AUTOMATIC RECOGNITION OF EVENTFULNESS AND PLEASANTNESS OF SOUNDSCAPE

Jianyu Fan
Simon Fraser University, SIAT
250-13450 102 Avenue
Surrey, Canada
jianyuf@sfu.ca

Miles Thorogood
Simon Fraser University, SIAT
250-13450 102 Avenue
Surrey, Canada
mthorogo@sfu.ca

Bernhard E. Riecke
Simon Fraser University, SIAT
250-13450 102 Avenue
Surrey, Canada
ber1@sfu.ca

Philippe Pasquier
Simon Fraser University, SIAT
250-13450 102 Avenue
Surrey, Canada
pasquier@sfu.ca

ABSTRACT
A soundscape is the sound environment perceived by a given listener at a given time and space. An automatic soundscape affect recognition system will be beneficial for composers, sound designers, and audio researchers. Previous work on an automatic soundscape affect recognition system has demonstrated the effectiveness of predicting valence and arousal on responses from one expert user. Thus, further validations of multi-users' data are necessary for testing the generalizability of the system. We generated a gold standard by averaging responses from people provided people agreed with each other enough. Here, we model a set of common audio features extracted from a corpus of 120 soundscape recording samples that were labeled for valence and arousal in an online study with human subjects. The contribution of this manuscript is threefold: (1) study the inter-rater agreement showing the high level agreement between participants' responses regarding valence and arousal, (2) train stepwise linear regression models with the average responses of participants for soundscape affect recognition, which obtains better results than the previous study, (3) test the correlation between the level of pleasantness and the level of eventfulness based upon the gold standard.

Categories and Subject Descriptors
H.5.5 [Sound and Music Computing]: Modeling

General Terms
Human Factors; Design, Measurement. Performance

Keywords
Soundscape Affect Recognition, Gold Standard

1. INTRODUCTION
A soundscape is all the sounds in an environment perceived by a given listener at a given time and space. Sound design, soundscape composition, and urban design research [1-4] have demonstrated the variety of approaches taken to investigate how soundscapes affect people for the creation of immersive experiences. These literatures show that mood is a significant characteristic of human perception of a soundscape. Our research aims towards an automatic soundscape affect recognition system with which soundscape composers can use to create emotional soundscape compositions to evoke a target mood. Sound designers will find it more streamlined workflow to add suitable sound effects for films. Engineers can design mood enabled recommendation systems for retrieval of soundscape recordings.

In this paper, we explore whether soundscape recordings evoke the same emotion to different listeners. Specifically, our study tested the effectiveness of a soundscape affect recognition system with data from multiple users. Furthermore, an analysis of the agreement between user ratings demonstrates a generalization of the model. Finally, we present the correlation between arousal and valence based upon the gold standard.

This paper is organized as follows. In Section 2, we cover related works that form the basis of the emotional model and prediction algorithm. Section 3 details the methods, including, dataset, audio features, online study, machine learning models, and implementations. Section 4 describes the evaluation results. Finally, we present our conclusion and future work in Section 5.

2. RELATED WORKS
Automatic music emotion recognition is an open problem [5]. The annual Music Information Research Evaluation eXchange (MIREX) is a community-based framework for formally evaluating music-IR systems and algorithms, which includes music mood recognition as a task for the first time in 2007 [5]. The MIREX tasks are limited to musical content, which is a tiny proportion of all possible soundscapes.

Two rigorously studied models of emotion are the discrete and dimensional models. The discrete model classifies a document into one of a finite number of categories [13]. On the other hand, the dimensional model, as proposed by Russell [10], is a continuous circumplex space of emotional attributes such as pleasantness and eventfulness. Research in the MIREX community has studied emotional ratings of music with both these models. For example, Eerola et al. [6] report on regression analysis of both a circumplex and discrete model for modeling music emotion responses. Similarly, Lu, Liu, and Zhang [7]

AM15, October 07-09, 2015, Thessaloniki, Greece
 ISBN 978-1-4503-3896-7/15/10$15.00
DOI: http://dx.doi.org/10.1145/2814895.2814927

studied mood detection based on a valence-arousal circumplex model.

