracting several Zipf-based measurements. These measurements serve as input for ANNs. We have successfully performed several validation experiments for author and style identification. In this paper, we describe a similar experiment, in the context of predicting music pleasantness.

The next sections discuss Zipf's law and its connection to music, and present results demonstrating how Zipf's law may be used to quantify music pleasantness. These results suggest that Zipf's law is a useful tool for developing fitness functions for evolutionary music.

1.1 Zipf's Law

Zipf's law reflects the scaling properties of many phenomena in human ecology, including natural language and music [5, 6]. Informally, it describes phenomena where certain types of events are quite frequent, whereas other types of events are rare. In English, for instance, short words (e.g., "a", "the") are quite frequent, whereas long words (e.g., "anthropomorphologically") are quite rare. In music, consonant harmonic intervals are more frequent, whereas dissonant harmonic intervals are quite rare, among other examples. In its most succinct form, Zipf's law is expressed in terms of the frequency of occurrence (quantity) of events, as follows:

$$F \sim r^{-a} \tag{1}$$

where F is the frequency of occurrence of an event within a phenomenon, r is its statistical rank (position in an ordered list), and a is close to 1.

Another formulation of Zipf's law is

$$P(f) \sim 1/f^{n} \tag{2}$$

where $P(f)$ denotes the probability of an event of rank f and n is close to 1. In physics, Zipf's law is a special case of a *power law*. When n is 1 (Zipf's ideal), the phenomenon is called *1/f* or *pink noise*. Interestingly, when rendered as audio, *1/f* (pink) noise is perceived by humans as balanced, whereas $1/f^{0}$ or *white noise* is perceived as too random, and $1/f^{2}$ or *brown noise* as too correlated [6].

In the case of music, we may study the frequency of occurrence of pitch events, duration events, melodic interval events, and so on. For instance, consider Chopin's "Revolutionary Etude." To determine if its melodic intervals follow Zipf's law, we count the different melodic intervals in the piece, e.g., 89 half steps up, 88 half steps down, 80 unisons, 61 whole steps up, and so on. Then we plot these counts against their statistical rank on log-log scale. This plot is known as *rank-frequency* distribution (see Fig. 1).

In general, the slope of the distribution may range from 0 to $-\infty$, with -1 denoting Zipf's ideal. This slope corresponds to the exponent n in (2). The R^2 value may range from 0 to 1, with 1 denoting a straight line. The straighter the line, the more reliable

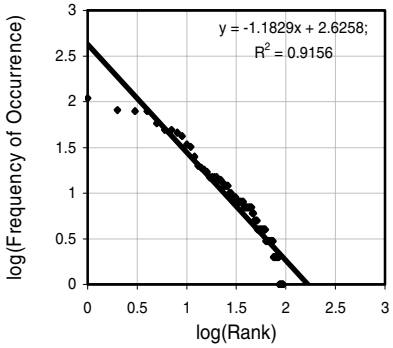

Fig. 1. The rank-frequency distribution of melodic intervals for Chopin's "Revolutionary Etude," Op. 10 No. 12 in C minor

the measurement. For example, melodic intervals in Chopin's "Revolutionary Etude" approximate a Zipfian distribution with slope of –1.1829 and R2 of 0.9156.

2 Experimental Studies

Earlier studies indicate that Zipfian distributions abound in socially-sanctioned music [7]. By *socially-sanctioned* we mean music that is sanctioned by a large enough musical subculture to be published/recorded, and thus survive over time; this is consistent with Zipf's use of the term (see [5], p. 329)

Currently, we have a set of 40 metrics based on Zipf's law [8]. We have used these metrics to extract features from MIDI-encoded music pieces. Specifically, these metrics count occurrences of various types of events and calculate the slope and R^2 value of the corresponding Zipf distribution. Table 1 shows a subset of these metrics.

