The Role of Loudness in Detection of Surprising Events in Music Recordings.

Piotr Holonowicz,[*1] Perfecto Herrera,[*2] Hendrik Purwins[*3]

[*]*Music Technology Group, Universitat Pompeu Fabra, Spain*
[1]piotr.holonowicz@upf.edu, [2]perfecto.herrera@upf.edu, [3]hendrik.purwins@upf.edu

ABSTRACT

The abrupt change of loudness is a salient event that is not always expected by a music listener. Therefore loudness is an important cue when seeking for events in a music stream that could violate human expectations. The concept of expectation and surprise in music has become recently the subject of extensive research, however mostly using symbolic data. The aim of this work is to investigate the circumstances when a change of sound intensity could be surprising for a listener. Then, using this knowledge, we aim to build a computational model that analyzes an audio stream and points to potential violations of human expectation. In order to check the quality of human prediction, an online (web-based) experiment, based on a gambling paradigm, has been performed. The subjects, after listening to an excerpt from real recordings of classical music, were asked to bet on how loud would be the next continuing seconds. . The results of the survey have shown that the moments of sudden changes of intensity more likely caused mispredictions, however it is obvious that the loudness itself is not the only factor that determines human expectations.

I. INTRODUCTION

From a biological point of view, sound events that carry a high amount of physical energy usually mean a threat. If the source of the threat is unknown or the event occurs at the moment that is unexpected, the organism is surprised and chances of its survival drop. Since listening to music is rather pleasant activity for humans, the chances of occurrence of a danger during listening are low, thus we rather do not expect a warning coming from the music itself. Instead, we expect the threat to come more frequently from the situations where someone listens to music in an potentially unsafe environment eg. at a street. Therefore loud moments in a music passage seem to be unexpected and cause frisson (Huron, 2006).

Most of the research effort in computational modeling of music expectations has focused on symbolic processing (Temperley, 2007; Huron, 2006; Hazan et all., 2008a) Modeling based on music transcription has several advantages: high accuracy of the input data, easiness to operate on polyphonic music and possibility of using the knowledge gathered through centuries about structure of music or composition rules (Narmour, 1990), as well as the possibility of application of the recent research about the statistical properties of music (Hazan et all, 2008a; Saffran, 1999). However, it does not take into account all the features of music, like the loudness, the timbre, or the expressivity introduced by performers.

Analysis of an audio signal needs to be performed at much lower level and information extracted from the recordings is much less accurate. The analysis itself requires application of statistical techniques or machine learning (Hazan et all, 2008b; Dubnov, 2008), which otherwise have also been applied for processing music in symbolic format (Kostek, Wojcik & Holonowicz, 2005).

Modelling surprise in music listening has received very scarce attention, specially from researchers working in music audio processing. Fortunately, there are recent promising proposals that deserve to be mentioned here. Itti and Baldi (Itti&Baldi, 2005) have proposed a quantitative scale of surprise based on Bayesian learning. The surprise is definied as the cross entropy (Kullback-Leibler divergence) between prior and posterior probabilities of a Bayesian model:

$$S(D,M)=KL(P(M|D),P(M))$$

$$S(D,M)=\int_{M_i} P(M_i|D)\log\frac{P(M_i|D)}{P(M_i)}dM_i$$

where M means the model as a whole and D means the data introduced to the model. M_i are the variables (nodes in case of graphical representation) of the model. The formula above is general and depends on what is understood behind the model and the data. Another advantage is that it can be applied to wide range of models provided that the prior and posterior probabilities are definied by the model. The surprise is undefinied if either the prior or the posterior are zeros (meaning impossible events).