The soundscape literature describes a similar methodological approach for eliciting and modeling emotion responses to auditory stimuli. One such case is in Berglund et al. [1], who describe a listener survey to ascertain the emotional attributes important for soundscape perception. Their results show that the first and the second principal component are pleasantness and eventfulness. Their metric space, defined by the following axes: pleasant-unpleasant, exciting-boring, eventful-uneventful, and chaotic-tranquil, successfully measured soundscape quality. Likewise Brocolini et al. [2] conducted a field survey for studying the relationship between sound pleasantness and auditory features. Their study demonstrated the acoustic scene has a significant contribution to one's evaluation of pleasantness. Therefore, it is possible to analyze soundscape apart from the landscape and visual aspect of the scene.

In our study, we use a circumplex model of emotion mapped to an affect grid (Figure 1) for acquiring user response data to a soundscape. The affect grid is a measurement instrument for acquiring continuous data to measure subject's reaction to stimulus. Our work is based on a previous study of the Impress system [3], which was designed specifically for automatic prediction of soundscape valence and arousal in real-time environments. In that system, a corpus of audio files is generated using an automatic segmentation algorithm [9][15] that searches the online Freesound [24] database for audio regions with consistent soundscape characteristic greater or equal to a specified duration. The system models audio features and expert user responses to soundscape recordings with multiple linear regression models. Evaluation of the model showed a good fit of features to responses of models of predicting valence (R^2 : 0.712) and arousal (R^2 : 0.71). The details are given in Section 4.

In our research, we explored the same method for modeling audio features and users' emotion responses. We extend that research [3] for generalizing soundscape emotion recognition to a larger group of users. Furthermore, we wish to discover how well the model performs on predicting valence and arousal with multiple subjects.

3. METHODS

3.1 Valence and Arousal Model

Valence and Arousal model has been widely used in psychology studies [8]. Valence represents the pleasantness of a stimulus, which in our case is used to report the perceived pleasantness of a soundscape. Arousal indicates the intensity of emotion provoked by a stimulus. To easily provide explicit affective ratings on our valence and arousal model, we made a circumplex ordering of affect by using axes separated by 45 degrees: pleasant-unpleasant, exciting- boring, eventful-uneventful, and chaotic-quiet (Figure 1).

3.2 Collection Stage

3.2.1 Corpus

According to Schafer, "sounds of the environment have referential meaning" [14]. Based on referential meanings, Schafer built six categories.

- Natural sounds: bird, chicken, rain, sea shore;
- Human sounds: laugh, whisper, shouts, talk, cough;
- Sounds and society: party, concert, grocery store;
- Mechanical sounds: engine, cars, air conditioner;
- Quiet and silence: wild space, silent forest;
- Sound as indicators: clock, doorbell, siren;

We selected audio clips following six categories according to Schafers taxonomy. As for the database, instead of using Freesound database [24], we used Sound Ideas corpus [11] and World Soundscape Project [12], which have consistently good audio recording quality.

Sound Ideas is the distributor of the largest available sound effect libraries [13] ". It is the world's leading publisher of professional sound effects, offering more than 272 distinct royalty-free collections to broadcast, post production and multimedia facilities".

The World Soundscape Project, founded by Murray Shafer in the late 1960s, initiated the modern study of Acoustic Ecology [14]. According to him, this project is to "find solutions for an ecologically balanced soundscape where the relationship between the human community and its sonic environment is in harmony". These corpuses include various audio clips, such as human talking, bell, footsteps and vehicle driving. The World Soundscape Project was digitalized by Barry Truax and Metacreation Lab[1] at Simon Fraser University.

After the preliminary testing, we decided to extract 6-second excerpts from each of the sounds we selected, which would leave enough time for participants to form an opinion of the valence and arousal for a soundscape. Each clip is monophonic. The sample rate is 44100 Hz. Regions were selected based on the length of the consistency of the audio characteristics using a segmentation algorithm by Thorogood and Pasquier [9][15].

3.2.2 Online Study

Participants were 20 students who are taking a sound design class from Simon Fraser University. There are 12 males and 8 females. The average age is 21.7. To adjust for any learning effects and allow users to calibrate their answers, we omitted the affective rating of the first five clips. Thus, we had 125 audio clips rated by each user in random order. The study was conducted online through a web browser[2]. Excerpts are played through an HTML5 audio player object, which allows participants to listen repeatedly to an excerpt. After a user had listened to an audio clip; the user used a mouse to click on the affective grid to enter their response (the interface is shown in Figure 3). Soundscape affects are represented by values of valence and arousal. The x-axis represents the level of valence. For instance, the highest value of the valence, pleasant, is shown on the right side of the figure. The value of arousal is represented along the y-axis.

