QUANTITATIVE ANALYSIS OF THE IMPACT OF MIXING ON PERCEIVED EMOTION OF SOUNDSCAPE RECORDINGS

Jianyu Fan
Simon Fraser University
jianyuf@sfu.ca

Miles Thorogood
Simon Fraser University
mthorogo@sfu.ca

Kıvanç Tatar
Simon Fraser University
ktatar@sfu.ca

Philippe Pasquier
Simon Fraser University
pasquier@sfu.ca

ABSTRACT

Sound designers routinely mix source soundscape recordings. Previous studies have shown that people agree with each other on the perceived valence and arousal for soundscape recordings. This study investigates whether we can compute the perceived emotion of the mixed-soundscape recordings based on the perceived emotion of source soundscape recordings. We discovered quantifiable trends in the effect of mixing on the perceived emotion of soundscape recordings. Regression analysis based on the trajectory observation resulted in coefficients with high R^2 values. We found that the change of loudness of a source soundscape recording had an influence on its weight on the perceived emotion of mixed-soundscape recordings. Our visual analysis of the center of mass data plots found the specific patterns of the perceived emotion of the source soundscape recordings that belong to different soundscape categories and the perceived emotion of the mix. We also found that when the difference in valence/arousal between two source soundscape recordings is larger than a given threshold, it is highly likely that the valence/arousal of the mix is in between the valence/arousal of two source soundscape recordings.

1. INTRODUCTION

Audio-based creative practices, such as sound design and soundscape composition, often use recordings to create musical works. A soundscape recording (or field recording) is "a recording of sounds at a given locale at a given time, obtained with one or more fixed or moving microphones" [1]. Often, sound designers select source soundscape recordings and carefully mix them together, which has a profound influence on meaning, significance, and perceived emotion. Together, the mixed-soundscape recordings create a rich, cohesive experience.

Previous studies demonstrate that people have a high level of agreement on the perceived emotion of source soundscapes recording [2]. It is also possible to build machine-learning models to predict the perceived emotion of soundscape recordings [3]. However, to our knowledge, no study has been presented regarding the effect of mixing on the perceived emotion of soundscape recordings.

In this study, we focus on the effect of mixing on the perceived emotion of soundscape recordings. We used Emo-Soundscapes, a dataset for soundscape emotion recognition that contains a group of annotated source soundscape recordings and annotated mixed-soundscape recordings [4]. The source soundscape recordings are selected following Schafer's taxonomy so as to cover the diversity of soundscapes as much as possible [5]. The perceived emotion is represented as the ranking of a two-dimensional vector of valence and arousal [6]. As identified by Thorogood and Pasquier [3], valence represents the pleasantness of a stimulus, which is used to report the perceived pleasantness of a soundscape recording. Arousal indicates the level of eventfulness.

Next, we convert the annotators' rankings to ratings and used regression models to determine the effect of mixing on the perceived emotion of soundscape recordings. Moreover, we analyzed the center of mass data plots to find the relationships between the perceived emotion of the mixed-soundscape recordings and perceived emotion of source soundscape recordings that are selected within Schafer's category. Last, we analyzed the likelihood of the perceived emotion of mixed-soundscape recordings lying between the perceived emotions of the two source soundscape recordings that are used for the mix.

2. RELATED WORKS

2.1 Taxonomy of Emotion and Affect Models

Emotional responses are subjective with people having possibly a different response to the same stimulus. According to previous studies [7], two types of emotions are involved when listening to soundscapes:

- *Perceived emotion*: emotions that are communicated and expressed by the source.
- *Induced emotion*: emotional reactions that the source provokes in an audience; it is what the audience feels from the source.

The perceived emotion is the emotion a source expresses. For example, 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 because of the internal interpretations and experiences of a listener. In this study, we focus on the perceived emotion of soundscapes.

