Journal of the Audio Engineering Society
Vol. 64, No. 09, September 2016 (© 2016)
DOI: http://dx.doi.org/10.17743/jaes.2016.0044

Automatic Soundscape Affect Recognition Using A Dimensional Approach

JIANYU FAN, MILES THOROGOOD, AND PHILIPPE PASQUIER
(jianyuf@sfu.ca) (mthorogo@sfu.ca) (pasquier@sfu.ca)

Simon Fraser University, SIAT, Canada

Soundscape affect recognition is essential for sound designers and soundscape composers. Previous work demonstrated the effectiveness of predicting valence and arousal of soundscapes in responses from one expert user. Based on this, we present a method for the automatic soundscape affect recognition using ground truth data collected from an online survey. An analysis of the corpus shows that participants have a high level of agreement on the valence and arousal of soundscapes. We generate a gold standard by averaging users' responses, and we verify the corpus by training stepwise linear regression models and support vector regression models. An analysis of the models shows our system obtains better results than the previous study. Further, we test the correlation between valence and arousal based on the gold standard. Last, we report an experiment of using arousal as a feature for predicting valence and vice versa.

0 INTRODUCTION

This study is inspired by research in the fields of soundscape studies and perception. Soundscape researchers have demonstrated the variety of approaches taken to investigate how soundscapes affect people for the creation of immersive experiences [1–4]. Our research tries to develop an automatic soundscape affect recognition system that soundscape composers can use to create emotional soundscape compositions to evoke a target mood. This system can offer sound designers a more streamlined workflow for creating suitable sound effects for films and can offer engineers a way to design mood-enabled recommendation systems for retrieval of soundscape recordings.

To build such a system, we conducted an online study to collect ratings of valence and arousal from multiple participants. Then, we explored whether soundscape recordings evoke the same emotions in different listeners by analyzing the agreement between user ratings. Next, we tested the effectiveness of a soundscape affect recognition system with data from multiple users. After that, we present the correlation between arousal and valence based upon the gold standard. Finally, we report the results of using valence as a feature for predicting arousal.

This paper is organized as follows. In Sec. 1 we discuss the related works in the domain of audio affect recognition. Next, in Sec. 2 we describe the creation of soundscape corpus and a gold standard obtained from an auditory perception experiment. Sec. 3 details the evaluation results. Finally, we present our conclusions and future work in Sec. 4.

1 RELATED WORKS

Both the discrete and dimensional models are widely used in the affective computing field. Russell proposed the dimensional model [9] [28] and describes an emotion using a continuous circumplex space of emotional attributes, which include pleasantness and eventfulness. The discrete model classifies an emotion into one of a finite number of categories [13].

Music Researchers have used these models to study emotional ratings of music. Eerola et al. [7] explored various regression techniques to analyze musical features using both a circumplex and discrete model for modeling music emotion responses. Lu, Liu, and Zhang [8] studied mood detection based on a valence-arousal circumplex model. Stockholm and Pasquier [22] modified the labels of dimensions to pleasure and energy. They then built a mood classification system based on reinforcement learning techniques. Van't Klooster and Collins [24] presented an emotion driven live perform system. It used a dimensional model for collecting users' responses for mood classification of piano music.

Likewise, soundscape researchers adopted a similar methodological approach for eliciting and modeling emotional responses to soundscapes. Berglund et al. [1] investigated how people perceived recordings of soundscapes categorized as "technological," "natural" or "human." They

This article is part of the Special Issue on Intelligent Audio Processing, Semantics, and Interaction. See the guest editors' note on page 464 of the 2016 July/August issue.

describe a listener survey to ascertain the important emotional attributes. One-hundred listeners were asked to evaluate 30-second recordings of 30 outdoor soundscapes with the help of 116 perceptual-emotional attribute scales. A principal component analysis selected two important dimensions: pleasantness (50%) and eventfulness (16%). They defined the metric space using the following axes: pleasant-unpleasant, exciting-boring, eventful-uneventful, and chaotic-tranquil. Their study found that "eventfulness was perceived to increase with increases in overall sound level, but this relationship was found to be weaker for pleasantness (Pearson's r = 0.4 for eventfulness and –0.7 for pleasantness)." They also indicate that soundscapes dominated by human sounds were perceived as more eventful than soundscapes without human sounds.