3.3 Data Analysis

3.3.1 Agreements between Participants

We used Intraclass Correlation Coefficient to present the reliability of measurements of ratings in both valence and arousal. It describes how strongly individual in a group resembles each other. In our case, both the valence index, with a 95% confidence interval of 0.866 to 0.915 and the arousal index, with a 95% confidence interval of 0.903 to 0.943, suggests that the ratings

[1] http://metacreation.net/

provided are reliable enough to use for building models. The higher index for arousal suggests that it is easier for observers to agree on the level of eventfulness than the level of pleasantness.

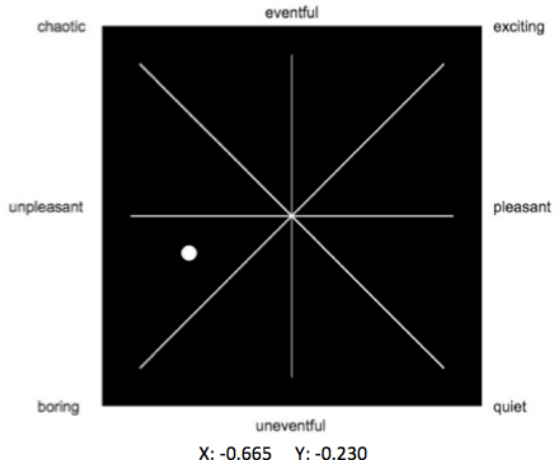

Figure 1. Online Study Interface: the X-axis represents valence (pleasantness). The Y-axis represents arousal (eventfulness). Users simultaneously evaluate on both dimensions by clicking their cursor on the grid, as illustrated by the white dot.

Because of the high agreements on affective responses towards soundscapes, we can build machine-learning models to predict the level of pleasantness and level of eventfulness. Given that this is a supervised machine-learning task, it is necessary to choose gold standard for the labels assigned to samples in the training set. We obtained the gold standard by averaging responses provided by 20 participants.

3.3.2 Audio Features for Modeling Soundscape

Based on the previous study [3], our audio feature vector contains Total loudness, Perceptual spread, Perceptual sharpness, Energy, Spectral Flatness, Spectral Flux, Spectral Roll-off, Spectral Slope, Spectral Variation and 40 MFCCs calculated using the BOF approach, which results in an 98 dimensions feature vector.

Total loudness is a feature that describes the psychological correlate of physical strength, (i.e., the sensation of intensity). The perception of loudness differed depending on the frequency of the sound [17]. Zwicker utilized principles such as equal loudness contours, critical band theory and the effect of sound fields. He considered the difference of perceptual loudness associated with each band along the Bark scale. The total loudness is the sum of the individual loudness from all bands [16].

The distance between the highest loudness value along the Bark scale and the total loudness is called perceptual spread. The perceptual equivalent to the spectral centroid but computed using the specific loudness of Bark bands is known as perceptual sharpness [17]. Energy is computed as root mean square of an audio Frame [17].

Spectral flatness is computed by using the ratio between geometric and arithmetic mean [17]. Spectral flux is the flux of the spectrum between consecutive frames [17]. Spectral roll-off is the frequency so that 99% of the energy is contained below [17]. Spectral slope is computed by linear regression of the spectral amplitude [17]. Spectral variation is the normalized correlation of the spectrum between consecutive frames [17].

Mel-frequency cepstrum coefficients (MFCCs) are common features in speech recognition systems recognizing people from their voices [18]. They have also been used in timbre recognition [19]. MFCCs are short-term spectral-based audio features. Mel-frequency is based upon the human auditory system, which does not have a linear perception of sound and maps different frequencies to perceived pitches.

Based on the previous study [3], we extracted audio features on sound recording regions formatted in AIF at a sample rate of 22500 Hz. Each excerpt is 6 seconds. Audio features were extracted with a 23ms Hanning window and a step size of 11.5 ms. The sliding window results in frames of 512 samples. Considering the similar processing methods of audio clips, we utilized the bag of frames technique [20], which considers frames that represent a signal having possibly different values.