The features extracted from these metrics (i.e., slope and R^2 values) have been used to train ANNs to classify these pieces in terms of composer, style, and pleasantness. To perform these classification studies, we compiled several corpora, whose size ranged across experiments from 12 to 758 music pieces [8]. These pieces are MIDI-encoded performances, the majority of which come from the Classical Music Archives [9]. We applied Zipf metrics to extract various features per music piece. The number of features per piece varied across experiments, ranging from 30 to 81.

These feature vectors were separated into two data sets. The first set was used for training the ANN. The second set was used to test the ANN's ability to classify new data. We experimented with various architectures and training procedures using the Stuttgart Neural Network Simulator [10].

In terms of author attribution, we conducted five experiments: *Bach vs. Beethoven*, *Chopin vs. Debussy*, *Bach vs. four other composers*, and *Scarlatti vs. Purcell vs. Bach vs. Chopin vs. Debussy* [11, 12]. The average success rate across the five author attribution experiments ranged from 95% to 100%.

Table 1. A sample of metrics based on Zipf's law [8]

Metric	Description
Pitch	Rank-frequency distribution of the 128 MIDI pitches
Chromatic tone	Rank-frequency distribution of the 12 chromatic tones
Duration	Rank-frequency distribution of note durations
Pitch duration	Rank-frequency distribution of pitch durations
Pitch distance	Rank-frequency distribution of length of time intervals between note (pitch) repetitions
Harmonic interval	Rank-frequency distribution of harmonic intervals within chord
Harmonic consonance	Rank-frequency distribution of harmonic intervals within chord based on music-theoretic consonance
Melodic interval	Rank-frequency distribution of melodic intervals within voice
Harmonic bigrams	Rank-frequency distribution of adjacent harmonic interval pairs
Melodic bigrams	Rank-frequency distribution of adjacent melodic interval pairs

We conducted several experiments for style identification tasks, using different ANN architectures and parameters. A detailed description and analysis of these results is awaiting publication. The average success rate across experiments, which required discerning between seven different styles, ranged from 91% to 95%.

These studies suggest that Zipf-based metrics may be used effectively for ANN classification, in terms of authorship attribution and style identification. These two tasks are relevant to evolutionary music composition, as it may contribute to fitness functions for composing music that is similar to a certain composer or music style. The next session presents ANN results related to music pleasantness.

3 Pleasantness Prediction

Much psychological evidence indicates that *pleasantness* and *activation* are the fundamental dimensions needed to describe human emotional responses [13]. Following established standards, we conducted an experiment in which we asked 21 subjects to classify music in terms of pleasantness and activation. The subjects were college students with varied musical backgrounds. The experiment was double blind, in that neither the subjects nor the people administering the experiment knew which of the pieces presented to the subjects were presumed as pleasant or unpleasant.

3.1 Data Collection Methodology

The subjects were presented with 12 MIDI-encoded musical performances. Our goal was to provide six pieces that an average person might find pleasant, and six pieces that an average person might find unpleasant. A member of our team with extensive music theory background helped identify 12 such pieces (see Table 2). From these pieces, we extracted excerpts up to two minutes long, in order to lessen fatigue for the human subjects and thus increase the consistency of the collected data.

Table 2. Twelve pieces used for music pleasantness classification study. Subjects rated the first six pieces as "pleasant", and the last six pieces as "unpleasant"

Composer	Piece	Duration
Beethoven	Sonata No. 20 in G. Opus 49. No. 2	(1:00)
Debussy	Arabesque No.1 in E (Deux Arabesques)	(1:34)
Mozart	Clarinet Concerto in A. K.622 (1. Allegro)	(1:30)
Schubert	Fantasia in C minor. Op.s 15	(1:58)
Tchaikovsky	Symphony 6 in B minor. Opus 36. Movement 2	(1:23)
Vivaldi	Double Violin Concerto in A minor. F.1. No. 177	(1:46)
Bartok	Suite. Op. 14	(1:09)
Berg	Wozzeck (trans. for piano)	(1:38)
Messiaen	Apparation de l'Eglise Eternelle	(1:19)
Schönberg	Pierrot Lunaire (5. Valse de Chopin)	(1:13)
Stravinksy	Rite of Spring. Movement 2 (tran. for piano)	(1:09)
Webern	Five Songs (1. Dies ist ein Lied)	(1:26)

While listening to the music, the subjects continuously repositioned the mouse in a 2D selection space to indicate their reaction to the music. The horizontal dimension represented *pleasantness* while the vertical dimension represented *activation* or arousal. The system recorded the subject's cursor coordinates once per second. Positions were recorded on 0 to 100 scales with the point (50,50) representing emotional indifference or neutral reaction.