2.2 Soundscape Emotion Studies

There are few studies that investigate modeling the perceived emotion of soundscape recordings. Thorogood and Pasquier [3] propose the Impress system, which uses a linear model predict the perceived pleasantness and eventfulness for soundscape recordings. Building on that research, Fan et al. [8] describe a corpus of audio files extracted from the Sound Ideas sound effects library and the World Soundscape Project library using an automatic segmentation algorithm [9]. A protocol maps audio features and expert user responses to soundscape recordings with stepwise linear regression models. Analysis of the protocol revealed a good fit of features for predicting eventfulness (R^2: 0.816) and pleasantness (R^2: 0.567). To further the design and evaluation of soundscape emotion research, Fan et al. designed a crowdsourcing listening experiment to collect ground truth annotations of 1,213 audio excerpts [4]. The authors used a ranking-based annotation method instead of rating-based methods [10]. The authors introduced baseline models and defined protocols to assess such models performance. The results of using support vector regressions are human competitive (eventfulness, R^2: 0.855; pleasantness, R^2: 0.629).

Lundén et al. [11] investigated another method of predicting the outcome of the soundscape assessment based on acoustic features. The authors extracted 120 excerpts (30 seconds) from 77 audio recordings (15 min) and asked 33 participants to move an icon into a 2D space to assess the pleasantness and eventfulness of soundscapes. The authors used the bag-of-frames approach [12] to represent the audio features. Then, they used a Gaussian mixture model to cluster the aggregate of features and used the resulting dissimilarity matrix to train two separate support vector regression models to predict soundscapes' pleasantness and eventfulness. The result indicates that the Mel-frequency cepstral coefficients (MFCCs) provide the strongest prediction for both eventfulness (R^2: 0.83) and pleasantness (R^2: 0.74).

2.3 Soundscape Taxonomy

Based on Fan et al. [2], we selected sound excerpts following Murray Schafer's soundscape taxonomy [5]. Schafer's referential taxonomy is widely used for the classification of soundscapes. Table 1 shows 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 part, silent forest
Sounds as indicators	Clock, church bells

Table 1. Murray Schafer's Taxonomy [2, 5].

3. DATASET

We use the Emo-Soundscapes dataset curated by Fan et al. [4]. Emo-soundscapes is a soundscape recording database for soundscape emotion recognition composed of 1213 soundscape excerpts downloaded from Freesound.org. The dataset also contains rankings of the perceived emotion of 1213 6-seconds long soundscape recordings in the 2D valence-arousal space. Fan et al. conducted a crowdsourcing study where 1182 trusted annotators from 74 countries did pairwise comparisons of all soundscape experts regarding perceived valence and perceived arousal. Each pair has been annotated by three annotators. Based on the pairwise comparisons, the database is sorted along the valence and arousal axis.

There are two sets in the Emo-Soundscapes dataset. The first set has 600 excerpts that are selected following Schafer's taxonomy [5] with 100 excerpts per category. The second set contains 613 excerpts that are mixed from the first set. We used the second subset in this study. As described in Tables 2, each mix consists of two or three audio excerpts selected within and between Schafer's soundscape categories. In this paper, we only focus on the mixed-soundscape recordings that are composed of two source excerpts. Before the mixing, each source excerpt is digitally attenuated by either −6 dB or −12 dB. We applied these attenuation levels to examine the influence of loudness changes of sources on the perceived emotion of the mix.

To examine mixing of different types of sounds, we mix excerpts as pairs both from within and between Schafer's categories. Table 2. shows the treatment given to these mixed pairs.

Categories	Excerpt Attenuation (A, B)		Number of Excerpts
Within Soundscape Categories	−6dB	−6dB	60
	−12dB	−6dB	60
	−6dB	−12dB	60
Between Soundscape Categories	−6dB	−6dB	75
	−12dB	−6dB	75
	−6dB	−12dB	75

Table 2. Mixed Audio Excerpts (Two Excerpts) [4].