3.3.3 Stepwise Multiple Linear Regression Models

In standard multiple linear regression models, all independent variables are used for analyzing at the same time. It straightforwardly explained the relationship between audio feature vectors and affect rating.

$$Y' = A + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \quad (1)$$

Where Y' is the predicted response, $X_{1\ldots k}$, are the predictor variables, A is the value of Y' when all $X_{1\ldots k}$ are equal 0, $\beta_{1\ldots k}$ are the regression coefficients. We built separate models for predicting valence and arousal. The stepwise multiple linear regression models take advantage of the stepwise method, which identifies the major predictors that influence the dependent variable.

Because of the effectiveness of the model in the previous work, we decided to use the same machine-learning model, a both-way stepwise multiple linear regression model, which combines the standard multiple linear regression model with stepwise selection methods. It selects the most effective set of predictors for fitting the pleasantness and eventfulness models to predict both valence and arousal values based on predictor variables, including audio features discussed in the previous section.

The forward stepwise multiple linear regressions begin with no variables. Iteratively, it selects the variable that increases R^2 the most. When none of the remaining variables are significant, the model will not add new variables. The backward stepwise multiple linear regressions begin with all variables. Iteratively, it removes the least significant variable. When there are no nonsignificant variables remain, the model will not remove variables.

Both-way stepwise multiple linear regressions combine above two stepwise models. It adds new variables. After each step, all variables in the model are checked to see if their significance has been reduced below the specified threshold. The model will remove nonsignificant variables, which also solves the problem of collinearity. Therefore, our model identifies the major predictors that influencing the dependent variable.

3.4 Implementation

We used SoX [21], a command line utility to convert audio formats from Wav to AIF. As for audio segmentation, we used the pydub library [22]. Audio features were extracted using the YAAFE [17] software package. We used Weka, a machine-learning tool [23] for training and testing the model. Other data analyses were implemented in Python.

4. RESULTS AND EVALUATION

In this section, we first present results of individual models of each participant, which is trained based on each participant's data. Then, we present the results of gold standard model, which is trained based on the average responses of 20 participants.

4.1 Evaluation Approach

We use the coefficient of determination (R^2) to evaluate the performance of our models. R^2 has been widely used for indicating how well the data fit a linear regression model. It represents the amount of variability explained by the regressors in the model. A R^2 is defined in (2), with f_i being the prediction made by the model and $\bar{y}$ being the mean of the ratings. R^2 describes the ratio of the variance of the model's predictions to the total variance. Results of the previous *Impress* system are used to compare with our results.

$$R^2 = 1 - \frac{SS_{res}}{SSSS_{tot}}$$

$$SS_{res} = \sum_i (y_i - f_i)^2 \quad (2)$$

$$SS_{tot} = \sum_i (y_i - \bar{y})^2$$

We used 10-fold cross validation and calculated the mean square error (MSE) to evaluate the prediction accuracy of the linear regressions. An MSE is defined in (3), where the $\hat{y}_i$ is the prediction made by the system and y_i is the user's rating. It indicates the error of the model in making prediction on unseen data.

$$MSE = \frac{1}{N}\sum_{i=1}^{n} (\hat{y} - y_i)^2 \quad (3)$$

4.2 Gold Standard Model

We built a gold standard model by training the model with the gold standard data, which is the average response of 20 participants. We used 10-fold cross validation to test our model of using six categories described in section 3.2. In addition, we used 10-fold cross validation to test our model without the category of "sound as indicators" in order to study the influence of semantic information on perceived pleasantness and eventfulness evoked by soundscapes. Furthermore, we tested our model by only using data from "sound as indicators", "natural sound", and "mechanical sound" individually. Due to the collection of excerpts of "sound and society", "quiet and silence", and "human sounds" having less than 20 items, these three categories were not tested separately.

Table 1. Results of Predicting Pleasantness Using the Gold Standard Model

Used Categories	R^2
All Six Categories	0.567
Without Sound as Indicators	0.715
Only Sound as Indicators	0.402
Only Natural Sound	0.989
Only Mechanical Sound	0.860

Table 2. Results of Predicting Eventfulness Using the Gold Standard Model

Used Categories	R^2
All Six Categories	0.816
Without Sound as Indicators	0.876
Only Sound as Indicators	0.737
Only Natural Sound	0.800
Only Mechanical Sound	0.983

The results of predicting pleasantness using the gold standard model are shown in Table 1. When we use all six categories, the R^2 for predicting pleasantness is 0.567, which indicates the model explains 56.7% of the variance of the data. Features that were identified as good predictors include mean of Total loudness, stdDev of Perceptual Sharpness, stdDev of MFCC5, mean of MFCC18, mean of MFCC32, and mean of MFCC23.

