2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)

Emo-Soundscapes: A Dataset for Soundscape Emotion Recognition

Jianyu Fan, Miles Thorogood, Philippe Pasquier
SIAT, Simon Fraser University
Vancouver, Canada
jianyuf, mthorogo, pasquier@sfu.ca

***Abstract*—Soundscape emotion recognition (SER) aims at the automatic recognition of emotions perceived in soundscape recordings. To benchmark SER, we propose a dataset of audio samples called Emo-Soundscapes and two evaluation protocols for machine learning models. We curated 600 soundscape recordings from Freesound.org and mixed 613 audio clips from a combination of these. The Emo-Soundscapes dataset contains 1213 6-second Creative Commons licensed audio clips. We collected the ground truth annotations of perceived emotion in these 1213 soundscape recordings using a crowdsourcing listening experiment, where 1182 annotators from 74 different countries rank the audio clips according to the perceived valence and arousal. This dataset allows studying SER and how the mixing of various soundscape recordings influences their perceived emotion. The dataset is at http://metacreation.net/emo-soundscapes/.**

1. Introduction

Soundscape emotion recognition (SER) aims at the automatic recognition of emotions perceived in soundscape recordings. The SER has received increasing interest from research communities, such as sound design [20], urban planning [29], and acoustic ecology [21]. Progress in machine learning for audio scene analysis has mainly been concerned with object detection [11] and scene context [2], while the problem of modeling the perceived emotion of a soundscape is not as well explored.

Perceived emotions are emotions that are represented and communicated by the source [1]. To accurately model and predict the perceived emotion of soundscapes, we require significant amounts of reliable ground truth data in the form of labeled audio recordings. To our knowledge, there is no publicly available soundscape database annotated with emotional labels. Further, there is no benchmark for SER as state of the art systems do not release datasets, or the size and diversity of samples are limited, thus making fair comparisons and reproducibility difficult. Therefore, we propose a new dataset, named Emo-Soundscapes, which contains 1213 audio clips released under Creative Commons license.

The paper is organized as follows. Section 2 covers background works regarding soundscape emotion analysis and soundscape databases. Section 3 provides the description of Emo-Soundscapes database. In Section 4, we describe the research instrument and experimental methodology for tagging clips. We present the analysis of the data in Section 5 and baseline machine-learning models in Section 6. We conclude with a summary and future work.

2. Background

2.1. Soundscape Emotion Studies

The soundscape research literature describes approaches for eliciting and modeling emotional responses to auditory stimuli. Berglund et al. [3] describe a listener survey to ascertain the critical emotional attributes people perceived in recordings of soundscapes categorized as 'technological,' 'natural' or 'human.' In their survey, 100 listeners were asked to evaluate 30-second recordings of 30 outdoor soundscapes on rating scales for 116 perceptual and emotional attributes. A principal component analysis of the listeners' annotation found the main principle components were pleasantness and eventfulness, explaining 50% and 18% of the total variance respectively. Furthermore, they found that eventfulness was perceived to increase with increases in overall sound level (Pearson's r = 0.4 for eventfulness and -0.7 for pleasantness).

Davies et al. [29] designed a survey for evaluating urban soundscapes based on subjective scales of preference. It showed that listeners rating along linear scales of pleasant–unpleasant, energetic–dull, calm–agitated, comforted–worried, and informed–confused could accurately evaluate the quality of urban soundscapes. Brocolini et al. [4] did a further survey of the relationship between sound pleasantness and environmental conditions (acoustic, visual and air quality). Their study demonstrated that the acoustic scene has a significant effect on one's evaluation of pleasantness.

Thorogood and Pasquier [5] proposed the Impress, which uses a linear model for automatic prediction of perceived pleasantness and eventfulness for soundscape recordings. Fan et al. [23] describe a corpus of audio files generated using a segmentation algorithm [14-15]. The system models audio features and experts' responses to soundscape recordings with stepwise regression models. Their evaluation showed a good fit of features to responses of models of predicting valence (R^2: 0.567) and arousal (R^2: 0.816).