Another way of measuring surprise has been recently proposed by Abdallah (Abdallah & Plumbey, 2007) is the *Instantaneous Predictive Information Rate* (IPIR), definied as follows:

$$I(x|z)=KL(P(Y|X=x,Z=z),P(Y|Z=z))$$

The IPIR is equal to cross entropy between the predictive distributions over Y before and after the event $X=x$, given that we already know $Z=z$ and works rather as a guide to how much attention needs to be directed towards the next event even before it happens, than the measure of surprise itself. As a direct measure of "surprisingness" of event x given the context z, Abdallah proposes the negative log probability:

$$L(x|z)=-\log p_{X|Z}(x|z)$$

where $p_{X|Z}$ is the distribution of the immediate prediction (Abdallah & Plumbey, 2007). This measure, however, is an entropy-type one, which can be seen as problematic (see (Itti & Baldi, 2005; Itti & Baldi, 2006) for a full discussion).

The measures above require knowledge of the priors and the posteriors. Thus, they are applicable for modeling. It is, however, not possible to obtain the exact values of the

probabilities from humans. Instead, the estimators need to be inferred from the statistical analysis of correlated variables, measured in some empirical way. The experiments that aim at measuring the surprise can be divided in two groups – the ones that measure it implicitly and the ones that try to measure it explicitly. These methods treat surprise as a violation of human expectations, so the subjects are asked to perform various tasks and the degree of surprise is estimated by measuring their accuracy. With the development of medical diagnosis tools, the explicit methods of measurement of expectations, have appeared. They are based on human body response to a violation of expectations. The earliest measure basic physiologic features of human body, like the heart rate or the resistance of the skin, eg. bradycardic response method, while the recent ones measure the time of reaction (Aarden, 2003) or the Evoked Response Potential (ERP) (Winkler, & Näätänen,1992; Czigler,Weisz & Winkler, 2006). In the latter case, the surprise is estimated by the presence of a specific electrical signature of brain processing called Mismatch Negativity (MMN - for a detailed description see, e..g., Winkler, 2007).

Among all the mentioned ways to estimate the surprise, the gambling or betting paradigm (von Hippel, Huron & Harnish, 1998) is particularly worth of attention, because it returns not only the amount of surprise but also the certainity of a choice, which allows to estimate at least the posteriors. Although long and tedious for the participants to perform, the gambling makes them to keep their attention on the task and generate a high amount of input data. Till now, the betting paradigm has been used for research, for example, on the entropy of music (Witten, Manzara & Conklin, 1993) and to compare melodic expectations for two different cultural groups (von Hippel, Huron and Harnish, 1998).

In both cases, the subjects were supposed to bet on the consecutive notes of a melody. Also in both cases, to estimate the certaininty of the subject's choices, the entropy measure was used. Although the experiments based on gambling have their disadvantages, for example the effect of learning usually affects the results (so the experiment cannot be too long), they have been proven to be a reliable way to probe musical expectancies (Huron, 2006).

II. AIMS

The main purpose of the experiment described in this paper is to check if and when, with the typical listening conditions, sudden changes in music loudness point to a surprise. Another question that the experiment may help to answer, is whether humans remember the loudness of the music or not and what factors affect their expectation about the instantaneous energy of the music they hear. The final goal is providing the ground truth data for the evaluation of a Bayesian music-surprise detection model that is currently under development.

III. METHOD

A. The experiment

The experiment reported here addresses musical surprise using the gambling paradigm. At the beginning the subject has to give his/her nickname and answer questions about age, preferred genres and the music education. Then an intitial amount of virtual money is assigned to him/her (100 credits). Afterwards, the experiment is performed with the following algorithm:

