Impress: A Machine Learning Approach to Soundscape Affect Classification for a Music Performance Environment

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

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

ABSTRACT

Soundscape composition in improvisation and performance contexts involves many processes that can become overwhelming for a performer, impacting on the quality of the composition. One important task is evaluating the mood of a composition for evoking accurate associations and memories of a soundscape. We present a new system called *Impress* that uses supervised machine learning for the acquisition and realtime feedback of soundscape affect. We used an audio features vector of audio descriptors to represent an audio signal for fitting multiple regression models to predict soundscape affect. A model of soundscape affect is created by users entering evaluations of audio environments using a mobile device. The same device then provides feedback to the user of the predicted mood of other audio environments. The evaluation of the *Impress* system suggests the tool is effective in predicting soundscape affect.

Keywords

soundscape, performance, machine learning, audio features, affect grid

1. INTRODUCTION

Soundscape composition is the creative practice of processing and combining sound recordings to evoke listeners associations and memories of audio environments. Composers make decisions on segmenting environmental sound recordings, arranging segments in temporal and spectral domains, and applying techniques to process the recordings. An important criterion for these decisions is the affect a composer is wanting to engender in listeners responses. Guastavino [8] suggests that salient features of the soundscape, such as periodicity and timbre, influence listeners psychological response of places. For example, a listener may feel an antipathy toward a soundscape filled with noisy machine sounds, but be more inclined to favour one with more tonal machine sounds.

The psychological response of a listener is of concern to soundscape composition practice. According to Truax [16], there are four important characteristics for composing a soundscape:

- Listener recognizability of the source material is maintained.
- Listener's knowledge of the environmental and psychological context is invoked.
- Composer's knowledge of the environmental and psychological context influences the shape of the composition at every level.
- The work enhances our understanding of the world and its influence carries over into everyday perceptual habits.

In our research, we address the quality dimensions of valence and arousal corresponding to the third characteristic of soundscape composition outlined by Truax. Specifically, the aim is to provide autonomous feedback of these qualities to performers.

Performance environments provide different challenges than studio production. Studio production of soundscape composition facilitates the composers ability to reflect upon the composition in a controlled environment. Given this environment, they are enabled to reconciling disparities between the composition product and intended listener response. Alternatively, if improvising in performance environments, a composer must make more immediate technical and aesthetic decisions. Consequently, the allocation of time for contemplation of intent becomes exceedingly restricted. Soundscape composers in performance environments would benefit from a tool that provides feedback on soundscape affect. At the the time of writing, no such tool exists.

We describe the *Impress* system for data acquisition and affect classification of audio environments. Specifically, a simple means for gathering data from a mobile device using an affect grid is examined. In addition, a method for the autonomous evaluation of soundscape composition in performance environments using a supervised machine learning algorithm is detailed. Our contribution is a system for composers to acquire personal soundscape affect data for providing visual feedback of compositions in performance environments using a supervised machine learning algorithm.

This paper is organized as follows. In Section 2. we cover related works that form the basis of the affect grid and soundscape affect prediction. Section 3. details the *Impress* system architecture, including the affect grid, visual interface, and supervised machine learning algorithm. Section 4. describes the evaluation of the system from a user study. We conclude and speak to the future aims of the research in Section 5.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
NIME'13, May 27 – 30, 2013, KAIST, Daejeon, Korea.
Copyright remains with the author(s).

2. RELATED WORKS

In our research, we use an affect grid for acquiring user response data of a soundscape. Furthermore, the same grid provides visual feedback of the predicted responses of audio environments. An affect grid is a tool for the quick assessment of affect along orthogonal dimensions of valence and arousal. Russell [13] was one of the first to detail an affect grid as a simple means of acquiring descriptive or subjective judgments in studies. According to Russell, arousal refers to the perceived activity of the stimulus. Whereas, valence refers to the degree of pleasantness. He suggest that a grid is more effective than other response forms in studies that require continuous or repeated observations. Therefore, it lends itself well to dealing with the rapid fluctuations of affect that occur in response to complex audio stimulus, such as soundscape, or music.

In the literature, examples of the affect grid being used as a means of collecting users response have tended to focus on mood classification of music[10, 9, 17]. Stockholm and Pasquier [14] use a variation of the affect grid for music mood classification. In their research, grid dimensions are labelled with pleasure and energy. They examine a method of reinforcement machine learning for mood classification of audio files based on listener response during an interactive performance.