Davies et al. [21] designed a listener response form for evaluating urban soundscapes based on subjective scales of preference. The study showed 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. [2] did a field survey to study the relationship between sound pleasantness and other subjective variables. Authors asked passers-by to rate the pleasantness of the surrounding sounds, visual environment, air quality, and overall environment. They then asked passers-by to evaluate soundscape characteristics, such as quiet-noisy, stable-changing, lifeless-lively, and surprising-familiar. Their study demonstrated that the acoustic scene has a significant effect on one's evaluation of pleasantness. Therefore, it is possible to analyze the soundscape apart from the landscape and visual aspect of the scene.

Our work is based on a previous study of the Impress system [3], which was an automatic soundscape affect prediction system designed for real-time environments. Thorogood and Pasquier [3] curated a corpus of audio files using an automatic segmentation algorithm [10] [15] that keeps audio regions with consistent soundscape characteristics. The segmentation algorithm was designed based on perceptual categories including background, foreground, and background with foreground sound. They used a circumplex model for collecting continuous data to measure a subject's reaction to soundscapes. Then, they used multiple linear regressions to model audio features and expert user affective ratings to soundscape recordings. Each excerpt is four seconds long. The model was trained with 250 data points. 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 Sec. 3.2.

Thorogood and Pasquier [3] built a soundscape affect recognition model based on one user's emotional responses. We extend that research [3] by conducting an online survey to collect ground truth data to build a soundscape affect recognition model.

2 METHODS

2.1 Circumplex Model

A circumplex model suggests that emotions are distributed in two-dimensional spaces [8]. Arousal represents the perceived activity of the stimulus. Valence refers to the degree of pleasantness [9]. A circumplex ordering of affect is made by a rotation of the axes of an affect grid [28]. Berglund et al. indicated that people classified soundscapes on scales of pleasant-unpleasant and eventful-uneventful. They state that their measurement model for soundscapes is "compatible with Russell's circumplex model of human emotions" [1].

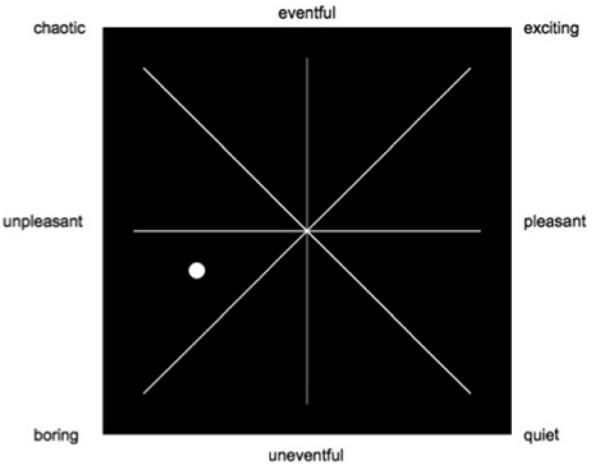


Fig. 1. The online study interface: the x-axis represents valence (pleasantness). The y-axis represents arousal (eventfulness). After a user had formed an opinion, the user clicks the cursor on the circumplex model [6].

The only formal two-dimensional system for eliciting responses to soundscapes was presented by Thorogood and Pasquier [3]. Their system, Impress, used the criteria "pleasant" and "unpleasant" to report the perceived pleasantness of a soundscape and used "eventful" and "uneventful" to report the feeling of arousal. On the axes of the affect grid, they applied the labels "exciting" for a pleasant and eventful sound, "quiet" for a pleasant and uneventful sound, "chaotic" for an unpleasant and eventful sound, and "boring" for an unpleasant and uneventful sound.

In this study we used the valence and arousal affect model. Valence in our case is the perceived pleasantness of a soundscape, and arousal indicates the intensity of emotion provoked by the soundscape. To easily provide explicit affective ratings on our valence and arousal model, we used the same circumplex model [3] with the axes separated by 45 degrees: pleasant-unpleasant, exciting-boring, eventful-uneventful, and chaotic-quiet (Fig. 1). Without a diagonal axis in the model, participants tend to rate the affect near the x and the y-axes [3].

2.2 Collection Stage

2.2.1 Corpus

Schafer's referential taxonomy was widely used for the classification of soundscapes. According to Schafer, "Sounds of the environment have referential meaning" [4]. He grouped soundscapes based on their context rather than content or physical characteristics. Table 1 shows Schafer's soundscape taxonomy.

Table 1. Schafer's soundscape taxonomy [6].