4. RESULTS AND ANALYSIS

4.1 Regression Analysis

We performed regression analysis on the data. We aim not to maximize absolute performance, but rather to study the relationship between the perceived emotion of two source soundscape recordings and the perceived emotion of the mix, and analyze the influence of the loudness of one source soundscape recording on its weight for the perceived emotion of the mixed-soundscape recording.

We convert the rankings to ratings by mapping the range of ranking values, 1 to 1213, to a range of rating values, 1.0 to −1.0, so that the highest ranked excerpt has the highest rating. 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 assumed that two dimensions are independent, and we hypothesized a linear relationship where soundscape recording A and soundscape recording B combine to yield a mixed-soundscape recording. The relationship is as follows:

$$MixedSound_{Affect} = \alpha \cdot A_{Affect} + \beta \cdot B_{Affect} \quad (1)$$

The subscript "*Affect*" is the dimension (arousal/valence) of emotion. A_{Affect} is the value of affect of soundscape recording A. B_{Affect} is the value of affect of soundscape recording B. $MixedSound_{Affect}$ is the value of affect of the mix. α and β are the weights optimized by the regression model, respectively.

We use the coefficient of determination (R^2) to evaluate the performance of our models. R^2 describes the ratio of the variance of the model's predictions to the total variance. The closer R^2 is to 1, the better the performance of the model. We obtained the R^2 based on 10-fold cross-validation. The results are summarized below.

Dimension	Excerpt Attenuation (A, B)	α	β	R^2
Arousal	−6dB, −6dB	0.597	0.518	0.751
	−12dB, −6dB	0.429	0.612	0.716
	−6dB, −12dB	0.668	0.276	0.724
Valence	−6dB, −6dB	0.535	0.359	0.647
	−12dB, −6dB	0.291	0.590	0.444
	−6dB, −12dB	0.576	0.288	0.526

Table 3. Regression results of predicting the perceived emotion of the mix (Within Schafer's categories).

Dimension	Excerpt Attenuation (A, B)	α	β	R^2
Arousal	−6dB, −6dB	0.667	0.476	0.660
	−12dB, −6dB	0.483	0.577	0.659
	−6dB, −12dB	0.786	0.353	0.739
Valence	−6dB, −6dB	0.527	0.401	0.216
	−12dB, −6dB	0.496	0.521	0.418
	−6dB, −12dB	0.771	0.271	0.514

Table 4. Regression results of predicting the perceived emotion of the mix (Between Schafer's categories)

From Tables 3 and 4, we can find correlations between change of loudness and change of weight. When the loudness of soundscape recording A goes from −6 dB to −12 dB, its weight goes down as well. Meanwhile, even though the loudness of the soundscape recording B stays at −6 dB, the weight of soundscape recording B goes up. The same pattern can be found when the loudness of the soundscape recording B goes down and the loudness of the soundscape recording A stay still. The correlation indicates that the loudness of a soundscape recording has a strong influence on its weight for the perceived emotion of the mixed-soundscape recording.

Comparing the performance of regression models for valence and arousal, we find that the prediction of arousal is more accurately modeled than the valence, confirming the findings in Fan et al. [2].

In general, the results of predicting the valence and arousal of mixed sound within soundscape categories are better than the results of predicting the mixed sound between soundscape categories. Specifically, when both excerpts are attenuated by −6 dB, the results of predicting the valence and arousal of mixed sound within soundscape categories are significantly better than the results of predicting the valence and arousal of mixed sound between soundscape categories. We believe this is because the texture of the mixed-soundscape recordings within the same categories is more homogenous. When mixing them together, it introduces less contrast so that the perceived emotion is more predictable.

4.2 Center of Mass Plots of Mixing Two Source Soundscape Recordings (Within Categories)

In Figures 1–4, we illustrate the center of mass of two source soundscape recordings for visual analysis of the all the data points of mixed sound within soundscape categories. Each figure has 10 mixed-soundscape recordings (combinations of 5 attenuated source recordings) of one soundscape category showing one attenuation condition; adding up to 180 mixed-soundscape recordings.