Lundén et al. [9] investigated yet another method of predicting the outcome of the soundscape assessment based on acoustic features. Ninety-three clips from 77 audio recordings were selected for the study, and thirty-three participants were asked to move an icon into a 2D space to assess the pleasantness and eventfulness of soundscapes. To build the model, the authors used the bag-of-frame approach to represent the audio features [19]. Then, they used a Gaussian Mixture model to cluster features and used the resulting dissimilarity matrix to train two separate support vector regression models to predict the pleasantness and eventfulness of soundscapes respectively. The authors

978-1-5386-0563-9/17/$31.00 ©2017 IEEE

processed the data by detecting the internal consistency of participants' answers to remove outliers. The result indicates the MFCCs provide the strongest prediction for both pleasantness (R^2: 0.74) and eventful-ness (R^2: 0.83). The ground truth model's results are better than the average of individual models.

To our knowledge, authors of these previous studies did not release the ground truth data with affective annotations.

2.2. Emotion Taxonomy

According to previous studies [1, 22], two types of emotions are at play when listening to soundscapes.

- Perceived emotion: Emotions that are communicated by the source.
- Induced emotion: Emotional reactions that the source provokes in listeners.

The perceived emotion is more abstract and objective. It is the emotion the source conveys. The perceived emotion of happy songs is always "happy". However, the induced emotion is more subjective. The same happy music may not necessarily induce happiness in the listener. In this study, we focus on the perceived emotion of soundscape because it is more objective.

2.3. Soundscape Taxonomy

Based on Fan et al. [23], we chose audios following Schafer's soundscape taxonomy [6]. Schafer's referential taxonomy is widely used for the classification of soundscapes. He indicates, "Sounds of the environment have referential meaning" [6]. Schafer grouped soundscape based on the identification of the source of the sound and the listening context rather than sonic characteristics. For example, in Schafer's taxonomy, a quiet forest is classed under quiet and silence, but without the listening context, it can be classified as natural sound. Theses categories are not mutually exclusive. Schafer's taxonomy is shown below.

Brown's taxonomy is based on the acoustic environment. The taxonomy is constructed regarding categories of places—indoor environment and outdoor environment. Within each environment, Brown el al. introduced four domains, including urban, rural, wilderness, and the underwater acoustic environment [16]. Within each domain, there are multiple subcategories. Comparing to Brown's taxonomy, Schafer's taxonomy has only six categories and considers both the identification of the source of the sound and the listening context. Thus, we choose the Schafer's taxonomy because of its simplicity and generality.

TABLE I. SCHAFER'S TAXONOMY

Categories	Examples
Natural sounds	Bird, thunder, rain, wind
Human sounds	Laugh, whisper, shouts
Sounds and society	Party, concert, store
Mechanical sounds	Engine, factory
Quiet and silence	Quiet park, silent forest
Sounds as indicators	Clock, church bells

2.4. Freesound Database

Though previous studies investigate the emotion recognition tasks regarding soundscape, none of the authors mentioned in Section 2.1 released their annotations and audio clips for further study. However, it is important to create an annotated soundscape dataset for SER and to set up a baseline for comparisons. In this section, we discuss the Freesound[1] database that we used for Emo-Soundscapes.

Unlike private soundscape recording databases, such as the World Soundscape Project[2] and Sound Ideas[3], Freesound is an online platform encouraging its 4 million users to upload and download sounds using a Creative Commons License[4]. The platform contains a large number of soundscape recordings that cover a broad range of subjects. Because Freesound is a crowd-sourced platform, care must be taken in the selection of sounds, as there is a wide variety of audio quality and indexing by semantic tags.

Leveraging the Freesound, Salamon et al. [18] released a database of short audio snippets, named UrbanSound8K. The dataset is comprised of 8732 clips of up to 4 seconds in duration extracted from recordings on Freesound. The authors define a taxonomy of urban sounds based on ten low-level sound sources including air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren and street music. However, the UrbanSound8K does not have emotion annotations, therefore, it cannot be directly used for SER.