Categories	Examples
Natural sounds	Bird, chicken, rain, sea shore
Human sounds	Laugh, whisper, shouts
Sounds and society	Party, concert, store
Mechanical sounds	Engine, cars
Quiet and silence	Wild space, silent forest
Sounds as indicators	Clock, doorbell

Three experts selected audio clips following six categories according to Schafer's taxonomy. It is straightforward to decide which category a sound is from (e.g., "birds in the forest" would be natural sound; "airplane engine" belongs to mechanical sounds). We selected 31 clips of natural sounds, 23 clips of mechanical sounds, 39 clips of sound as indicators, 11 clips of quiet and silence, 16 clips of sounds and society, and no human sounds. We used the Sound Ideas corpus [11] and World Soundscape Project [12], which have consistently good audio recording quality.

Sound Ideas 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 [11]." The World Soundscape Project (WSP) was established by Murray Schafer at Simon Fraser University in the late 1960s [4]. The project is to "find solutions for an ecologically balanced soundscape where the relationship between the human community and its sonic environment is in harmony. [4]" Later on, Barry Truax and the Metacreation Lab[1] digitized the WSP.

The previous study used 4-second excerpts. However, after our preliminary testing we decided to extract 6-second excerpts from each of the sounds we selected, which would give participants more time to form an opinion of both valence and arousal for a soundscape. Each clip is monophonic. The sample rate is 44100 Hz. Regions were selected based on a soundscape audio signal segmentation algorithm using listeners' perception of background and foreground sound [10] [15].

2.2.2 Online Study

Twenty students who took a sound design class from Simon Fraser University participated the online study. The mean age of all the participants was 21.7 years. There were 12 males and 8 females. The study was done in a lab environment. Participants used lab computers, each of which had the same monitor and same sound card. Participants used circumaural headphones to listen to the audio clips. We asked participants to adjust the volume to a comfortable level instead of asking them to adjust the volume to the same loudness. This could add robustness to our findings. We gave a tutorial of the task. We omitted the affective ratings of the first 5 clips to allow users to calibrate their answers for practices. Thus, we had 120 audio clips summed over all 6 categories. Participants rated them in random order. The study was conducted online through a web browser. We used an HTML5 audio player object to play audio excerpts. There was no timer, so participants were allowed to listen to an excerpt repeatedly. After a user had listened to an audio clip and formed an opinion of affect, the user used a mouse to click on the circumplex model (the interface is shown in Fig. 1) to enter their response. In Fig. 1, the x-axis represents the level of valence, and the y-axis represents the level of arousal.

2.3 Data Analysis

2.3.1 Agreement between Participants

To build a gold standard model, we need to demonstrate participants' high level of agreement on the valence and arousal of soundscapes. The intraclass correlation coefficient (ICC) was used to measure the reliability of measurements of ratings in both valence and arousal. 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, suggest that participants highly agree with each other regarding soundscape affect. The higher index of arousal suggests that it is easier for observers to agree on valence than arousal. We obtained the gold standard by averaging responses provided by the 20 experimental participants.

2.3.2 Audio Features for Modeling Soundscape

We selected the audio features based on the previous study [3]. Audio features were extracted using the YAAFE [17] software package. We used the bag-of-frames (BOF) approach proposed by Aucouturier et al. [20], which represents signals as the long-term statistical distribution of local spectral features. We ended up with a 98-dimension feature vector. We resampled the audio from 44100 Hz AIF format to 22050 Hz and applied a 23 ms Hanning window of 512 samples. The mean and standard deviation of features is calculated.

Total loudness is a feature that describes the psychological correlate of physical strength (i.e., the sensation of intensity) [17]. It is the sum of the individual loudness from all bands along the Bark scale [16].The distance between the highest loudness value along the Bark scale and the total loudness is called perceptual spread. The perceptual sharpness is computed using the specific loudness of Bark bands [17]. Energy is computed as the root mean square of an audio frame [17].

Spectral flatness is computed by using the ratio between the geometric and arithmetic means [17]. Spectral flux is the flux of the spectrum between consecutive frames [17]. Spectral roll-off is the frequency below which 99% of the energy is contained [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]. MFCCs are common features in speech recognition systems that recognize people by their voices [18]. They have also been used in timbre recognition [19]. MFCCs are short-term spectral-based audio features. Mel-frequency is based on the human auditory

[1] http://metacreation.net/

system, which does not have a linear perception of sound and maps different frequencies to perceived pitches.