Similar methods for continuous recording of emotional response to music have been used elsewhere [14].

3.2 ANN Training Methodology

For the ANN experiment, we divided each music excerpt into segments. All segments started at 0:00 and extended in increments of four seconds. That is, the first segment extended from 0:00 to 0:04 seconds, the second segment from 0:00 to 0:08 seconds, the third segment from 0:00 to 0:012 seconds, and so on. We applied Zipf metrics to extract 81 features per music increment. Each feature vector was associated with a target output vector (x, y), where x and y ranged between 0.0 and 1.0. Target vectors were constructed from the exact ratings (averaged over subjects) at each point in time in the piece. Target vector (1.0, 0.0) corresponded to most pleasant, (0.0, 1.0) corresponded to most unpleasant, and (0.5, 0.5) corresponded to neutral. This generated a total of 210 training vectors.

We conducted a 12-fold, "leave-one-out," cross-validation study. This allowed for 12 possible combinations of 11 pieces to be "learned" and 1 piece to be tested. The ANN had a feed-forward architecture with 81 elements in the input layer, 18 in the hidden layer, and 2 in the output layer. Internally, the ANN was divided into two 81x9x1 "Siamese-twin" pyramids both sharing the same input layer. One pyramid was trained to recognize pleasant music, the other unpleasant. Classification was based on the average of the two outputs.

Table 3. ANN results from all 12 experiments for the human-training, human-testing condition

Composer	Cycles	Test Rate	Test MSE	Train Rate	Train MSE
Beethoven	11200	100.00%	0.002622	100.00%	0.011721
Debussy	151000	100.00%	0.086451	100.00%	0.001807
Mozart	104000	100.00%	0.003358	100.00%	0.005799
Schubert	194000	100.00%	0.012216	100.00%	0.002552
Tchaikovsky	4600	100.00%	0.002888	100.00%	0.019551
Vivaldi	2600	100.00%	0.002026	94.05%	0.046553
Bartók	20200	100.00%	0.006760	100.00%	0.008813
Berg	4600	80.95%	0.100619	100.00%	0.015412
Messiaen	35200	100.00%	0.001315	100.00%	0.008392
Schönberg	4400	100.00%	0.013170	99.49%	0.024644
Stravinksy	10800	100.00%	0.000610	100.00%	0.015685
Webern	6400	100.00%	0.006402	100.00%	0.015366
Average	45750	98.41%	0.019870	99.46%	0.014691
Std	66150	0.0549	0.034775	0.0171	0.012118

During each training cycle the ANN was presented with every training vector once, in random order. Using back-propagation, the ANN weights were adjusted to reduce output mean standard error (*train MSE*). Every 200 cycles, the ANN was tested against the test data keeping track of the output mean standard error (*test MSE*). If the test MSE did not improve after a number of cycles, the ANN was considered stuck at a local minimum. Using a simulated annealing schedule, the ANN weights were "jogged" (adjusted by adding small amounts of random noise to the original weights). This forced the ANN to explore neighboring areas in the search space. The ANN weights were jogged with decreasing frequency as training progressed. The back-propagation part of the training focused on minimizing the train MSE, whereas the simulated-annealing part focused on minimizing the test MSE. By combining back-propagation with simulated annealing, we aimed at finding the best possible fit of the training data given the test data.