1. *Let the subject calibrate the headphones using a reference tone - ask him/her to set maximum volume to his headset that does not yet cause audible distortions and is plausible. The subject is then instructed not to tamper with the volume afterwards.*
2. *Randomize the playlist*
3. *For each playlist item:*
    1. *start playing the introduction part and display how many seconds left to the end. When the playback enters the zone of special attention, notify the subject by a blinking of the display. The introduction is allowed to be played back only once.*
    2. *stop playing*
    3. *ask the subject to bet from 10 % to 90 % of his credits if the loudness of the next 3 seconds will change noticeably (at least 2 times ~= 10 dB, in both directions - upwards or downwards). There are 3 options: the loudness of the continuation may be Louder, Quieter or Not Change.*
    4. *right after making the bet, user is asked to enter the degree of familiarity with the excerpt. The levels of familiarity look as follows:*
       F1 - "I did have not heard it before"
       F2 - "the melody sounds familiar"
       F3 - "I know the author or the title"
       F4 - "I know both, the author and the title"
       F5 - "I can sing or hum the melody"
    5. *if the user has guessed the loudness of the continuation correctly, he earns the bet amount of credits he has bet. If not, he loses the amount he has bet.*
    6. *After being informed about the result of the bet, user can hear the continuation as many times as he/she wants.*
4. *At the end, when the subject has made bets on all the excerpts, the total score is shown.*

The game ends with the end of the last excerpt. Each time the account status of the participant drops below 1 credit, his account is recharged with 10 credits.

B. The conditions

A web-based experiment was decided in order to maximize the number of possible subjects. The participants were instructed to keep the volume unchanged during the experiment, and their honesty in obeying this, has been assumed. The subjects were also instructed to exclusively use headphones while participating, in order to minimize the influence of background noise. In this unmonitored conditions, although the subjects might "cheat" by tampering with the volume gauge of their amplifiers, it should be clear that the experiment trury measures the *relative* loudness between two fragments of music. For any methodological issues related to web-based experiments, see Honing et al., (2008)

C.The data

The data collection consists of 51 excerpts from classical music, stored as the digital recordings in the MP3 format (audio compressed at 128 kbps). Each one has up to 20 seconds, so total time of experiment should be around 30 minutes per participant. The collection is constructed according to the following guidelines / assumptions:

- the recordings are chosen to minimize probability of being familiar for the subjects.
- the recordings are chosen to contain at least one prominent change in the dynamics.
- The types are distributed uniformly so each type has equal number of representants. The collection finally contains 17 excerpts of each type.

The reason for using exclusively classical music is the fact that in opposite to the other genres that the dynamics has much higher impact on the listener experience thus the music is recorded the way the dynamics is preserved. That allows assuming that the instantaneous energy of the sound in a moment t is much closer to its perceptual loudness than in music recordings of other genres like for example pop, where the energy is kept more less on the same level but the mastering allows to hear some instruments louder than the others. The instantaneous energy of a sound is much easier to extract than any perceptual feature, which is later useful for introducing the data to the Bayesian model.

Each of the excerpts is labeled manually, independently by two experts, using the "Audacity" sound edition tool. The whole excerpt constists of two parts, the introduction and the continuation. The introduction has a label that points to the place where the subject should take special attention while listening (the zone of special attention), expressed as the time left to the end. The continuation has the label which says whether it is *Louder (**L**), Quieter (**Q**)* or *with the same loudness(**N**)* than the zone of special attention pointed in the introduction. By the *type of the excerpt,* the label of the continuation is considered.

D.Subjects

The time when the experiment was avaliable for public was relatively short (1 week). During that time, 33 participants, aged from 19 to 36 years, finished the survey. The length of their music education was diverse, as well as the genres they prefered to listen to, but only 4 of them admitted that they have no experience with playing any instrument. 13 persons declared that classical music belonged to their preferred genres, while the remaining 20 did not include it into their favourities. No additional information was gathered although, because of the way subjects were recruited, most of them were studying either music or computer science.

During the experiment, the following data are gathered:

- the participant's nickname, the age, the preferred genres of music (5 categories: classical, jazz, pop/hip hop,electronic, rock/metal) and the length of musical education.
- for each excerpt: the participant's answer, bet amount in percent then its absolute value in credits, current account status and the participant familiarity with the excerpt in scale 1 to 5.