2.3.3 Stepwise Linear Regression Models

We decided to use the same machine-learning model in [3], a bidirectional stepwise multiple linear regression model. It combines the standard multiple linear regression models with stepwise selection methods. It selects the most effective predictors to predict both valence and arousal values.

The model will remove nonsignificant variables, which also solves the problem of collinearity. Therefore, our model identifies the major predictors that influence the dependent variable. We did not set a threshold for the number of features; we kept all variables that are significant.

2.3.4 Support Vector Regression

Support vector regression has been widely used in affective computing fields, including music emotion recognition [25] and video affect recognition [26]. We used the support vector regression (SVR) option in the Weka software [23], which uses the sequential minimal optimization algorithm in Smola [27] to train a support vector regression using polynomial kernels. It guarantees that the optimal solution will be found. Induced by the selected kernel, the model maps the input data into a higher-dimensional feature space using nonlinear mapping and builds a linear model in this feature space to do prediction.

3 RESULTS AND EVALUATION

In this section we first present the results of stepwise linear regression models trained by gold standard data. Then we describe the results of support vector regression models trained by gold standard data. Next, we present results of individual models of each participant, which is trained based on each participant's data. Last, we present a correlation test between valence and arousal data in a gold standard, an experiment of using arousal as a feature for predicting valence and vice versa.

3.1 Evaluation Approach

We use the coefficient of determination (R^2) to evaluate the performance of our models. R^2 represents the amount of variability explained by the regressors in the model. If the R^2 is close to 1, it means the regression model is well fitted. We used 10-fold cross-validation partitioning our dataset into ten subsets and iteratively performing the learning on nine subsets and validating the model on the other subset. We calculated the mean squared error (MSE) to evaluate the prediction accuracy of the linear regressions.

3.2 Gold Standard Model

3.2.1 Stepwise Linear Regression Models

3.2.1.1 Stepwise Regression Model Based on Low-Level Audio Features. After obtaining the gold standard data, which is the average response of 20 participants, we used it to train the stepwise regression model to build a gold standard model.

Table 2. Results of predicting valence using the stepwise regression model with low-level audio features [6].

Categories used	R^2
All six categories	0.567
Without sounds as indicators	0.715
Only sounds as indicators	0.402
Only natural sounds	0.989
Only mechanical sounds	0.860

First, we used 10-fold cross-validation to test our model of using the six categories described in Sec. 2.2.1. Second, we used 10-fold cross-validation to test our model without the "sounds as indicators" category to study the influence of semantic information on valence and arousal evoked by soundscapes. Third, we tested our model by only using data from "sounds as indicators," "natural sounds," and "mechanical sounds" individually. We did not test categories of "sounds and society," "quiet and silence," and "human sounds" separately because the collection of excerpts of these three categories were less than 20 items.

Table 2 shows the results of predicting valence using the stepwise regression model and gold standard data. When we use all six categories, the R^2 for predicting valence is 0.567. When we remove the category of "sounds as indicators," the results of predicting valence indicates the model explained 71.5% of the variance ($R^2 = 0.715$, $F(7, 72) = 29.35$, $p < 0.001$). It is significantly higher than using data including all six categories. As Orians et al. described in [30], "Indicators serve as clues that something more fundamental or complicated is happening than what is measured by them." We assume the category of "sounds as indicators" carries strong semantic information, which plays an important role in evoking valence to listeners. The low R^2 (0.402) from only using data from "sounds as indicators" supports this assumption. Our model performs very well in predicting valence when only using "mechanical sounds" or "natural sounds."

For the model trained with data points of all six categories, significant predictors include the mean of loudness, standard deviation of perceptual sharpness, standard deviation of MFCC5, mean of MFCC18, mean of MFCC32, and mean of MFCC23. The equation for predicting valence is given.

$$\begin{aligned} Valence = {} & 0.231 + (-0.433) \times Loud_{Mean} \\ & + (-0.937) \times PSH_{Std} \\ & + 0.808 \times MFCC5_{Std} \\ & + 0.626 \times MFCC18_{Mean} \\ & + (-2.046) \times MFCC32_{Mean} \\ & + 0.732 \times MFCC23_{Mean} \end{aligned} \quad (1)$$

Table 3 gives the standardized coefficients of each feature, which shows the relative contribution. We can see

Table 3. Standardized coefficients for predicting valence.

Predictors	Standardized coefficients
Mean of loudness	−0.658
StdDev of perceptual sharpness	−0.325
StdDev of MFCC5	0.268
Mean of MFCC18	0.262
Mean of MFCC32	−0.227
Mean of MFCC23	0.171

Table 4. Results of predicting arousal using the stepwise regression model with low-level audio features [6].