The labelling of dimensions on an affect grid represent some *qualities* of a domain. We use an affect grid with orthogonal dimensions *unpleasant-pleasant* and *uneventful-eventful*, which have a greater specificity for soundscape. At time of writing, no formal two dimensional system for eliciting responses to soundscape was available in the literature. However, much work on people's preference of soundscapes, especially in urban design studies, provides perceptual measurement scales that could be used for an affect grid. Birgitta et al. [2] conducted listening experiments, finding people classified soundscapes on scales of *pleasant-unpleasant*, and, *eventful-uneventful*. Davies et al. [4] developed a listener response form for evaluating urban soundscapes that included subjective scales of preference. Their research found that an accurate evaluation of a soundscape could be obtained by listeners rating along linear scales of *unpleasant-pleasant*, *agitated-calm*, and *gloomy-fun*. Brocolini et al. [3] modelled the relationship of listeners responses of two soundscapes with a system similar to Birgitta et al. and Davies et al. The subjective scales in their research are, namely, *unpleasant-pleasant*, *quiet-noisy*, *not loud-loud*, *not present-present*.

For soundscape classification based on subjective responses, research has tended to focus on questionnaire based analysis instead of modelling the audio features of the soundscape. An increasing number of studies have found that audio features and machine learning techniques are an effective in classifying environmental sounds, such as those related to soundscape. Aucoutier et al. [1] propose the *bag of frames* (BOF) technique. They use an audio features vector of Mel Frequency Cepstral Coefficients for similarity comparisons of audio environments, such as a park, or a urban square. Their approach suggests audio signals may be better represented by a number of frames with different values, which makes it an attractive method for representing audio environments that evolve over longer durations.

3. SYSTEM ARCHITECTURE

Impress was designed to provide a visual interface for the data acquisition and feedback of soundscape affect. There are two stages for operating *Impress*. First, audio and response data is acquired in the collection stage. Second, the data is modelled and the affect of new soundscapes are predicted.

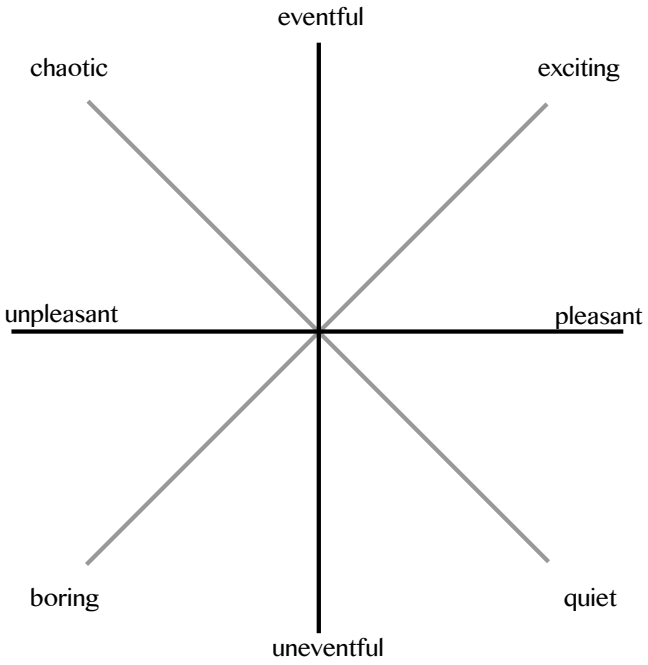

Figure 1: The soundscape affect grid with a circumplex ordering of affect labels.

3.1 Affect Grid

Responses to soundscapes in *Impress* are represented by evaluative labels of affect around a two-dimensional continuous scale. The dimensions of this affect grid closely resemble the evaluative responses as outlined in the soundscape literature. In particular, on the dimension of valence, *pleasant*, and *unpleasant* are used to report the perceived pleasantness of a soundscape. Similarly, for the reported *feeling* of arousal, *eventful*, and *uneventful* are positioned orthogonal to the pleasantness dimension.

A circumplex ordering of affect is made by a rotation of the axes of an affect grid [12]. In our research, labels attributed to the rotation are *exciting* for a *pleasant* and *eventful* sound, *quiet* for a *pleasant* and *uneventful* sound, *chaotic* for a *unpleasant* and *eventful* sound, and, *boring* for a *upleasant* and *ueventful* sound. The affect grid used in *Impress* is shown in Figure 1.