IV. Results

A. Introduction

The outcome could be split in two groups, the data that described particpant's behavior and the data that described the participant itself. For both groups the most important variable was the number of mistakes (mispredictions) done by the participans, as it allows to estimate the level of surprise. A mistake is considered to be a lost bet (in other words, a disagreement between the participant's belief and the opinion of the experts who rated the excerpt). The primary question the experiment supposed to answer to, was if the type of excerpt influences the variable. But then it was necessary to estabilish if the other factors : participant's familiarity with the excerpt and the amount of the bet the participant has made (which shows the overview of the strategy of gambling for a particular subject) also affects the variable. Finally, the influence of particular human features on the variable has been investigated.

B. Answers and number of mistakes versus loudness of the continuation

The purpose of the analysis was to find the dependency between the type of the excerpt and the answers of the participants as well as the mistakes they have done. In order to find the dependency, one-way ANOVA has been applied, after checking the prequisites (see Table 1). The experiment design was balanced, so the number of samples is the same for each category (that is the number of participants = 33). The number of answers as well as the number mistakes were divided by the total number of excerpts to obtain normalized results. In both cases the variances of the populations were equal (checked with the Levene test for homoscedacity). The distribution of the residuals turned out to be normal, checked with the Shapiro-Wilk test for normality, at the significance level = 0.05. That also proves that the samples are independent. All the prequisites for ANOVA were fulfilled.

Table 1: ANOVA for subject answers vs. type.

	Degrees of freedom	Sum Sq	Mean Sq	F value
Loudness	2	2336,42	1168,21	67,09
Residuals	96	1671,58	17,41	H0 rejected

Table 2: ANOVA for subject mistakes vs. type.

	Degrees of freedom	Sum Sq	Mean Sq	F value
Loudness	2	2275,7	137,85	30,08
Residuals	96	429,21	4,47	H0 rejected

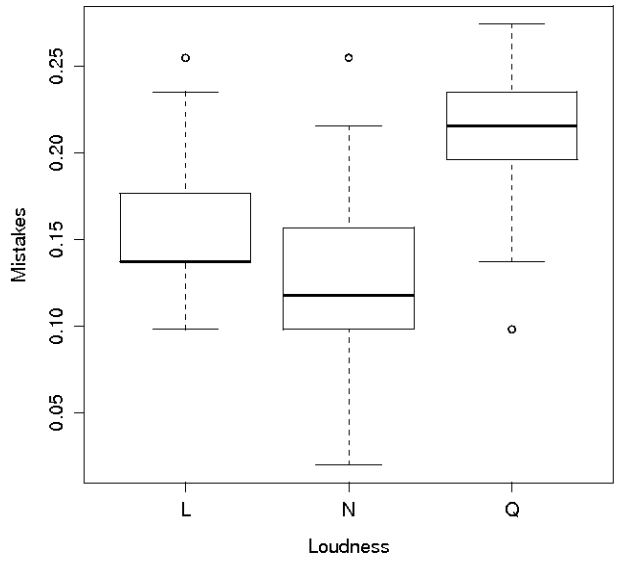

Figure 1: number of mistakes vs. excerpt type

The ANOVA indicates that the null hypothesis stating an equality of means for the samples can be rejected in both cases. For mistakes, the value of the normality test was on the border of α. Thus, additional Kruskal-Wallis procedure was applied to confirm ANOVA results. The null hypothesis was rejected (p<2.9e-09), the medians for the population are not equal, so the result of ANOVA was confirmed (see Table 2). Figure 1 graphically details the dependency found between mistakes and excerpt type. Whereas the difference in error rates for the types "Louder" and "No change" are negligible, the "More Quite" excerpts have generated a significantly higher amount of errors. It might happen because of two reasons: first it could be that the kind of ending of the introductory part presented some "closure" and actually the subject had no clue about what could happen next. However, the excerpts had miscellaneous endings, with no special preference for closures. The other reason, pictured at Figure 2 is the general preference of the subjects, who expected mostly that loudness would not change at all or, if ever, it would increase. So it seems that the excerpts with quieter continuations were more surprising than the others on average.