3.3 Experimental Results

The ANN performed extremely well with an average success rate of 98.41%. All pieces were classified with 100% accuracy, with one exception: Berg's piece was classified with 80.95% accuracy (see Table 3). The ANN was considered successful if it rated a music excerpt within one standard deviation of the average human rating; in other words it came within 68% of the range of human responses (i.e., 32% of the humans were outside of this range). There are two possibilities for the decrease in accuracy of the ANN with regard to Berg: Either our metrics fail to capture some essential aspects of Berg's piece, or the other 11 pieces do not contain sufficient information to enable the interpretation of Berg's piece.

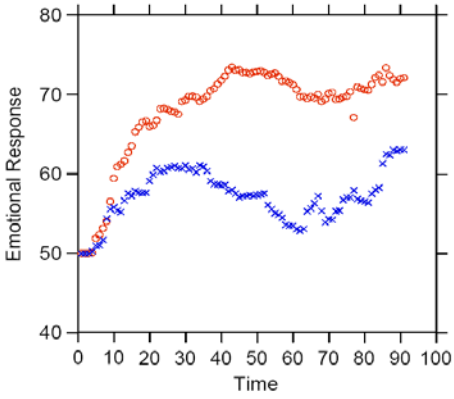

Fig. 2. The average pleasantness (o) and activation (x) ratings from 21 human subjects for the first 1:30 seconds of Mozart's "Clarinet Concerto in A" (K.622). A rating of 50 denotes neutral response

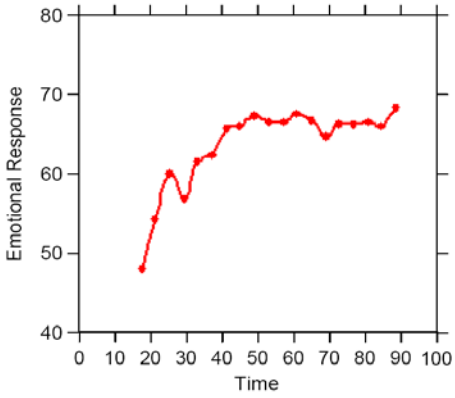

Fig. 3. Pleasantness classification by ANN of the same piece having been trained on the other 11 pieces

Fig. 2 displays the average human ratings for the excerpt from Mozart's "Clarinet Concerto in A" K.622. Fig. 3 shows the pleasantness ratings predicted by the ANN for the same piece. The ANN prediction approximates the average human response.

Additionally, we performed three control experiments to validate the results produced by the ANN. In specific, all values in the Human-training, Human-testing (HH) data were replaced by values generated using a uniform-distribution random number generator. These and the original values were then combined into three data sets for the control experiments: Random-training and Random-testing (RR), Random-training and Human-testing (RH), and Human-training and Random-testing (HR). Each of the control experiments was a complete 12-fold cross-validation study, just like the human data experiment.

Fig. 4 shows the Test MSE per piece across all four conditions (HH, RR, RH, and HR). The reader should recall that the first six pieces were pleasant and the last six unpleasant. Fig. 5 shows the average Test MSE across the four experiments.

3.4 Discussion

The ANN was able to discover strong correlations between the human pleasantness data and Zipf-based metrics (HH condition). Also, as expected, the ANN did not discover any correlations between random data and Zipf-based metrics (HR and RR conditions).

However, the ANN performed relatively well when trained against random data and tested against human data (RH condition). This may be surprising at first, however, it simply demonstrates the effect of *peeking at the test data* while training (see [15], p. 661) – as mentioned above, we used simulated annealing to "jog" the weights when the ANN appeared stuck in local minima relative to the test MSE. In other