When we remove the category of "sound as indicators", however, the results of predicting pleasantness is 0.715, which indicates the audio features vector explained 71.5% of the variance ($R^2 = .715$, $F(7, 72) = 29.35$, $p < 0.001$). We assume semantic information plays an important role in evoking pleasantness of listeners, which improve the performance in our case. The low R^2 (.402) from only using data from "sound as indicators" also supports our assumption. Table 1 also shows us that when only using "mechanical sound" or "natural sound", our model obtains great prediction results of predicting pleasantness.

Table 2 shows the results of our model of predicting eventfulness. The R^2 of predicting eventfulness using all six categories is 0.816. Features that were identified as good predictors include mean of Total loudness, stdDev of Total loudness, mean of Energy, stdDev of Spectral Flatness, mean of Spectral Roll-off, mean of Spectral Variation, mean of MFCC2, mean of MFCC28, stdDev of MFCC5, and stdDev of MFCC26.

As for using five categories without "sound as indicators", the R^2 is 0.876, which indicates the features vector explained 87.6% of the variance ($R^2 = .876$, $F(12, 67) = 47.634$, $p < 0.001$). When we only tested "sound as indicators", the R^2 decreases to 0.737. Only using "mechanical sound" or "natural sound" obtains great predicting prediction results.

In general, the performance of the gold standard model for predicting eventfulness ($R^2 = .816$) is better than the one for predicting pleasantness ($R^2 = .567$). It is most likely because the energy is more easily differentiable than the emotion. This result also echoes the results for the Intraclass Correlation Coefficient of valence and arousal described in section 3.3.1. The category of "sound as indicators" influent on the general performance because of the semantic information. Nonetheless, this influence is not reflected by eventfulness as much as pleasantness, which explains a weak relationship between semantic information with energy but a strong relationship between semantic information with emotion. Our gold standard model performs better than the expert user's results in [3]. The application of this study can be founded online[2].

[2] http://142.58.183.142/impress

4.3 Performance of the Model of Individual Participants

We used the same model and correlated it with each participant's ratings individually. In this section, we present the performance of the model of individual participants. Previous section shows the improvement of the performance when not including "sound as indicators". Thus, we removed audio clips classified as "sound as indicators", which involves semantic information.

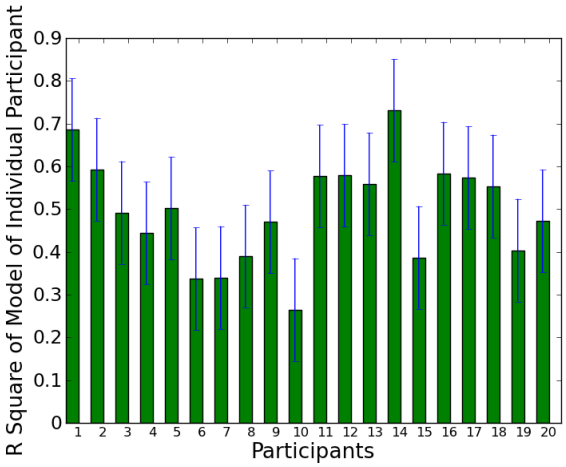

Figure 2. R^2 of individual participant's model for predicting valence, the error bars represent the standard deviation

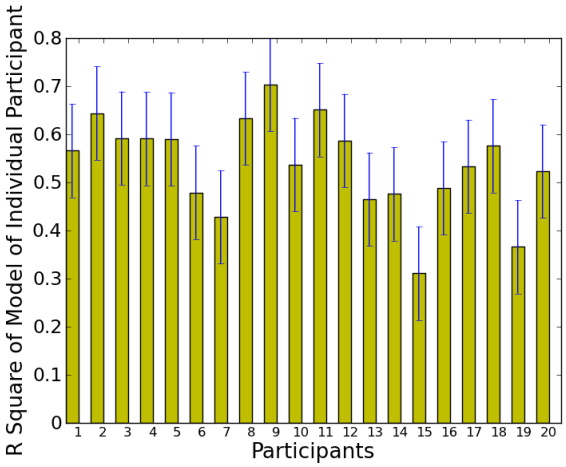

Figure 3. R^2 of individual participant's model for predicting arousal, the error bars represent the standard deviation

recognition system with multiple user data. Figure 2 shows the R^2 of all 20 participants models that predicting the pleasantness. Our average R^2 is 0.467, suggesting that on average; our models explain approximately 46.7% of the variability in the ratings of valence. (Mean = 0.467, stdDev = 0.132).