C.Mistakes versus familiarity

Here the impact of prior knowledge about the data has been checked. Since the goal was using stimuli that were as unfamiliar as possible, the expected outcome was a lack of dependency between familiarity and number of mistakes, which is reflected in Figure 3.

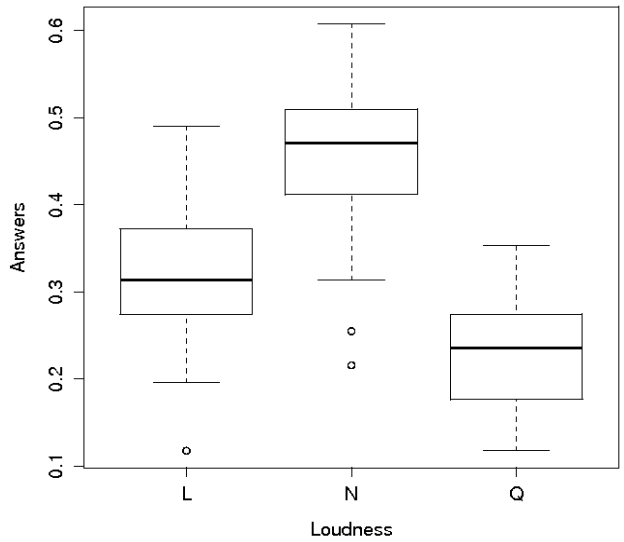

Figure 2: subject answers vs. excerpt type

The Shapiro-Wilk test and Levene test show that unfortunately both prequisites for ANOVA are not fulfilled. Although it is still possible to test the data using procedures for non-parametric, heteroscedastic distributions (eg. Welch, Brown-Forsythe), the plots show clearly that the most errors

have been done in the first group - “I haven't heard it before”. That happened because the most of the answers (not pictured here) belonged to that group, so in general the subjects were unfamiliar with the excerpts.

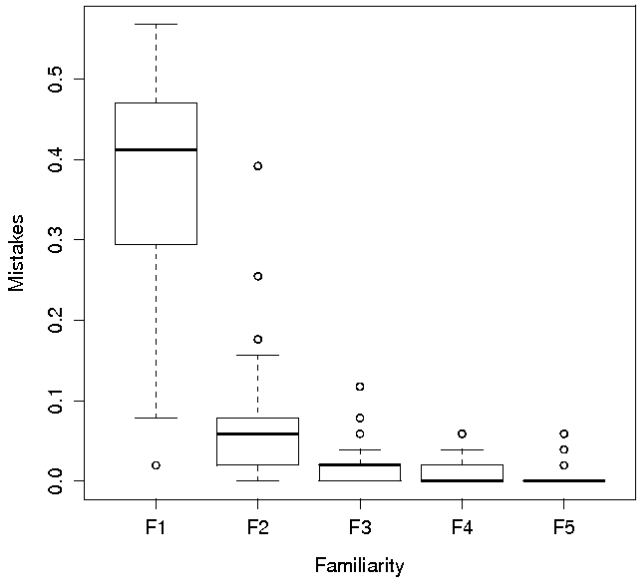

Figure 3: mistakes vs. familiarity

D. Mistakes versus bet amounts

The correlation between the amounts people bet and their choices is highly important for this experiment. The expected outcome of the analysis was then a clear dependency between the bet amount and the answers. Generally, the higher was bet, the lower number of mistakes should occur.

As with the case of familiarity, the analysis with ANOVA does not make sense since the prequisites are not met.

Unfortunately, Figure 4 shows that the dependency between a percentage of the bet and the number of mistakes is not monotonous. Basically it seems that the subjects have applied two gambling strategies : either to bet carefully (below 50 %) or to bet full amount allowed (90 %). In both cases the medians and the quartiles are almost equal, so both strategies caused similar number of errors. That means that the persons gambling more carefully were the winners, since they lost much less credits. The number of mistakes in the groups B60 to B80 is caused by general lack of bets of these amounts - if someone was fairly certain of the answer (he wanted to bet above 50%) he put the bet of 90 % in order to maximize the win. This result makes usage of the bet percents as an estimates of the certainity of the subjects choices, useless.