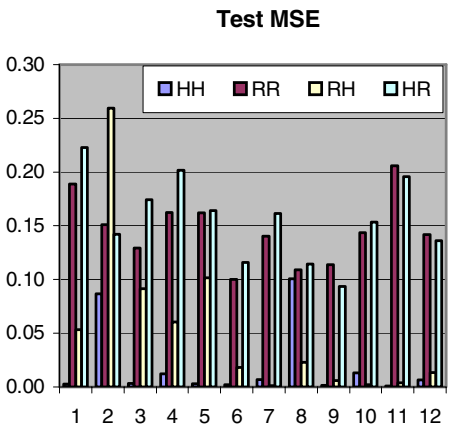

Fig. 4. ANN Test MSE for each piece across all conditions

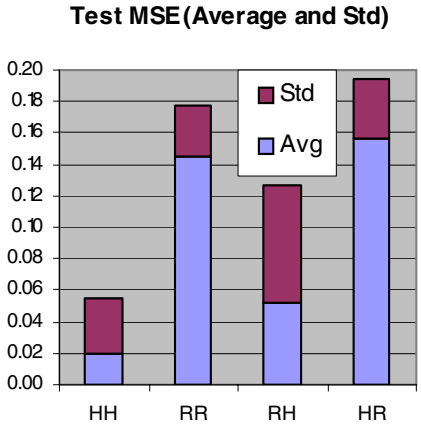

Fig. 5. Average Test MSE and standard deviation across all conditions

words, the ANN was trained to minimize both the test and the train MSEs. This indicates that the ANN is actually able to learn something about the human data, even though it was trained on random "noise.". While the ANN does succeed in classifying the data, its error rate is more than double than when it was trained with actual human data.

Reassuringly, this peeking effect produced no convergence in all 12 experiments of the RR condition (random training, random testing). This strongly suggests that there is a correlation between Zipf metrics and human pleasantness data, and no correlations with random data.

Analysis of the ANN weights associated with each metric suggests that *harmonic consonance* and *chromatic tone* were consistently relevant for "pleasantness" prediction, across all 12 experiments. Other relevant metrics include chromatic-tone distance, pitch duration, harmonic interval, harmonic and melodic interval, harmonic bigrams, and melodic bigrams.

The HH (and RH) results indicate that the ANN is able to identify patterns that are relevant to human reporting of pleasantness. The feature extractor and ANN evaluator used in this experiment can easily be incorporated into an evolutionary music system as part of fitness evaluation. Our results suggest that such a fitness function has strong potential to guide the evolutionary process towards music that sounds pleasant to humans. However, given the statistical nature of the metrics, we expect that additional structural, music-theoretic metrics may be required to discourage evolution from finding shortcuts – ways of creating false maxima. In other words, we suspect that ANN-based fitness functions, such as the one reported in the pleasantness study, at best, define a necessary but not sufficient (pre)condition for pleasant music. To evaluate this hypothesis, we are in the process of developing an evolutionary music system, called NevMusE, that will be used to generate music guided by such ANN-based "pleasantness" fitness functions.

4 Conclusions

The experimental results attained show that the considered set of metrics captures important music attributes, facilitating not only accurate prediction of author and style, but also pleasantness of musical pieces.

We propose that this approach may be applied successfully in the scope of a fully- or partially-automated system to assign fitness according to:

- compliance to a given musical style or styles;
- similarity to the works of some composer(s); and
- predicted pleasantness of the piece

There are several differences, and potential advantages over previous works dealing with the automation of fitness assignment. For instance, by using a set of well-known pieces instead of ones generated through IE, we ensure that the training set is unbiased towards the scores typically generated by the system. Also, the tasks of author and style identification do not involve subjective criteria. The output vector of the ANN can be seen as a set of distances to particular styles and authors, which opens new possibilities in terms of fitness assignment. Finally, the ANNs trained for predicting the pleasantness of pieces appear to capture fundamental principles of aesthetics. This contrasts with other approaches where the ANNs, when successful, capture only some of the preferences of an individual user.

Similarly to other approaches there is always the possibility of errors in classification and prediction. As such, using a totally automated system may result in convergence to false optimums. Taking into account the current state of development, we believe that it is probably wiser and more interesting to use a partially interactive system. The system would run on its own using the ANNs to assign fitness. However, the user can interfere at any point of the evolutionary run assigning fitness to the individuals, thus overriding the automatic evaluations.