3. Emo-Soundscapes Corpus

Three soundscape composers selected 600 audio clips following Schafer's taxonomy. There are 100 clips per category. Therefore, the Emo-Soundscapes corpus consists of 600 audio clips extracted from soundscape recordings from Freesound and 613 mixed sounds that are mixed using the clips. The items in the corpus are tagged with labels of perceived emotion derived from a listening experiment.

3.1. Audio Clips in Emo-Soundscapes

As this database is open for the research community, we need to ensure licensing allows for modification and distribution. We selected recordings from Freesound.org that permit such allowances. In the selection process, first, we automatically downloaded soundscape recordings that have higher ratings for quality and a greater number of downloads which indicates the higher standards. Then, we manually selected downloaded recordings to ensure the quality of recordings is varied but high. Next, the selected recordings were automatically segmented by BF-classifier that keeps audio regions with consistent characteristics [14-15]. The segmentation algorithm was designed based on perceptual categories, including background, foreground, and background with foreground sound. After the segmentation, we selected clips that can represent the name of the soundscape recording the most. That is, there is a high correlation with the sounds in the recording and semantic tags.

[1] http://freesound.org/
[2] https://www.sfu.ca/~truax/wsp.html
[3] https://www.sound-ideas.com/
[4] https://creativecommons.org/licenses/

The selected segments are cohesive and coherent enough so that the emotion is perceived as being stable throughout the clips. We keep 100 clips for each Schafer's category, which are 600 in total. This is the first subset of Emo-Soundscapes.

TABLE II. AUDIO CLIPS THAT ARE MIXED OF TWO CLIPS

Categories	Clip attenuation (A, B)	Number of Clips
Within Schafer's Categories	-6dB, -6dB	60
	-12dB, -6dB	60
	-6dB, -12dB	60
Between Schafer's Categories	-6dB, -6dB	75
	-12dB, -6dB	75
	-6dB, -12dB	75

TABLE III. AUDIO CLIPS THAT ARE MIXED OF THREE CLIPS

Categories	Clip attenuation (A, B, C)	Number of Clips
Within Schafer's Categories	Mixed sounds	48
Between Schafer's Categories	Mixed sounds	100

To study the impact of mixing and sound design on perceived emotion of soundscape recordings, we created a second subset of Emo-Soundscapes by manually mixing sounds of the first subset. As described in Tables II and III, each mix consists of mixing two or three selected audio clips selected within and between Schafer's soundscape categories and modulating and attenuation levels. Before the mixing, each audio clip is attenuated by either -6 dB or -12 dB. If it is within Schafer's categories, two audio clips that are from the same category are mixed. If it is between Schafer's categories, two audio clips that are from different categories are mixed.

Regarding the duration of each audio clip, previous SER studies indicated that 6 seconds is long enough for annotators to perceive the essence of the recording and form an opinion of both valence and arousal for a soundscape recording [7, 23]. Therefore, we set up parameters in BF-classifier to make the output audio clips are around 6 seconds. In Emo-Soundscapes, all the audio clips are 6 seconds (Mean: 6.17s, Std: 0.07s).

4. Data Annotation

4.1. Rating or Ranking

In the affective computing community, affective ratings instruments are tools for collecting annotations. Such tools are used in video emotion recognition [27], music emotion recognition [25-26], and soundscape emotion recognition [23]. However, recent research indicates that rating-based instruments are unreliable for collecting ground truth data [30]. Due to the different contextual situation and cultural backgrounds, the meaning of each level on a rating scale may change across annotators. Therefore, both ratings from different annotators and the same annotator may not be consistent [17, 31].

Ranking is an alternative approach for eliciting responses from subjects that circumvent many of these reliability problems [32]. Yang et al. [32] found that ranking-based approach simplifies the annotation process and enhances the reliability of the ground truth. This is essential in crowdsourcing because the simpler the task is, the better annotation that annotator will provide [13].