E. Number of mistakes vs. participant features.

Table 3:ANOVA for mistakes vs. age.

	Degrees of freedom	Sum Sq	Mean Sq	F value
Age	16	0,0837	0,0052	1,8924
Residuals	16	0,0442	0,0028	H0 accepted

As the ANOVA indicates, the null hypothesis that the means are equal is accepted. That basically means that the number of errors can be considered as independent from the participant age. Since the p-value of the normality test was low(*=0.1970*), the Kruskal-Wallis test has been additionaly applied in parallel to ANOVA, confirming it.

Figure 5 pictures the dependency of the participants errors on the experience and preferred genres. The experience value was collected as the number of years a participant played an instrument (“No” means 0 years of playing, while “Yes” means “at least 1 year of playing an instrument). For the preferred genres we included in the group “Classical” all the participants that marked “Classical” among the genres they preferred to listen to, and the remaining subjects were asigned to the group “Popular”. After standard check-up for the prequisites, the ANOVA was performed for both variables.

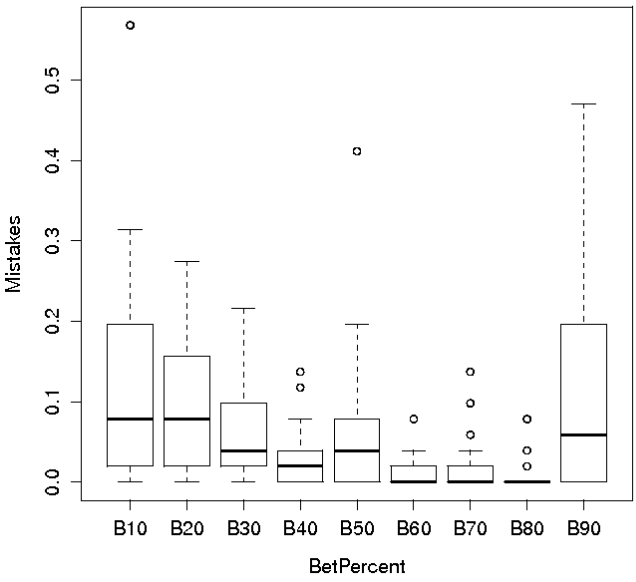

Figure 4: mistakes vs. percent of the bet.

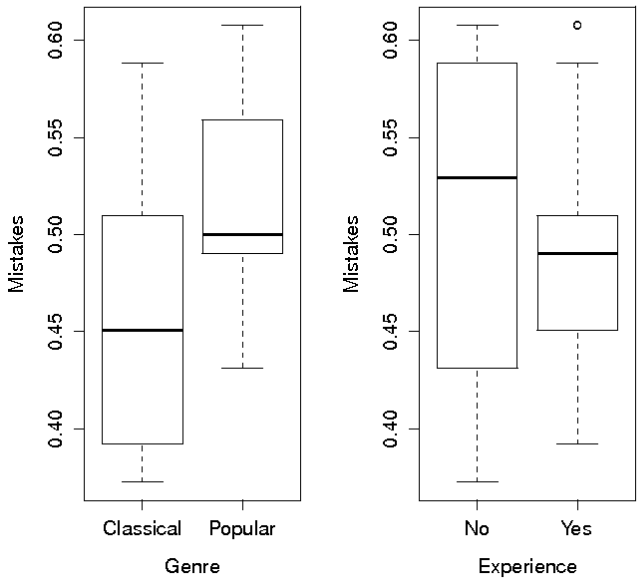

Figure 5: mistakes vs. preferred genre (left) and experience with playing an instrument (right)

Table 4: ANOVA for mistakes vs. experience.