On these center of mass charts, each green dot represents a source soundscape recording; red stars represent mixed-soundscape recordings. Finally, the influence of the mixing is shown as the trajectory through data points from a source soundscape recording (green circle) through a mixed-soundscape recording (red star) to another source soundscape recording (green circle).

Figure 1 shows the center of mass plots of mixed sound within the categories of "natural sounds" and "quiet and silence." From Figure 1, we see that when source soundscape recordings have a low level of arousal and a high level of valence, it is the same for the mixed-soundscape recording.

Figure 3 shows the center of mass plots of mixed sound within the categories of "mechanical sounds." It indicates that when the source soundscape recordings have a high level of arousal and a low level of valence, it is highly likely the case for the mix.

Figure 2 shows the center of mass plots of mixed sound within the categories of "human sounds" and "sounds and society." In comparison to "natural sounds," "quiet and silence," and "mechanical sounds," the distribution of soundscape recordings on the two-dimensional emotion space is more scattered and the valence/arousal values are more diverse.

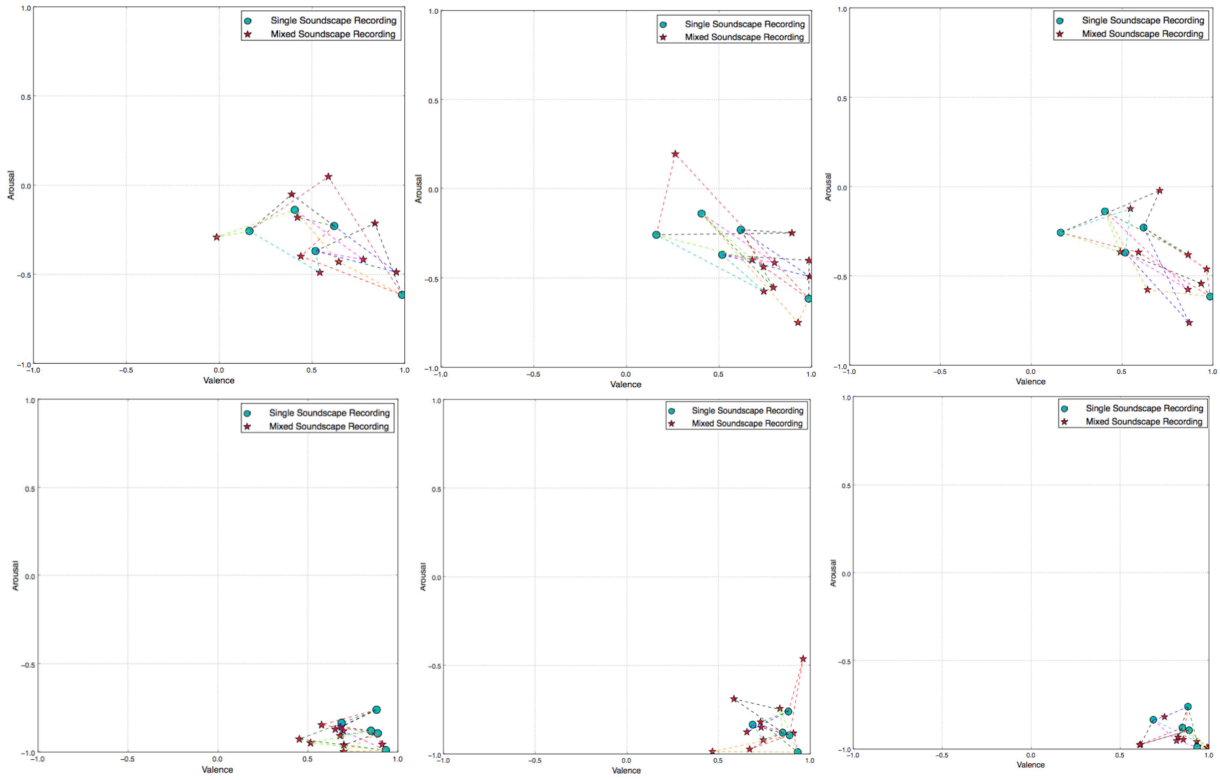

Figure 1. Center of mass of mixed "natural sounds" (Top) and mixed "quiet and silence" (Bottom). The left column shows the attenuation of −6 dB and −6 dB. The middle shows the attenuation of −6 dB and −12 dB. The right column shows the attenuation of −12 dB and −6 dB. (Arousal is the Y-axis, Valence is the X-axis)

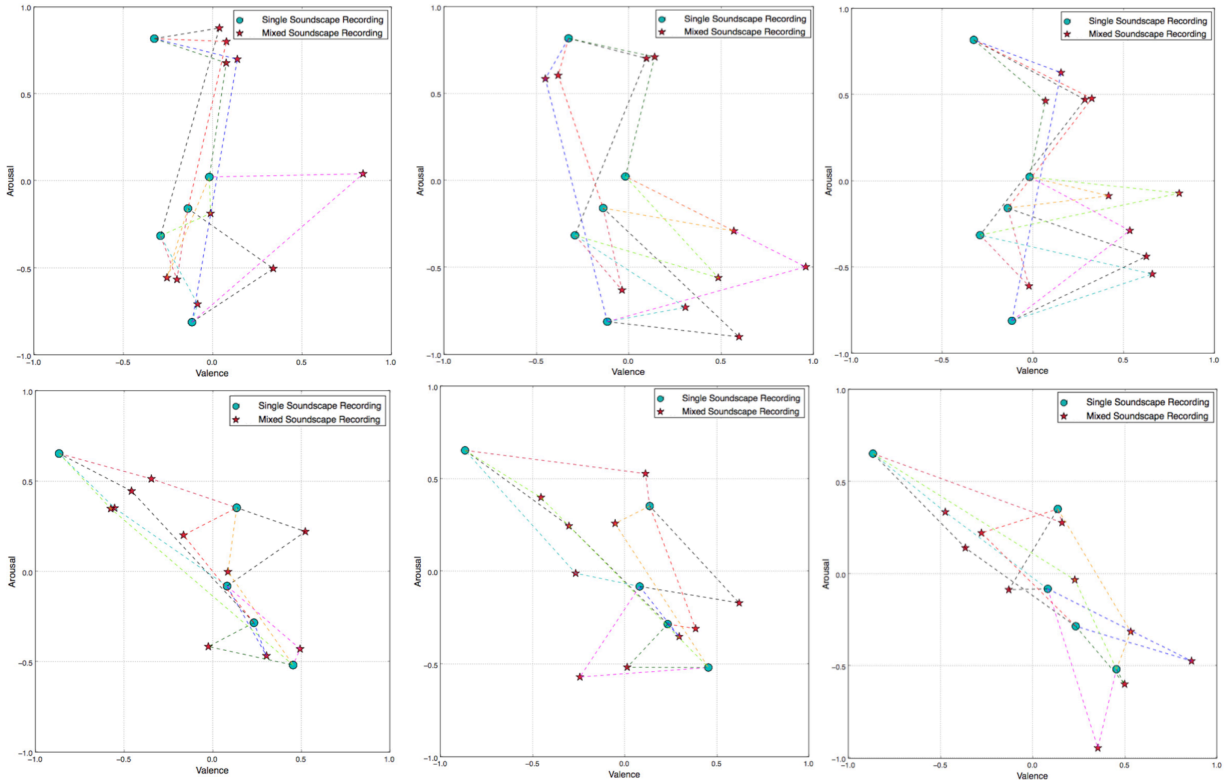

Figure 2. Center of mass of mixed "human sounds" (Top) and mixed "sounds and society" (Bottom). The left column shows the attenuation of −6 dB and −6 dB. The middle shows the attenuation of −6 dB and −12 dB. The right column shows the attenuation of −12 dB and −6 dB. (Arousal is the Y-axis, Valence is the X-axis)