Yannakakis et al. [8] proposed *AffectRank*, a rank-based real-time annotation tool. They conducted a study where annotators use the FeelTrace [12], a continuous annotation tool, and the proposed *AffectRank*, a discrete rank-based annotation tool, on the arousal-valence 2D plane. Yannakakis et al. used Krippendorff's alpha to measure the inter-rater agreement of both rank-based annotation and rating-based annotation, and addressed that the inter-rater agreement of the ordinal data is significantly higher than using the nominal data.

Hence, we designed a ranking-based questionnaire where annotators made a pairwise comparison between two audio clips based on perceived valence and arousal.

4.2. Select Comparisons

Baveyes et al. found that collecting three annotations per comparison is a good compromise between the cost and the accuracy [13]. Therefore, we decided to collect three annotations for each pairwise comparison. To be more efficient, we used traditional quick sort algorithm to select comparisons [13]. For the first iteration, the algorithm randomly selects one audio clip as the pivot. Every other clip needs to be compared with the pivot so that the algorithm generates 1212 comparisons. For each comparison, after all the three annotations were collected, we determine the result to be the one that provided by at least two annotators. The first iteration ends up dividing the dataset into two subgroups. For the second iteration, the algorithm randomly selects one audio clip as a pivot in each subgroup and generates pairwise comparisons. We repeated this process until each audio clip received a rank of valence and a rank of arousal from 1 to 1213. The computational complexity of the quick sort algorithm is O(*NlogN*). The number of comparisons is determined by the selection of the pivot.

4.3. Experimental Design

To gain a large amount of data from people online, we used CrowdFlower[5], a crowdsourcing company, which allows users to access an online workforce of millions of people to label data. This method of data collection has been used for collecting affective data [24]. Our goal is to sort the 1213 audio clips based on valence and arousal independently. Therefore, we launched two tasks on CrowdFlower: one for valence and another for arousal.

At the beginning of the annotation process, subjects are provided with the terminology of arousal and valence. We use valence to describe perceived pleasantness of the sound. High valence describes a positive and pleasant percept, whereas low valence describes a negative, sad, unpleasant percept. For example, happiness is higher on the valence scale than

[5] https://www.crowdflower.com/

sadness; excitement is higher than boring, calm and serenity are higher than fear. We use arousal to describe eventfulness of the sound perceived. High arousal represents an eventful and energetic percept, whereas low arousal describes an uneventful and low energy percept. For example, chaotic is higher on the arousal scale than boring, excitement is higher than quiet, and energetic is higher than sleepiness.

To assist with subjects mapping of states of arousal to levels represented in the study, we applied the Self-Assessment Manikin [34] (see Figure 1.). The Self-Assessment Manikin [34] is a pictorial system used in experiments to represent emotional valence and arousal axes. Its non-verbal design makes it easy to use regardless of age, educational or cultural background. We modified the pictorial system by adding arrows to inform annotators that we were collecting perceived emotion.

We request the subject to follow a tutorial to get familiar with the annotation interface. Subjects are notified that they are required to use headphones to listen to the audio clips. We asked them to turn the volume up to a comfortable level given a test signal. The subject is then presented with a quiz, where five gold standard comparisons are provided. These comparisons are easily comparable regarding valence and arousal, which were carefully selected by experts. The annotator can continue to the task only if s/he achieves an 80% of accuracy in the quiz.

To enable tracking of the quality during the annotation process, we follow annotators' performance by inserting gold standard comparisons throughout the tasks. Similar to the comparisons in the quiz, these gold standard comparisons are easily comparable regarding valence/arousal. If annotators' answers are not the same as the default answer, they will be noticed by a pop our window. If annotators have strong reason to explain their answer, they can message the reason to us. Using the gold standard questions, we dismiss and do not remunerate any worker who answers over 20% wrong. This process also affects annotators' reputation on the CrowdFlower platform. Thus, workers have little interest in answering the questions at random.