	Degrees of freedom	**Sum Sq**	**Mean Sq**	**F value**
Experience	1	0,0014	0,0014	0,3545
Residuals	31	0,1265	0,0041	H0 accepted

From the ANOVA results we accept the equality of the error means for the experienced and unexperienced subjects, meaning that the number of errors does not depend on the music education (see Table 4). This result is somewhat surprising because it has been shown that the listening preferences, not playing experience, was what really ruled the generation of expectancies.

Table 5: ANOVA for mistakes vs. genre.

	Degrees of freedom	**Sum Sq**	**Mean Sq**	**F value**
Genre	1	0,0309	0,0309	9,8914
Residuals	31	0,0970	0,0031	H0 rejected

The ANOVA summarized in Table 5 led to reject the null hypothesis, meaning that the number of errors and the familiarity withclassical music are bound. At the Figure 5 this binding is visible: people that had classical music mentioned in their preferences performed about 5 % better than people that did not like this genre.

F. Analysis per excerpt.

The initial goal for this experiment was to provide the ground truth data for a Bayesian model capable of detecting music surprise. Two criteria had to be true simultaneously to consider an excerpt as surprise-generating :

1. a high rate of errors done by the subjects by betting for the excerpt (over 50 % of all the bets done for the excerpt)
2. the bets concentrated on a wrong answer (over 50 % of the bets was put on the wrong answer).

If both criteria were fulfilled, the excerpt was considered as *surprising*. If the distribution of the answers were close to the uniform, regardless of the error rate the excerpt was considered as *confusing*. After the analysis of the data 10 from 51 excerpts were annotated as surprising and another 10 from the 51 excerpts were annotated as confusing. Among the surprising ones, 5 were type “Q”, 2 of type “L” and 3 of type “N”, so the excerpts with the quieter continuation were more surprising than the excerpts from remaining groups. The distribution of confusing excerpts vs. type is uniform, which may suggest that for these excerpts the differences between loudness level were not salient enough. This problem will be further adressed in the future work.

V. Discussion and conclusions

The experiment has shown that the dynamics of music has aclear impact on human expectations in the case of classical music. The choice of the data allowed us to observe that a human exposed to an unknown piece of music does not expect changes of the loudness, at least not frequently. And if humans expect a change of loudness, they more probably expect an increase than a decrease. It might be inferred that people does not expect suddent drops of the loudness at all, since subjects performed nearly on a chance level in case of excerpts marked as “Quiet”. However, music loudness almost never drops without being first followed by an increase. As the average length of the introductory part was not exceeding 15 seconds, it is possible that the context was simply not long enough to make broader inferences about the loudness. Additionally, the perceptual loudness is more than the bare RMS energy of the audio signal (Zwicker & Fastl, 1999).

In the case of the experiment performed by Huron with American and Balinese musicians the entropy calulation based on the bets was used to estimate the certainity of the choices (Huron, 2006). However, our experiment has proven that the amounts of bet put by the participants strongly depend on their individual gambling strategies, thus the priors reflected by the certainity of the subject choices must be extracted another way.

Finally, the experiment has shown that the participants that like classical music perform better in guessing the loudness, which was a fairly difficult task for the people that does not prefer to listen to this genre. It seems that they judge probability of change of the energy to be a bit higher than the people that listens only to modern, popular genres.

All this informations allowed to construct a preliminary probabilistic model that is able to detect potentially surprising

moments in music. The results have also provided the ground truth data for its evaluation. Currently the model is fully developed and it will be described in a future paper.

VI. REFERENCES

Aarden, B. J. (2003). *Dynamic melodic expectancy*. PhD Dissertation, The Ohio State University.

Abdallah, S.A, Plumbley, M. D. (2007). *Information Dynamics.* Technical Report, Centre for Digital Music: Queen Mary University of London.