Categories used	R^2
All six categories	0.816
Without sounds as indicators	0.876
Only sounds as indicators	0.737
Only natural sounds	0.800
Only mechanical sounds	0.983

Table 5. Standardized coefficients for predicting arousal.

Predictors	Standardized coefficients
Mean of loudness	0.444
StdDev of loudness	0.384
Mean of spectral roll-off	0.374
StdDev of MFCC26	0.201
StdDev of MFCC5	0.197
Mean of MFCC2	−0.115
Mean of MFCC28	−0.114

that mean of loudness gives a high negative contribution to perceived valence.

Table 4 shows the R^2 of predicting arousal using all six categories is 0.816. When we remove the category of "sounds as indicators," the results of predicting valence indicates the model explained 87.6% of the variance ($R^2 = 0.876, F(12, 67) = 47.634, p < 0.001$). Similar to predicting valence, the R^2 significantly decreases when we only tested "sounds as indicators." The R^2 decreases to 0.737.

For the model trained with data points of all six categories, significant predictors include the mean of loudness, standard deviation of loudness, mean of spectral roll-off, mean of MFCC2, mean of MFCC28, standard deviation of MFCC5, and standard deviation of MFCC26. The equation for predicting arousal is given.

$$\begin{aligned} Arousal = -1.441 &+ 0.317 \times Loud_{Mean} \\ &+ 0.556 \times Loud_{Std} \\ &+ 4.064 \times E^{-5} \times Sroll_{Mean} \\ &+ (4.296) \times MFCC26_{Std} \\ &+ 0.64 \times MFCC5_{Std} \\ &+ -0.038 \times MFCC2_{Mean} \\ &+ -0.604 \times MFCC28_{Mean} \end{aligned} \quad (2)$$

Table 5 gives the standardized coefficients of each feature, which shows the relative contribution. We can see that mean of loudness, mean of spectral roll-off, and the standard deviation of loudness gives the major positive contribution to perceive arousal.

Table 6. Results of predicting valence using the support vector regression model with low-level audio features.

Categories used	R^2
All six categories	0.542
Without sounds as indicators	0.711
Only sounds as indicators	0.384
Only natural sounds	0.861
Only mechanical sounds	0.753

Table 7. Results of predicting arousal using the support vector regression model with low-level audio features.

Categories used	R^2
All six categories	0.735
Without sounds as indicators	0.817
Only sounds as indicators	0.703
Only natural sounds	0.892
Only mechanical sounds	0.943

The result of predicting arousal ($R^2 = 0.816$) is better than the one for predicting valence ($R^2 = 0.567$). We assume that pleasantness of soundscapes is less differentiable than the eventfulness. The results for the ICC also showed that it is easier for observers to agree on valence than arousal.

The category of "sounds as indicators" carries strong semantic information, which has influence on the affect recognition task. Tables 2 and 4 show that this influence is not reflected by arousal as much as valence. We assume it is because there is a stronger relationship between semantic information and pleasantness of soundscape than the one between semantic information and eventfulness of soundscapes. Our gold standard model performs better than the expert user's results in [3]. The application is online.[2]

3.2.2 Support Vector Regressions

3.2.2.1 Support Vector Regression Models Based on Low-Level Audio Features. We built a gold standard model by training the support vector regression model with the gold standard data, which is the average response of 20 participants. The test is the same as the one described in Sec. 3.2.1.1. The results of predicting valence using the support vector regression model are shown in Table 6. The results are better than the previous study with Impress [3]. However, the results from using SVR are not as good as those obtained using stepwise linear regression.

The results of predicting arousal using the support vector regression model are shown in Table 7. The results are better than those obtained in the previous study [3]; however, they are not as good as those obtained using stepwise linear regression.