We have already used this scheme in a partially interactive visual art evolutionary system [16]. The experimental results show that user intervention was enough to overcome the deficiencies of the fitness assignment scheme, which, in that case, where quite severe. Nevertheless, due to the generic properties of the extracted features, it is expected that, in the case of music, our approach results in more generic and robust fitness assignment.

Acknowledgements

This project has been partially supported by an internal grant from the College of Charleston and a donation from the Classical Music Archives. We thank Timothy Hirzel, Robert Davis and Walter Pharr for various comments and contributions. William Daugherty and Marisa Santos helped conduct the ANN experiments. Giovanni Garofalo helped collect human emotional response data for the ANN pleasantness experiment.

References

1. Burton, A. R., Vladimirova, T.: Applications of Genetic Techniques to Musical Composition. Computer Music Journal, Vol. 23, 4 (1999) 59-73
2. Spector, L., Alpern, A.: Induction an Recapitulation of Deep Musical Structure. IJCAI-95 Workshop on Artificial Intelligence and Music (1995) 41-48
3. Biles, J. A., Anderson, P. G., Loggi, L.W.: Neural Network Fitness Function for a Musical GA. International ICSC Symposium on Intelligent Industrial Automation (IIA'96) and Soft Computing (SOCO'96) (1996) B39-B44
4. Burton, A. R., Vladimirova T.: Genetic Algorithm Utilising Neural Network Fitness Evaluation for Musical Composition. 1997 International Conference on Artificial Neural Networks and Genetic Algorithms (1997) 220-224
5. Zipf, G.K.: Human Behavior and the Principle of Least Effort. Addison-Wesley Press, New York (1949)
6. Voss, R.F., and Clarke, J.: 1/f Noise in Music and Speech. Nature, Vol. 258 (1975) 317-318
7. Manaris, B., Vaughan, D., Wagner, C., Romero, J. and Davis, R.B.: Evolutionary Music and the Zipf–Mandelbrot Law – Progress towards Developing Fitness Functions for Pleasant Music. EvoMUSART2003 – 1st European Workshop on Evolutionary Music and Art, Essex, UK, LNCS 2611, Springer-Verlag (2003) 522-534
8. Manaris, B., Romero, J., Machado, P., Krehbiel, D., Hirzel, T., Pharr, W., and Davis, R.B.: Zipf's Law, Music Classification and Aesthetics. Computer Music Journal, Vol. 29, 1, MIT Press, Cambridge, MA (2005)
9. Classical Music Archives [online]: http://www.classicalarchives.com (2004).
10. Stuttgart Neural Network Simulator [online]: http://www-ra.informatik.uni-tuebingen.de/SNNS/ (2004)
11. Machado, P., Romero, J., Manaris, B., Santos, A., and Cardoso, A.: Power to the Critics - A Framework for the Development of Artificial Critics. Proceedings of 3rd Workshop on Creative Systems, 18th International Joint Conference on Artificial Intelligence (IJCAI 2003), Acapulco, Mexico (2003) 55-64
12. Machado, P., Romero, J., Santos, M.L., Cardoso, A., and Manaris, B.: Adaptive Critics for Evolutionary Artists. EvoMUSART2004 – 2nd European Workshop on Evolutionary Music and Art, Coimbra, Portugal, *Lecture Notes in Computer Science*, Applications of Evolutionary Computing, LNCS 3005, Springer-Verlag (2004) 437-446
13. Barrett, L.F., and J. A. Russell, J.A.: The Structure of Current Affect: Controversies and Emerging Consensus. Current Directions in Psychological Science, Vol. 8, 1 (1999) 10-14
14. Schubert, E.: Continuous Measurement of Self-report Emotional Response to Music. In Music and Emotion – Theory and Research, Juslin, P.N. and J.A. Sloboda (eds). Oxford University Press, Oxford, UK (2001) 393-414
15. Russell, S., and Norvig, P.: Artificial Intelligence – A Modern Approach, 2nd ed. Prentice Hall, Upper Saddle River, NJ (2003)
16. Machado, P., Cardoso, A.: All the Truth about NEvAr. Applied Intelligence, Vol. 16, 2 (2002) 101–119