Audio clips are played through an HTML5 audio player object (see Fig. 1.), which allows participants to listen repeatedly to a clip. After a subject had listened to both audio clips, the option to enter their response is presented in the form of an input button. The subject can pause the annotation process at any time and continue at a later stage. We implemented the pause feature for easing the fatigue that increases naturally during manual data annotation [17] in an attempt to minimize possible effects in our data collection. A CrowdFlower worker had to rank 5 pairs of audio clips before being paid US$0.05 and was able to exit the task at any time.

Because of the mechanism of the quick sort algorithm, in the first iteration of the experiment, subjects compare all the other 1213 audio clips to one pivot, which might cause learning issue because of the repetitions. To tackle the issue, we proposed a method called parallel ranking by which three corpora of audio clips are ranked in parallel. To account for the parallel ranking, we created another two corpora: Soundscape-Music Corpus and EMusic Corpus [28]. We used the same method (see Section 4.2) to select comparisons for ranking valence and arousal of audio clips in these two corpora. We end up having three pivots in the first iteration. This avoids the bias caused by repetition of only one pivot. In this paper, we do not analyze the results of other two corpora.

During the experiment, we display each pairwise comparison of audio samples to annotators until three annotations for each sample have been collected.

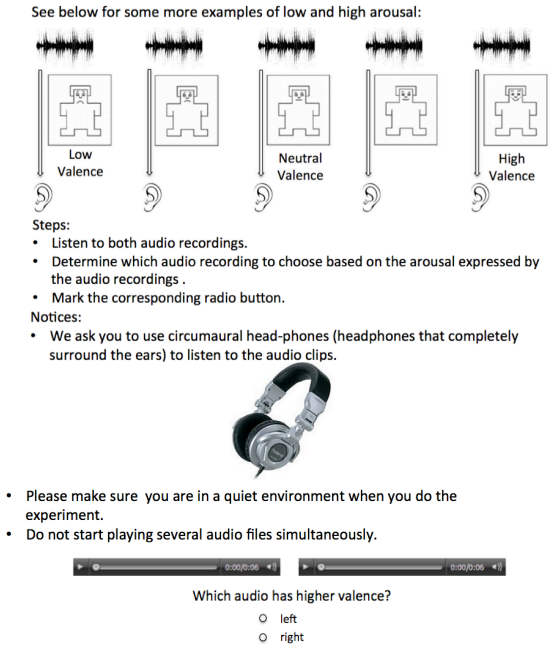

Figure. 1. Interface displayed to workers

5. Analysis of Annotations

1182 trusted annotators performed the task from 74 different countries. Most of the workers are Venezuelans (30.37%), Brazilian (6.69%), Serbian (6.35%), Russian (6.35%), Bosnians (4.82%), with the remaining workers globally dispersed. These 1182 trusted annotators had a gold accuracy of 92.18% for the quiz, and provided 69477 comparisons. Each trusted annotator took approximately 13s to perform a pairwise comparison. This also shows that annotators carefully listened to both soundscape clips.

5.1. Inter-Subject Reliability

Having created a corpus of tagged audio clips for the Emo-Soundscapes dataset, we wish to demonstrate participants' high level of agreement on the valence and arousal. Inter-subject reliability is used to measure whether independent subjects participate in an experiment reach the same conclusion despite the subjectivity of the task.

We used percent agreement and Krippendorff's alpha to measure the inter-subject reliability. Percent agreement is widely used and intuitive but overestimates inter-subject reliability since it does not take into account the agreement expected by chance. However, Krippendorff's alpha does take into account observed disagreement and expected disagreement but is sensitive to trait prevalence: it considers that annotators have a priori knowledge of the quantity of cases that should be distributed in each category [10]. The inter-subject reliabilities are displayed in Table IV. Regarding percent agreement, the value is between [0.0, 1.0]. As for Krippendorff's alpha, the value is in the range of [-1.0, 1.0]. A Krippendorff's alpha below 0 indicates that disagreements are

systematic and exceed what can be expected by chance; a value equal to 0 indicates the absence of reliability, and a value higher than 0 indicates an agreement between annotators (1 for perfect reliability) [10]. The percent agreement indicates that annotators agreed on 81.99% and 80.47% of comparisons. The value of Krippendorff's alpha ranges from 0.21 to 0.40, which indicates a fair agreement.