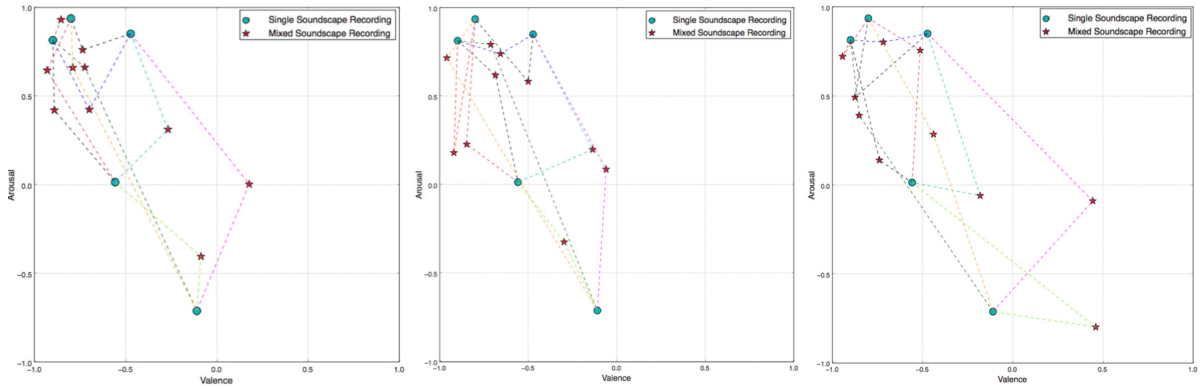

Figure 3. Center of mass of mixed "mechanical sounds". The left chart shows the attenuation of −6 dB and −6 dB. The middle chart shows the attenuation of −6 dB and −12 dB. The right chart shows the attenuation of −12 dB and −6 dB.

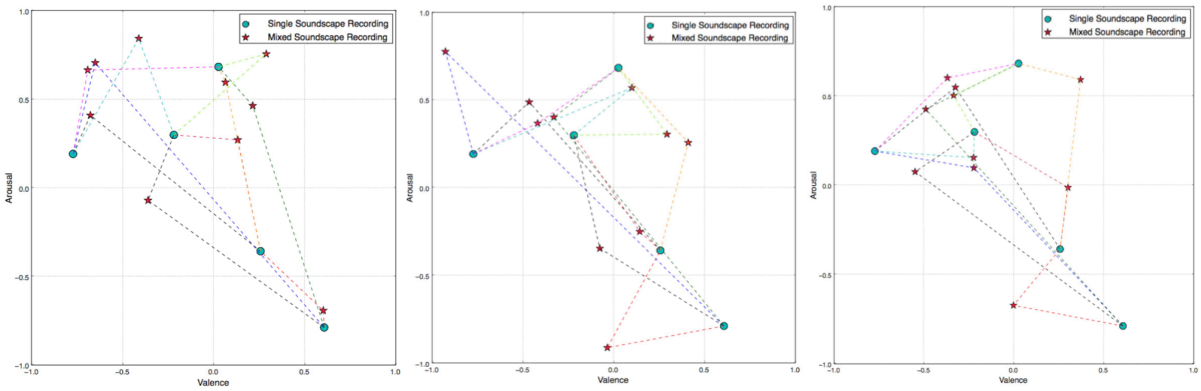

Figure 4. Center of mass of mixed "sounds as indicators". The left chart shows the attenuation of −6 dB and −6 dB. The middle chart shows the attenuation of −6 dB and −12 dB. The right chart shows the attenuation of −12 dB and −6 dB.