In general, the results of predicting valence and arousal using the support vector regression model are not as good as those obtained using stepwise linear regression. We think

[2]http://audiometaphor.ca/impress/index.html

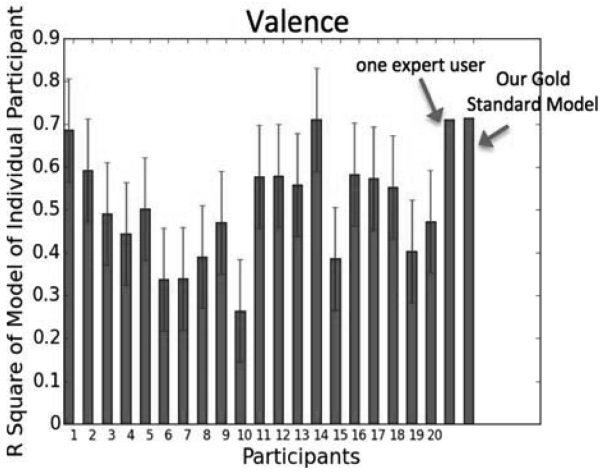


Fig. 2. R^2 of individual participants' models for predicting valence. The error bars represent the standard deviation [6].

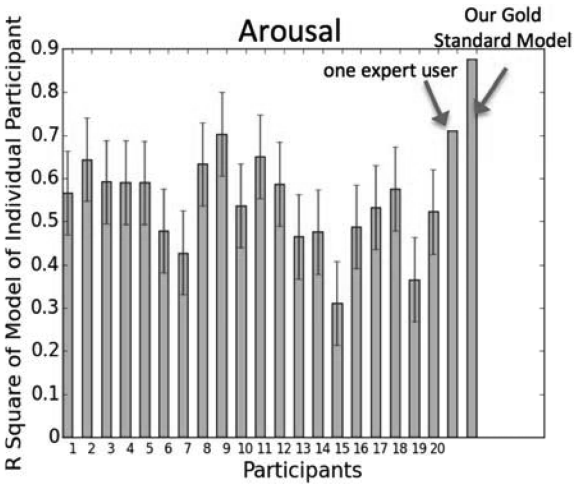


Fig. 3. R^2 of individual participants' models for predicting arousal. The error bars represent the standard deviation [6].

it is because the polynomial kernel function creates many higher level features based on low-level audio features we provided. Because our dataset is not huge, the combination of created features and existing features would cause the overfitting problem. This problem is avoided in the stepwise regression models by making feature selection and using 10-fold cross-validation.

3.3 Performance of the Stepwise Regression Models of Individual Participants

In the previous section, we demonstrated that the performance of stepwise regression was better than the performance of SVR. Therefore, in this section, we present the performance of the stepwise regression models of individual participants. Sec. 3.2.1.1 shows the improvement when not including "sounds as indicators." Because the model performed better without "sounds as indicators" (as a result of the semantic information this category contained) we removed this category. Fig. 2 shows the R^2 of all 20 participants' models that predict the valence of soundscapes. It also includes the R^2 of the expert user's model in [3] and our gold standard model.

Fig. 3 shows the R^2 of all 20 participants' models that predict the arousal of soundscapes. It also includes the R^2 of the expert user's model in [3] and our gold standard model.

The individual models produce an average MSE of 0.182 for valence and 0.129 for arousal. As it is shown in Figs. 2 and 3, our gold standard model described in Sec. 3.2 performs better than the results of the previous Impress system [3] (valence: R^2: 0.712, arousal: R^2: 0.71), which is also better than the performance of the current model of the individual participants.

3.4 Correlation between Valence and Arousal

Researchers have studied the relationship between valence and arousal [21] [29]. However, to our knowledge no empirical research exists addressing the question of how valence and arousal of soundscapes correlate to each other. Tsai et al. found that there is a preference for high arousal positive affect in western cultures and lower arousal positive affect in eastern cultures [29]. They suggested that valence varied inversely with arousal in Asian culture. In our study, 9 participants came from Asia and 11 participants came from North America. We ran a Pearson correlation test on the average value over the 20 participants' responses of valence and arousal. There are 120 data points. Each data point represents a value of valence and a value of arousal ranging from –1 to 1. Our Pearson correlation coefficient is –0.453 ($p < 0.01$), which indicates there is a moderate negative correlation between the two dimensions. This indicates that sounds that were rated as having higher arousal were rated as having lower valence. Considering the indication given by Tsai et al. in the scenario of soundscape affect we assume human listeners think a quiet and peaceful soundscape is more pleasurable.

Previous studies have also considered valence as a function of arousal. For example, Berlyne indicated that the pleasantness stimuli is maximized at an intermediate level of arousal [14]. We tested the effectiveness of including arousal as a new feature to predict valence and vice versa. We used data including all six categories to train stepwise regression models. The performance of the models is shown in Tables 8 and 9 ($p < 0.001$). The results indicate that when

Table 8. Results of predicting valence using the stepwise regression model with low-level audio features and arousal

	Just include low-level audio features	Include arousal as a new feature
R^2	0.563	0.570

Table 9. Results of predicting arousal using the stepwise regression model with low-level audio features and valence.