Czigler, I., Weisz, J., & Winkler, I. (2006). ERPs and deviance detection: Visual mismatch negativity to repeated visual stimuli. *Neuroscience Letters, 401,* 178-182.

Carlsen, J. C., Divenyi, P. L., Taylor, J. A. (1970). A preliminary study of perceptual expectancy in melodic configurations. *Council for Research in Music Education Bulletin*, *22*, 4-12.

Dubnov, S. (2008). Unified View of Prediction and Repetition Structure in Audio Signals With Application to Interest Point Detection. *Audio, Speech, and Language Processing, IEEE Transactions on*, *16*(2), 327-337.

Greenberg,G.Z., Larkin,W.D. (1968). Frequency-response characteristic of auditory observers detecting signals of a single frequency in noise: The probe-signal method. . *Journal of the Acoustical Society of America*, *44*(6), 1513-1523.

Hazan, A., Holonowicz, P., Salselas, I., Herrera, P., Purwins, H., Knast, A., et al. (2008). Modeling the Acquisition of Statistical Regularities in Tone Sequences. *30th Annual Meeting of the Cognitive Science Society*.

Hazan, A., Marxer, R., Brossier, P., Purwins, H., Herrera, P., & Serra, X. (accepted). What/when causal expectation modelling applied to audio signals. *Connection Science*.

Honing, H., & Ladinig, O. (2008). The potential of the Internet for music perception research: A comment on lab-based versus Web-based studies. *Empirical Musicology Review, 3 (1*), 4-7.

Huron, D. (2006). *Sweet anticipation: Music and the psychology of expectation*. Cambridge, MA: MIT Press.

Itti, L., Baldi, P. (2006) Bayesian surprise attracts human attention, *Proceedings of the Neural Information Processing Systems 2005*

Itti, L., Baldi, P. (2005). A Principled Approach to Detecting Surprising Events in Video, Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 631-637, Jun 2005.

Kostek, B., Wojcik, J., Holonowicz, P. (2005). Estimation the Rhythmic Salience of Sound with Association Rules and Neural Networks. *Proc. of the Intern. IIS: IIPWM*, 531–540.

Krumhansl, C, Shepard, R. N. (1979) Quantification of the hierarchy of tonal functions within a diatonic context. *Journal of Experimental Psychology: Human Perception and Performance 5(4): 579-594*

Manzara, L. C., Witten, I. H., & James, M. (1992). On the Entropy of Music: An Experiment with Bach Chorale Melodies. *Leonardo Music Journal*, *2*(1), 81-88.

Narmour, E., & Chicago, U. O. (1990). *The analysis and cognition of basic melodic structures: The implication-realization model*. University of Chicago Press.

Saffran, J., Johnson, E., Aslin, R., & Newport, E. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 52, 27.

Team, R. D. C. (2009). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. Retrieved from http://www.R-project.org.

Temperley, D. (2007). *Music and Probability* (s. 256). The MIT Press.

von Hippel, P., Huron,D., & Harninsh, D. (1998). Melodic expectation for a Balinese melody: a comparison of Balinese and American musicians. *Unpublished manuscript*.

Weiss, M.J, P.R. Zelazo, I.U. Swain (1988). Newborn response to auditory stimulus discrepancy. *Child Development 59, 530-541.*

Winkler, I. (2007). Interpreting the mismatch negativity (MMN). The *Journal of Psychophysiology*, *21(3–4),147–163.*

Winkler, I. & Näätänen, R. (1992). Event-related potentials in auditory backward recognition masking: a new way to study the neurophysiological basis of sensory memory in humans. *Neuroscience Letters, 140*, 239-242.

Witten, I. H., Manzara, L. C., & Conklin, D. (1994). Comparing Human and Computational Models of Music Prediction. *Computer Music Journal*, *18*(1), 70-80. doi: 10.2307/3680523.

Zwicker, E., & Fastl, H. (1999). *Psycho-acoustics Facts and Models.* Berlin Heidelberg New York: Springer-Verlag.