TABLE IV. INTER-ANNOTATOR RELIABILITY

Measures	Arousal	Valence
Percent Agreement	0.820	0.805
Krippendorff's alpha	0.289	0.233

6. A Baseline Model for SER

The goal of this section is to introduce a baseline model, similar to what can be found in the state of the art on similar emotion recognition tasks [13]. We also wish to define two separate protocols to assess such models performance using Emo-Soundscapes. These reproducible protocols will allow fair comparisons between future models and the baseline.

We convert the rankings to ratings to train regression models. This procedure has two assumptions. First, the distances between two successive rankings are equal. Second, the valence and arousal are in the range of [-1.0, 1.0]. We map the range of ranking values, 1 to 1213, to a corresponding rating range of 1.0 to -1.0, respectively.

6.1. Support Vector Regression

Support vector regression (SVR) has been widely used in affective computing tasks, such as music emotion recognition [35] and video emotion recognition [13]. 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. We built two independent SVRs to model arousal and valence. We selected the RBF kernel and used a grid search method to find the parameters C and gamma.

6.2. Feature Extraction and Selection

We investigate a large number of audio features, including energy, attack, fluctuation, spectral flatness, spectral rolloff, spectral kurtosis, summarized fluctuation, spectral flux entropy, Chromagram, MFCCs, novelty, roughness, brightness, regularity, zero cross rate, etc. Each audio clip is monophonic. The sample rate is 44100 Hz. We applied a 23ms Hanning window with 50% overlapping. Both MIRToolbox [36] and YAAFE [33] software package are used for the feature extraction.

We used the mean and standard deviation of each feature to represents signals as the long-term statistical distribution of local spectral features. This approach is well explored in the literature [19]. Using this method, we obtain a 122-dimension feature vector. Next, all features are normalized between [0, 1.0]. We then select specific features by examining the variance of a feature across the corpus. Those features whose variance is lower than a threshold are eliminated. The threshold of variance is 0.02, which is chosen as a heuristic value. We end up having a 39-dimension feature vector.

6.3. Protocols

We set up standard protocols for making possible comparisons of different models using the Emo-Soundscapes database. We used the *Mean Squared Error* (*MSE*) and R^2 to evaluate the performance of the model under each protocol

1) *Protocal A: Shuffle :* In this protocol, the annotated Emo-Soundscapes database is shuffled 10 times. Each time, 20% of the database is randomly selected for testing, and the remaining 80% is for training the SVR model. The R^2 and *MSE* are shown in Table V.

TABLE V. PERFORMANCE FOR PROTOCAL A

Metrics	Arousal		Valence	
	Mean	***Std***	***Mean***	***Std***
R^2	0.853	0.015	0.623	0.014
MSE	0.049	0.006	0.128	0.005

2) *Protocol B: Leave-One-Out:* In this protocol, one clip is selected for testing in each iteration while the rest is used for training. The number of iteration equals the number of data points in the dataset. After all the iterations, we calculated R^2 and *MSE*. Therefore, we obtained one R^2 and one *MSE*.

TABLE VI. PERFORMANCE FOR PROTOCAL B

Metrics	Arousal		Valence	
	Mean	***Std***	***Mean***	***Std***
R^2	0.855	N/A	0.629	N/A
MSE	0.048	N/A	0.124	N/A

The results from our experiment expressed with the R^2 statistic in Table V. and Table VI. are superior to previous research of modeling the arousal and valence of soundscapes. The previous model results using a multiple linear regression demonstrated 81.6% and 56.7% for arousal and valence, respectively [7]. Regarding protocol A, we can account for 85.3% and 62.3% of the variance for arousal and valence around the regression line, respectively. Regarding protocol B, we can account for 85.5% and 62.9% of the variance for arousal and valence around the regression line, respectively.