The collection stage of the *Impress* system involves logging audio analysis data and the user response of an audio environment. This data is obtained when the device is put into *listening mode* through a button in the GUI. Figure 3. shows the device engaged in the collection stage, as it would be used in the field. Audio analysis data is derived from an 4 second audio signal buffer that is updated FIFO. A buffer of this length is applied to capture the complex and slowly evolving properties of a soundscape signal, and is computationally feasible for the system. An audio-signal is represented with the mean and standard deviation of low-level audio features, similar to the *bag of frames* approach (BOF). A BOF considers that frames representing a signal have possibly different values, and the aggregation of the frames provides a more effective representation than a singular frame. This method was chosen because it is one of the most practical for representing complex audio signals of audio environments.

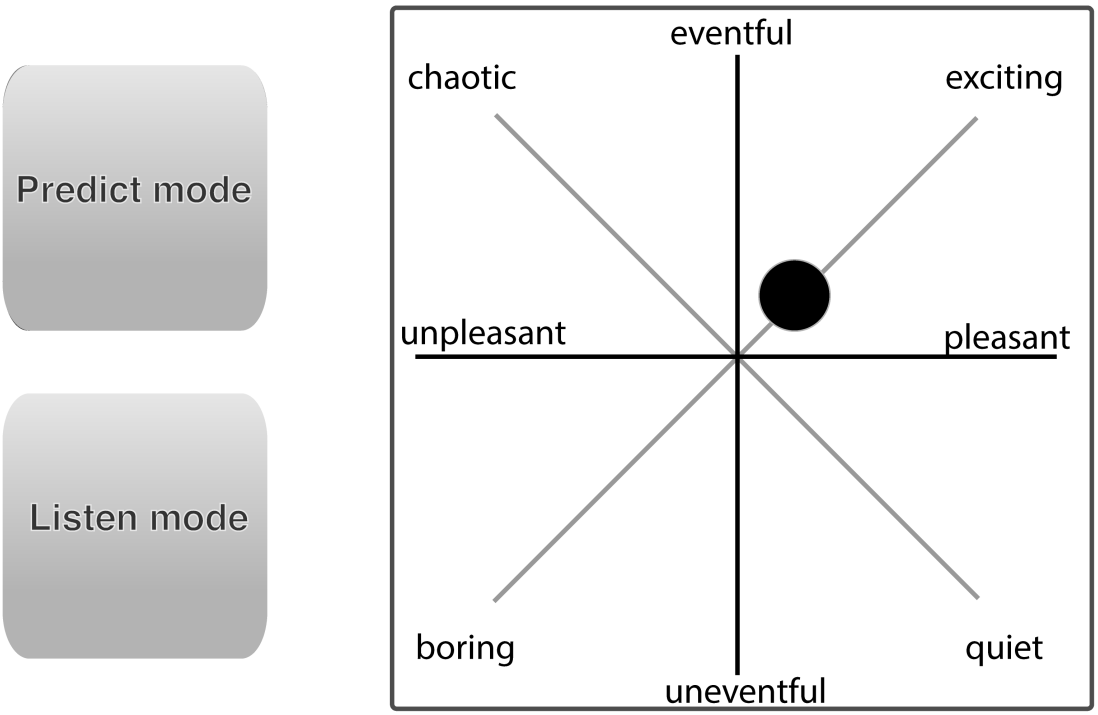

Figure 2: The system interface showing the affect grid and *listening*, and *predict* mode buttons.

Figure 3: The system as it would be used in the field engaged in the *listening* mode.

3.2 Collection Stage

3.2.1 Acquisition of Response and Features Vector

The aim of the collection stage is to acquire audio analysis and user affect response data associated with a soundscape. Soundscape affect is represented along the x and y axes of the grid corresponding to the dimensions of pleasantness and eventfulness. When the user enters their response on the affect grid, those coordinates are logged and the system exits *listening mode*. Thereafter, audio features are extracted from the signal buffer and modelled using the BOF. The response coordinates audio features vector are logged to a database for further analysis in the *prediction mode*.