	Just include low-level audio features	Include valence as a new feature
R^2	0.816	0.817

using either valence or arousal as a new high-level feature, the performance of the model would be improved.

4 CONCLUSIONS AND FUTURE WORK

We conducted an online study to obtain ratings from participants. Our analysis shows participants have a high level of agreement on the valence and arousal of soundscapes. Then, we found the performance of the stepwise linear regression was better than the performance of the support vector regression. Next, we built a gold standard model using stepwise linear regressions and gold standard data. Our model performed better than the expert user model and any of the individual study participants. Moreover, we tested the correlation between responses of valence and arousal using gold standard data and found a moderate negative correlation between theses two dimensions. Finally, we reported the results of using arousal as a feature to predict valence, and vice versa.

For the next stage, we plan to study the ability of this model to predict the responses of people with different cultural backgrounds. We will use deep learning methods to do soundscape affect recognition.

5 REFERENCES

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

[2] L. Brocolini, L. Waks, C. Lavandier, C. Marquis-Favre, M. Quoy and M. Lavender, "Comparison between Multiple Linear Regressions and Artificial Neural Networks to Predict Urban Sound Quality," *Proceedings of the 20th International Congress on Acoustics*, pp. 2121–2126, Nantes, France (2010).

[3] M. Thorogood and P. Pasquier, "Impress: A Machine Learning Approach to Soundscape Affect Classification for a Music Performance Environment," *Proceedings of the International Conference on New Interfaces for Musical Expression*, pp. 256–260, Daejeon, Republic of Korea (2013).

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

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

[6] J. Fan, M. Thorogood, and P. Pasquier, "Automatic Recognition of Eventfulness and Pleasantness of Soundscape," *Proceedings of the 10th Audio Mostly*, Thessaloniki, Greece (2015). http://dx.doi.org/10.1145/2814895.2814927.

[7] T. Eerola, O. Lartillot and P. Toiviainen, "Prediction of Multidimensional Emotional Ratings in Music from Audio Using Multivariate Regression Models," *Proceedings of the 10th International Symposium on Music Information Retrieval*, pp. 612-626, Kobe, Japan (2009).

[8] L. Lu, D. Liu and J. Zhang, "Automatic Mood Detection and Tracking of Music Audio Signals," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 5–18 (2006). http://dx.doi.org/10.1109/TSA.2005.860344

[9] J. A. Russell, A. Weiss and G. A. Mendelssohn, "Affect Grid: A Single-Item Scale of Pleasure and Arousal," *J. Personality and Soc. Psych.*, vol. 57, no. 3, pp. 493–502 (1989). http://dx.doi.org/10.1037/0022-3514.57.3.493

[10] M. Thorogood and P. Pasquier, "Computationally Generated Soundscapes with Audio Metaphor," *Proceedings of the 4th International Conference on Computational Creativity*, pp. 1–7, Sydney, Australia (2013).

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

[12] B. Truax World Soundscape Project, Tape Library (2015). Available online at http://www.sfu.ca/sonic-studio/srs/index2.html; visited on January 12, 2015.

[13] T. Eerola and J. K. Vuoskoski, "A Comparison of the Discrete and Dimensional Models of Emotion in Music," *Psychology of Music,* vol. 39, pp. 18–49 (2011). http://dx.doi.org/10.1177/0305735610362821

[14] D. E. Berlyne, "Conflict, Arousal and Curiosity," *British J. Psychiatry*, vol. 108, no. 452, pp. 109–110 (1962). http://dx.doi.org/10.1192/bjp.108.452.109-a

[15] M. Thorogood, J. Fan, and P. Pasquie, "BF-Classifier: Background/Foreground Classification and Segmentation of Soundscape Recordings," *Proceedings of the 10th Audio Mostly*, Thessaloniki, Greece (2015). http://dx.doi.org/10.1145/2814895.2814926

[16] E. Wicker, "Subdivision of the Audible Frequency Range into Critical Bands," *J. Acous. Soc. Amer.*, vol. 33, no. 2, pp. 248 (1961). http://dx.doi.org/10.1121/1.1908630

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

[18] T. Ganchev, N. Fakotakis, and G. Kokkinakis "Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task," *Proceedings of the 10th International Conference on Speech and Computer*, pp. 191–194, Patras, Greece (2005). http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.75.8303