7. Conclusions and Future Work

We presented the Emo-Soundscapes dataset, an annotated soundscape database for soundscape emotion recognition. The 1213 6-second audio clips that make up the database are selected based on Schafer's soundscape taxonomy to be representative of soundscapes. The database is sorted along the valence and arousal axis through a crowdsourcing listening experiment, ensuring the quality of labels by tracking annotators' performance. To evaluate this and future models of SER on the Emo-Soundscapes dataset, we provided two protocols and demonstrated baseline SVR models.

The relative rankings along the dimensions of valence and arousal do not explain if the extreme cases with the lowest or highest ranks express extreme emotions. To describe the

emotional space that the Emo-Soundscapes dataset represents, we are designing an experiment to collect absolute ratings for valence and arousal of the clips and testing the correlation between the absolute ratings and the relative rankings.

References

[1] K. Kallinen, N. Ravaja, "Emotion Perceived and Emotion Felt: Same and Different," Musicae Scientiae, vol. 5, no. 1, pp. 123-147, 2006.

[2] B. Truax, Environmental Sound and its Relation to Human Emotion, Canadian Acoustics, vol. 44, no.3, 2016.

[3] B. Berglund, M. Nilsson, and O. Axelsson, "Soundscape Psychophysics in Place," in International Congress and Exhibition on Noise Control Engineering, 2007, pp. 3704–3712.

[4] L. Brocolini, L. Waks, C. Lavandier, C. Marquis- Favre, M. Quoy, and M. Lavandier, "Comparison between Multiple Linear Regressions and Artificial Neural Net works to Predict Urban Sound Quality," in International Congress on Acoustics, 2010, pp. 2121–2126.

[5] M. Thorogood and P. Pasquier, "Impress: A Machine Learning Approach to Soundscape Affect Classification for a Music Performance Environment," in International Conference on New Interfaces for Musical Expression, 2013, pp. 256–260.

[6] R. M. Schafer, The Soundscape: Our Sonic Environment and the Tuning of the World, Rochester, VT: Destiny Books, 1993.

[7] J. Fan, M. Thorogood, and P. Pasquier, "Automatic Recognition of Eventfulness and Pleasantness of Soundscape," in Audio Mostly, 2015.

[8] G. N. Yannakakis and H. P. Matínez, "Grounding Truth via Ordinal Annotation," in International Conference on Affective Computing and Intelligent Interaction, 2015.

[9] P. Lundén, O. Axelsson, M. Hurtig, "On Urban Soundscape Mapping: A Computer can Predict the Outcome of Soundscape Assessments," in International Congress and Exposition on Noise Control Engineering: Towards a Quieter Future, 2016, pp. 4725-4732.

[10] K. Krippendorff, "Estimating the Reliability, Systematic Error and Random Error of Interval Data," Educational and Psychological Measurement, vol. 30, no. 1, pp. 61–70, 1970.

[11] B. D. Barkana and I. Saricicek, "Environmental Noise Source Classification Using Neural Networks," in International Conference on Information Technology: New Generations, Las Vegas, NV, 2010, pp. 259-263.

[12] R. Cowie, E. DouglasCowie, S. Savvidou, E. McMahn, M. Sawey, and M. Schro¨der, "'FEELTRACE': An Instrument for Recording Perceived Emotion in Real Time," in ISCA Tutorial and Research Workshop on Speech and Emotion, 2000.

[13] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "LIRIS-ACCEDE: A Video Database for Affective Content Analysis," IEEE Transactions on Affective Computing, vol. 6, no. 1, pp. 43–55, 2015.

[14] M. Thorogood, J Fan, and P Pasquier. "Soundscape Audio Signal Classification and Segmentation Using Listeners Perception of Background and Foreground Sound," Journal of the Audio Engineering Society, vol. 64, no.7/8, pp. 484-492, 2016.

[15] M. Thorogood, J. Fan and, P. Pasquier, "BF-Classifier: Background/Foreground Classification and Segmentation of Soundscape Recordings," in Audio Mostly, 2015.