3.2.2 Audio Features for Modelling Soundscape

Audio features extracted for analysis are Total loudness, Perceptual spread, Perceptual sharpness, and Mel Frequency Cepstral Coefficients (MFCC). These features are perceptually motivated, which is a key consideration in soundscape studies. Total loudness is the characteristic of a sound associated with the sensation of intensity. The human auditory system effects the perception of intensity of sound at different frequencies. The model of loudness provided by Fasl and Zwicker [6], takes into account the disparity of loudness at different frequencies along the Bark-scale, which correspond to the critical bands of hearing. A specific loudness is the loudness associated at each of these bands. The total loudness is the sum of individual specific loudness in all bands.

Perceptual spread is the spread of the loudness coefficients computed as the distance from the largest specific loudness value to the total loudness. Similarly using the Bark-scale, perceptual sharpness is the sharpness of the loudness coefficients, computed as a distribution of frequencies with probabilities of observing these as the normalized specific loudness at the critical bands of hearing.

MFCCs are commonly used in speech recognition systems, and are found to be an effective feature in music and environmental sound classification. MFCCs represent the short time spectrum of an audio signal that are spaced along the a perceptual scale of pitches that models the response of human pitch perception. The common representation of the MFCC is of filter bank values linearly spaced at frequencies below 1000Hz, and logarithmic spaced filters at higher frequencies.

Audio features were extracted using the YAAFE [11] software package. We use a feature vector of the density distribution of Total loudness, Perceptual spread, Perceptual sharpness, and 40 MFCC calculated using the BOF approach, which results in an 86 dimension feature vector.

3.3 Prediction Mode

The aim of *Impress*, whilst in *prediction mode*, is to use a model trained with data from the collection stage for predicting the position on the affect grid that represents an

audio signal. *Impress* is put in prediction mode through a check box in the interface. When entering this mode, collection stage data is retrieved from the database. The data is then used to build a multiple regression model for both axes of the affect grid.

3.3.1 Multiple Linear Regression Predictor

We used a multiple linear regression (MLR) for modelling the relationship between response variables and predictor variables. We chose this relatively straightforward model to observe any positive linear relationships between audio features vectors and soundscape affect responses. This type of model is suited to problems with continuous variables, such as that we described in our research. The goal of MLR is to find a vector of coefficients representing the strength of the linear relationship between the response variable and predictor variables. The regression model fits a linear function to a set of data points. The form of the function is:

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

Where Y' is the predicted response, $X_{1\ldots k}$ are the predictor variables, A is the value of Y' when all $X_{1\ldots k} = 0$, and $\beta_{1\ldots k}$ are the regression coefficients.

In our case we build a separate model for the x and y axes of the affect grid, where Y' are the predicted responses for pleasantness and eventfulness, $X_{1\ldots k}$ are the features vector for predicting Y', and $\beta_{i1\ldots k}$ are the regression coefficients for each axis.

3.3.2 Prediction Updating

After training is complete, *Impress* records an audio signal into a 4 second buffer that is continuously updated FIFO. *Impress* iteratively copies and processes the buffer to make predictions from an audio features vector. An audio features vector extracted with the same method as in Section 3.2 is used to compute a prediction of soundscape affect on both axes respectively. After each prediction the GUI is updated by moving a black dot on the affect grid to the reflect the predicted response. The interface updates are smoothed using a simple in-out cubic easing algorithm applied to the movement of the dot.

4. PREDICTOR EVALUATION

The soundscape mood predictor described here uses a supervised machine learning approach for making predictions of an audio signal. Specifically, we use multiple linear regressions to predict the correct response along two axes of an affect grid given a vector of audio features. We evaluated the regression model with 250 data points using a corpus of 4 second sound recordings obtained from a user study. The user in the experiment was familiar with soundscape and had been actively involved with soundscape composition.

We developed a tool for building a corpus of audio files. First, audio files from the Freesound [7] audio repository are downloaded given search keywords, a number of files requested, and a duration range of files. Next, sections of audio recordings are cut for the desired duration of corpus items. This was achieved using a segmentation algorithm to search recordings for regions with a consistent soundscape characteristic greater or equal to the required duration [15]. The middle section the region is copied and stored for further analysis. Lastly, audio features are extracted and logged to a database with the corresponding file name.