Figure 3 shows the R^2 of 20 participants models that predict ratings of arousal. Our average R^2 is 0.512, suggesting that on average; our models explain approximately 51.2% of the variability in the ratings of arousal. (Mean = 0.512, stdDev = 0.106). The individual models produce an average mean squared error (MSE) of 0.182 for valence and 0.129 for arousal.

Figure 2 and Figure 3 shows that the performance of the model of individual participant's is not as goods as the model in the previous study, which was initially trained based on an expert's data [3]. However, our golden standard model described in section 4.2 performs better than the expert user's results in [3] and better than any of the individuals.

4.4 Correlation between Valence and Arousal

We run a Pearson correlation test on average value over 20 participants' responses of pleasantness and eventfulness. There are 120 data points distributed in the affect grid (Figure 4). Each data point represents a value of pleasantness and a value of eventfulness ranging from -1 to 1.

Pearson correlation coefficient measures the linear correlation between pleasantness and eventfulness, giving a value between +1 and -1 inclusive, where 1 is total positive correlation, 0 is no correlation, and -1 is total negative correlation. Our Pearson correlation coefficient is -0.453, which means there is medium negative correlation between the two dimensions. This indicates that sounds that were rated as having higher arousal were rated as having lower valence. We assume this is because human listeners think a quiet and peaceful soundscape are more pleasurable.

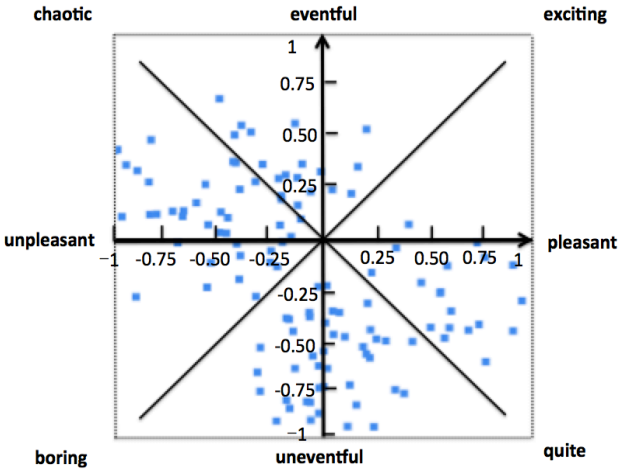

Figure 4. Correlation Area in the Affect Grid

To further test where the negative correlation exists in an affect grid (Figure 4), we run Pearson correlation tests on values of pleasantness and eventfulness of data points within the affect grid. The result is shown in Figure 4. This correlation indicates that pleasantness and eventfulness within the affect grid are not independent to each other.

5. CONCLUSION AND FUTURE WORK

We conducted an online study to obtain ratings from participants, evaluated the agreements between user affect ratings, built a machine learning model based on gold standard data and tested the performance of our automatic soundscape affect recognition system. We demonstrated the high-level agreement between participants' responses regarding two quality dimensions, valence, and arousal. This facilitates the investigation of the generalizability of categories of soundscape affects. Moreover, we created a better prediction model using gold standard data. This model performed better than the expert user model and any of the individuals. Finally, we tested the correlation between responses of pleasantness and eventfulness using gold standard data. This

study will benefit researchers in various fields, including sound studies, psychology, composition, and information retrieval.

For the next stage, we will test whether this model can be used for musical affect recognition and for different genres of music. Also, we plan to study the performance of this model for people with different cultural backgrounds.

6. REFERENCES

[1] Berglund, B., Nilsson. M. and Axelsson. O.2007. Soundscape Psychophysics in Place, *In Proceedings of the 36th International Congress and Exhibition on Noise Control Engineering,* page 3704–3712, Istanbul, Turkey.