[16] A. L. Brown, J. A. Kang, T. Gjestland, "Towards Standardization in Soundscape Preference Assessment," Applied Acoustics. vol. 72, no. 6, pp. 387-392, 2011.

[17] A. Metallinou and S. Narayanan, "Annotation and Processing of Continuous Emotional Attributes: Challenges and Opportunities," in IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, 2013, pp. 1–8.

[18] J. Salamon, C. Jacoby, and J. P. Bello, "A Dataset and Taxonomy for Urban Sound Research," in ACM International Conference on Multimedia, 2014.

[19] J. J. Aucouturier and B. Defreville, "Sounds Like a Park: A Computational Technique to Recognize Soundscapes Holistically, Without Source Identification," in International Congress on Acoustics, 2007.

[20] M. Thorogood and P. Pasquier, "Computationally Generated Soundscapes with Audio Metaphor," in International Conference on Computational Creativity, 2013, pp 1-7.

[21] W. Ma and W. F. Thompson, "Human Emotions Track Changes in the Acoustic Environment," PNAS, vol.112, no. 47, pp. 14563-14568, 2015.

[22] A. Kawakami, K. Furukawa, K. Katahira and K. Okanoya, "Sad music induces pleasant emotion," Front Psychol vol. 4, no. 311, 2013.

[23] J. Fan, M. Thorogood, and P. Pasquier, "Automatic Soundscape Affect Recognition Using A Dimensional Approach," Journal of the Audio Engineering Society, vol. 64, no. 9, pp. 646-653, 2016.

[24] R. Morris and D. McDuff, "Crowdsourcing Techniques for Affective Computing," in Handbook of Affective Computing, Oxford University Press, 2014.

[25] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H.-H. Chen, "A Regression Approach to Music Emotion Recognition," IEEE Transactions on Audio, Speech and Language Processing , vol. 16, no. 2, pp. 448–457, 2008.

[26] F. Weninger, F. Eyben, and B. Schuller, "On-line Continuous-time Music Mood Regression with Deep Recurrent Neural Networks," in IEEE Internatioanl Conference Acoustics, Speech and Signal Processing, 2014.

[27] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi, "A Supervised Approach to Movie Emotion Tracking," in IEEE Internatioanl Conference Acoustics, Speech and Signal Processing, 2011, pp. 2376–2379.

[28] J. Fan, K. Tatar, M. Thorogood, and P. Pasquier, "Ranking-based Emotion Recognition for Experimental Music," in International Symposium on Music Information Retrieval, 2017.

[29] B. Davies, et al. "A Positive Soundscape Evaluation Tool," in Proceedings of the 8th European Conference on Noise Control-Abstracts. Institute of Acoustics, 2009.

[30] G. N. Yannakakis and H. P. Mart´ınez, "Ratings are Overrated!" Frontiers on Human-Media Interaction, 2015

[31] I. Sneddon, G. McKeown, M. McRorie, and T. Vukicevic, "Cross-Cultural Patterns in Dynamic Ratings of Positive and Negative Natural Emotional Behaviour," PLoS ONE, vol. 6, no. 2, p. e14679, 2011.

[32] Y.-H. Yang and H. Chen, "Ranking-based Emotion Recognition for Music Organization and Retrieval," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 762– 774, 2011.

[33] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard, "Yaafe, an Easy to Use and Efficient Audio Feature Extraction Software," in International Symposium on Music Information Retrieval, 2010, pp. 441–446.

[34] M. M. Bradley and P. J. Lang, "Measuring Emotion: the Self-Assessment Manikin and the Semantic Differential," Journal of Behavior Therapy and Experimental Psychiatry, vol. 25, no. 1, pp. 49–59, 1994.

[35] B.-J. Han, S. Rho, R. B. Dannenberg, and E. Hwang, "SMERS: Music Emotion Recognition Using Support Vector Regression," in International Conference on Music Information Retrieval, . 2009, page 651–656.

[36] O. Lartillot, P. Toiviainen, and T. Eerola, "A Matlab Toolbox for Music Information Retrieval," In: Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, Heidelberg,2008.