Feature extraction was performed on sound recording regions formatted in AIF and a sample rate of 22500Hz. Audio features were extracted at the frame-level with a 23 ms Hanning window and a step size of 11.5 ms. Computing the means and standard deviations of these frames resulted in an 86 dimension feature vector for each sound region. We then used a both-ways stepwise regression to identify predictor variables for fitting the regression models, resulting in subsets of predictors for fitting the *pleasantness* and *eventfulness* models.

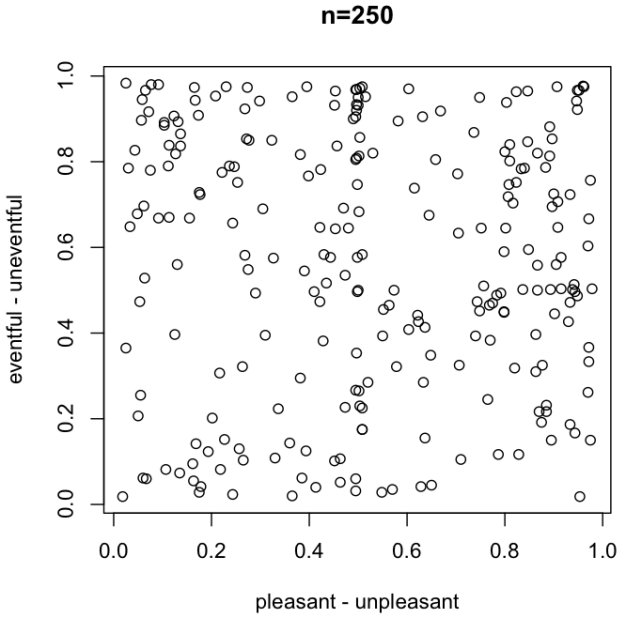

Figure 4: The spread of 250 data points of users affect response to soundscape recordings from an evaluation study along the *pleasantness* and *eventfulness* axes.

In this evaluation, each 4 second sound recording was obtained from separate audio clips tagged as *field-recording*, downloaded from the Freesound audio repository. The recordings were curated to remove erroneous items from the corpus, such as music loops. The user was asked to listen to each recording using a desktop computer and headphones. After the recording finished playing, they input a response on the affect grid (see Section 3) presented on the computer monitor by using a mouse. The responses were logged to a database for further analysis.

4.1 Evaluation Results

The regression models where then fitted with feature vectors and corresponding participant responses. Multiple regression analysis was used to test if the audio features vector significantly predicted participant's ratings of soundscape affect on two dimensions. The results of the regression for the dimension of *pleasantness* indicated the audio features vector explained 71.2% of the variance ($R_2 = .712, F(35, 214) = 15.1, p < .001$). Whereas, on the dimension of *eventfulness*, the results of the regression indicated the audio features vector explained 71% of the variance ($R_2 = .71, F(45, 204) = 11.0, p < .001$).

We conducted a k-fold validation and calculated the mean square error (MSE) to evaluate the prediction accuracy of the linear regressions. Specifically, a 10-fold cross validation strategy was used. This technique involved randomizing the data set and splitting it into equal sized partitions. Thereupon, one partition was separated and the model built with the remaining partitions. This process was repeated for all partitions. The MSE over 25 folds for the *pleasantness* regression model was 0.0392. Similarly, the *eventfulness* regression model MSE over 25 folds was 0.0348.

4.2 Discusion

The methodology for representing soundscape data with small number of audio descriptors was presented. We showed an evaluation of the multiple regression models for predicting soundscape affect based upon a user study. The disparity of the audio descriptors found to fit each model suggests that a different criteria is employed by listeners when responding to affect along these dimensions. This statement is reinforced by the similar prediction accuracy and MSE of both models even though different subsets of predictors were used.

However, the study results suggest that the models were not perfect predictors, eventhough a good correlation between the explanatory and response variables was demonstrated. Moreover, it means further that there are other independent variables, not studied, that effect the response variable. Our nascent hypothesis from these results is that other cultural factors, such as automobile traffic or more natural sounds, contributed to the affects of pleasantness and eventfulness even though the spectral and temporal characteristics of the sound may be similar.

5. CONCLUSIONS AND FUTURE WORK

We have shown a system for predicting two quality dimensions of soundscapes using a simple approach to acquiring and modelling soundscape affect. This work promises to provide autonomous feedback on soundscape compositions in performance environments. Performers adopting *Impress* will benefit by relegating the task of evaluating the mood of the composition to the machine. As a result, giving greater attention to other performance tasks.