[2] Brocolini, L., Waks, L., Lavandier. C., Marquis-Favre. C., Quoy, M. and Lavandier. M. 2010. Comparison between Multiple Linear Regressions and Artificial Neural Net works to Predict Urban Sound Quality, In *Proceedings of the 20th International Congress on Acoustics*, page 2121–2126, Nates, France.

[3] Thorogood, M. and Pasquier, P. 2013. Impress: A machine learning approach to soundscape affect classification for a music performance environment. In *Proceedings of the International Conference on New Interfaces for Musical Expression,* page 256–260, Daejeon, Republic of Korea.

[4] Schafer, M. 1997. Our Sonic Environment and the Soundscape: The Tuning of the World. *Destiny Books*.

[5] Kim, Y. E., Schmidt. E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., Speck, J. A. and Turnbull. D. 2010. Music Emotion Recognition: A State of the Art Review. In *Proceedings of the 11th International Symposium on Music Information Retrieval*, page 255–266, Utrecht, The Netherlands.

[6] Eerola, T., Lartillot, O. and Toiviainen, P. 2010. Prediction of Multidimensional Emotional Ratings in Music From Audio Suing Multivariate Regression Models, In *Proceedings of the 11th International Symposium on Music Information Retrieval*, Utrecht, The Netherlands.

[7] Lu, L., Liu, D. and Zhang. H. J. 2006. Automatic Mood Detection and Tracking of Music Audio Signals, *IEEE Transactions on Audio, Speech and Language Processing*, 14.1, 5-18.

[8] Russell, J. A., Weiss, A. and Mendelsohn, G. A. 1989. Affect Grid: A Single-Item Scale of Pleasure and Arousal, *Journal of Personality and Social Psychology*, 57.3, 493-502.

[9] Thorogood, M. and Pasquier, P. 2013. Computationally Generated Soundscapes with Audio Metaphor, In *Proceedings of the 4th International Conference on Computational Creativity*, page 1-7, Sydney, Australia.

[10] Russell, J. A. 1980. A circumplex model of affect, *Journal of Personality and Social Psychology*, 39.3. 1161-1178.

[11] Sound Ideas, Available online at http://www.sound-ideas.com/ , visited on Oct 22th 2014.

[12] World Soundscape Project, Available online at http://www.sfu.ca/ truax/wsp.html, visited on Oct 22th 2014

[13] Eerola, T. and K. Vuoskoski. 2011. A Comparison of the Discrete and Dimensional Models of Emotion in Music, *Psychology of Music* 39, 18-49.

[14] Schafer, M. 1977. The Tuning of the World, *Random House Inc*.

[15] Thorogood, M., Fan, J. and Pasquier, P. 2015. BF-Classifier: Background/Foreground Classification and Segmentation of Soundscape Recordings. *In Proceedings of the 10th Audio Mostly*, Thessaloniki, Greece.

[16] Zwicker, E. 1961. Subdivision of the Audible Frequency Range into Critical Bands, *The Journal of the Acoustical Society of America*, 33.2, 248.

[17] Mathieu, B., Essid, S., Fillon, T., Prado, J. and Richard, G. 2010.Yaafe, an Easy to Use and Efficient Audio Feature Extraction Software, In *Proceedings of the 11th International Symposium on Music Information Retrieval*. page 441-446, Utrecht, The Netherlands.

[18] Ganchev, T., Fakotakis, N., and Kokkinakis, G. 2005. Comparative evaluation of various MFCC implementations on the speaker verification task. In *Proceedings of the 10th International Conference on Speech and Computer*, page 191-194, Patras, Greece.

[19] Logan, B. 2000. Mel Frequency Cepstral Coefficients for Music Modeling, In *Proceedings of the 1st International Symposium on Music Information Retrieval*.

[20] Aucouturier, J. J. and Defreville, B. 2007. Sounds Like a Park : A Computational Technique to Recognize Soundscapes Holistically, Without Source Identification. In *Proceedings* of the *10th International Congress on Acoustics*., Madrid, Spain.

[21] SoX. Available online at http://sox.sourceforge.net/; visited on Oct 22th 2014

[22] Pydub. Available online at https://github.com/jiaaro/pydub; visited on Oct 22th 2014

[23] Berglund, B., Nilsson, M. E. and Axelsson, O. 2009. The WEKA Data Mining Software: An Update, SIGKDD Explorations, Newsl 11.1, 10-18.

[24] Freesound. Available online at http://www.freesound.org/; visited on October 22th 2014.