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

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

[21] P. Kuppens, "Individual Differences in the Relationship between Pleasure and Arousal," *J. Research in Personality*, vol. 42, pp. 1053–1059 (2008). http://dx.doi.org/10.1016/j.jrp.2007.10.007

[22] J. Stockholm and P. Pasquier, "Eavesdropping: Audience Interaction in Networked Audio Performance,"

Proceedings of the 16th ACM international Conference on Multimedia, pp. 559–568, New York, NY, USA (2008). http://dx.doi.org/10.1145/1459359.1459434

[23] B. Berglund, E. Nilsson and O. Axelsson, The WEKA Data Mining Software: An Update, SIGKDD Explorations, Newsl 11.1, pp. 10–18 (2009). http://dx.doi.org/10.1145/1656274.1656278

[24] A. van't Klooster and N. Collins, "In a State: Live Emotion Detection and Visualization for Music Performance," *Proceedings of the 14th International Conference on New Interfaces for Musical Expression*, pp. 545–548, London, United Kingdom (2014).

[25] F. Weninger, F. Eyben, W. Schuller, M. Mortillaro and K. Scherer, "On the Acoustics of Emotion in Audio: What Speech, Music, and Sound Have in Common," *Frontiers in Psychology*, vol. 4, pp. 1664–1078 (2013). http://dx.doi.org/ 0.3389/fpsyg.2013.00292

[26] I. Kanluan, M. Grimm and K. Kroschel, "Audio-Visual Emotion Recognition Using an Emotion Space Concept," *16th European Signal Processing Conference*, Lausanne, Switzerland (2008).

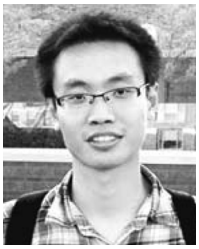
[27] A. Smola and B. Schoelkopf, "A Tutorial on Support Vector Regression," *NeuroCOLT2 Technical Report Series (1998).*

[28] J. A. Russell, "A Circumplex Model of Affect," *J.Personality and Social Psych.*, vol. 39, pp. 1161–1178 (1980). http://dx.doi.org/10.1037/h0077714

[29] J. L. Tsai, B. Knutson and H. H. Fung, "Cultural Variation in Affect Valuation," *J. Personality and Social Psych.*, vol. 90, pp. 288–307 (2006). http://dx.doi.org/10.1037/0022- 3514.90.2.288

[30] G. H. Orians, M. Dethier, C. Hirshman, A. Kohn, D. Patten, and T. Young, "Sound Indicators: A Review for the Puget Sound Partnership," Washington Academy of Sciences (2012).

THE AUTHORS



Jianyu Fan

Miles Thorogood

Philippe Pasquier

Jianyu Fan is currently a Ph.D. student at the School of Interactive Arts and Technology, Simon Fraser University. He holds a bachelor's degree from Beihang University and a master's degree from Dartmouth College. He is interested in the area of artificial intelligence, conducting research aimed at endowing machines with creative autonomous behavior. In particular, he focuses on music information retrieval systems and generative systems for music and videos. He has presented his papers in various international scientific venues. His artworks were presented at conferences and art festivals.

•

Miles Thorogood is a Ph.D. candidate at Simon Fraser University and lecturer in computer science and digital media at the University of British Columbia, Okanagan. He received his master of applied arts in signal modeling and human factors in 2010 from Emily Carr University of Art and Design. His doctoral dissertation deals with computationally creative systems to assist in human creative tasks. His academic research has been applied in industry and is regularly featured in public displays, such as the Vancouver Olympics in 2010 and City of Vancouver public works. His research interests include soundscape signal analysis, music information retrieval, modeling sound design process, and interdisciplinary research.

•

Philippe Pasquier is associate professor in interactive arts and technology and an adjunct professor in cognitive science at Simon Fraser University. He is both a scientist and a multi-disciplinary artist. His contributions range from theoretical research in artificial intelligence and multi-agent systems to applied artistic research and practice in computer music and generative art. Philippe is the chair and investigator of the AAAI International Workshops on Musical Metacreation (MUME), the MUME concert series, the International Workshops on Movement and Computation (MOCO), and he was director of the International Symposium on Electronic Arts (ISEA2015). He has co-authored over 120 peer-reviewed contributions, presented in forums ranging from the most rigorous scientific venues to the most subjective artistic ones.