We demonstrated the user interaction process for operating *Impress*. In particular, how the an affect grid is used for soundscape evaluation, and, secondly, visual feedback on the prediction of soundscape. This interface, coupled with a mobile device, facilitates quick and repeated acquisition of soundscape data in real-world conditions.

The system describe here is a tool for building a model of soundscape affect based upon personal evaluations of audio environments. Primarily, the application of *Impress* is performance contexts of soundscape composition, where it facilitates the artistic point of view of the composer. However, we would like to see the level of agreement between human subjects of the affect of the soundscapes. Consequently, investigating how generalizable categories of soundscape affect are. That research will contribute toward future work of applying soundscape affect prediction in information retrieval of audio recordings. Specifically, it will aim at furthering previous work that classified audio recordings based upon measures of timbral proximity [5]. By further adding a measure of affect users will be able to request files with an additional dimension of quality.

6. ACKNOWLEDGMENTS

This research was partly funded by a grant from the Natural Sciences and Engineering Research Council of Canada.

7. REFERENCES

[1] J.-J. Aucouturier and B. Defreville. Sounds like a park: A computational technique to recognize soundscapes holistically, without source identification. *19th International Congress on Acoustics*, 2007.

[2] B. Berglund, M. E. Nilsson, and O. Axelsson. Soundscape psychophysics in place. In *InterNoise*, 2007.

[3] L. Brocolini, L. Waks, C. Lavandier, C. Marquis-Favre, M. Quoy, and M. Lavandier. Comparison between multiple linear regressions and artificial neural networks to predict urban sound quality. In *Proceedings of 20 th International Congress on Acoustics*, 2010.

[4] W. J. Davies, M. D. Adams, N. S. Bruce, A. Carlyle, and P. Cusack. A positive soundscape evaluation tool. *Changes*, 2009.

[5] A. Eigenfeldt and P. Pasquier. Real-time timbral organisation: Selecting samples based upon similarity. *Org. Sound*, 15(2):159–166, 2010.

[6] H. Fastl and E. Zwicker. *Psychoacoustics: Facts and Models*. Springer series in information sciences. Springer, 2007.

[7] Freesound. Available online at `http://www.freesound.org/`; visited on April 12th 2013.

[8] C. Guastavino. The ideal urban soundscape: Investigating the sound quality of french cities. *Acta Acustica united with Acustica*, 92(6):945–951, 2006.

[9] P. Hunter and E. Schellenberg. Music and emotion. In M. Riess Jones, R. R. Fay, and A. N. Popper, editors, *Music Perception*, volume 36 of *Springer Handbook of Auditory Research*, pages 129–164. Springer New York, 2010.

[10] C. Laurier, O. Meyers, J. Serrà, M. Blech, P. Herrera, and X. Serra. Indexing music by mood: design and integration of an automatic content-based annotator. *Multimedia Tools and Applications*, 48:161–184, 2010.

[11] B. Mathieu, S. Essid, T. Fillon, J.Prado, and G.Richard. YAAFE, an Easy to Use and Efficient Audio Feature Extraction Software. In *Proceedings of the 2010 International Society for Music Information Retrieval Conference (ISMIR)*, Utrecht, Netherlands, 2010. ISMIR.

[12] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[13] J. A. Russell, A. Weiss, and G. A. Mendelsohn. Affect Grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology*, 57(3):493–502, 1989.

[14] J. Stockholm and P. Pasquier. Eavesdropping: audience interaction in networked audio performance. In *Proceedings of the 16th ACM international conference on Multimedia*, MM '08, pages 559–568, New York, NY, USA, 2008. ACM.

[15] M. Thorogood and P. Pasquier. Computationally Generated Soundscapes with Audio Metaphor. In *Proceedings of the Fourth International Conference on Computational Creativity*, Sydney, Australia, 2013. ICCC.

[16] B. Truax. Soundscape Composition, 1996-2012. Available online at `http://www.sfu.ca/~truax/scomp.html`; visited on January 23rd 2013.

[17] Y.-H. Yang and H. H. Chen. Machine recognition of music emotion: A review. *ACM Trans. Intell. Syst. Technol.*, 3(3):40:1–40:30, 2012.