From Figure 2, we can find that the soundscape recordings that have high arousal or low valence (located in the third quadrant) usually have a bigger impact on the valence and arousal of the mixed-soundscape recording, especially when the difference between two source soundscape recordings regarding valence/arousal is large. Mixed soundscapes within the "human sounds" category (−6 dB, −6 dB), for instance, are an example of the hypothesized effect of one source soundscape recording influencing the emotion of the mix. One possible explanation is that the emotion of the high-arousal and low-valence soundscape recordings drew listeners' attention from the mixed-soundscape recordings.

Moreover, for "human sounds" and "sounds and society," note that the center of mass for the mixed-soundscape recording is situated on a path between two source soundscape recordings. With only a few exceptions, the mixed-soundscape recording's valence/arousal ratings lie on a smooth trajectory from one source's rating to another source's rating.

Regarding "sounds as indicators," the relationship between mixed-soundscape recordings' ratings and source soundscape recordings' ratings is more complex. Figure 4 shows the center of mass plots for "sounds as indicators." The fact that "sounds as indicators" is difficult to model confirms the finding in the previous study [2]. "Indicators serve as clues that something more fundamental or complicated is happening than what is measured by them" [13]. They carry strong semantic information, which is important for perceived emotions.

When we converted rankings to ratings, we also found the following. When the difference between arousal of two soundscape recordings' ratings is larger than a given threshold (0.5), it is highly likely that the rating of arousal of the mixed-soundscape recording lies in between the rating of arousal of two source soundscape recordings. This is also true for valence. This corresponds to the finding that the mixed-soundscape recording occurs on a trajectory from one soundscape recording to another one. Tables 5 and 6 show the probability of occurrence of the above statement for mixed sound within soundscape categories and mixed sound between soundscape categories.

We also tested the probability of occurrence of the above statement when we removed the soundscape recordings that belong to "sounds as indicators." Table 5 shows the results, which indicate that the probability increases when we removed "sounds as indicators." This means the patterns in "sounds as indicators" are more complex. A similar explanation for this is that the semantic information increases the complexity of modeling this category.

Dimension	Excerpt Attenuation (A, B)	Probability	Probability (Not include "sounds as indicators")
Arousal	−6dB, −6dB	80.00%	89.47%
	−12dB, −6dB	84.00%	89.47%
	−6dB, −12dB	84.00%	89.47%
Valence	−6dB, −6dB	92.86%	100.00%
	−12dB, −6dB	71.43%	87.50%
	−6dB, −12dB	78.57%	87.50%

Table 5. The probability that the rating of the perceived emotion of the mix lies between the ratings of the perceived emotion of sources that are selected within Schafer's categories.

Dimension	Excerpt Attenuation (A, B)	Probability
Arousal	−6dB, −6dB	82.67%
	−12dB, −6dB	88.24%
	−6dB, −12dB	84.31%
Valence	−6dB, −6dB	75.56%
	−12dB, −6dB	68.89%
	−6dB, −12dB	73.33%

Table 6. The probability that the rating of the perceived emotion of the mix lies between the ratings of the perceived emotion of sources that are selected between Schafer's categories.

5. CONCLUSION

We analyzed the relationship between the perceived emotion of mixed-soundscape recordings and source soundscape recordings that are used for mixing. Our analysis shows that there is a correlation between the loudness of a source soundscape recording and its weight that contributes to the perceived emotion of the mix. From the center of mass charts, we found the consistency of perceived emotion of source soundscape recordings and the perceived emotion of mixed-soundscape recording under certain circumstances. Moreover, when the difference of perceived emotion is larger than a given threshold, we found that there is a high likelihood that the perceived emotion of mixed-soundscape recordings lies between the perceived emotions of the two source soundscape recordings that are used for the mix.

The aim of this research is to move toward a formal definition of complex sound design mixing decisions. In doing so, we plan to investigate computational tools that provide suggestions and automate different sound design tasks. One application of this work is in the development of emotion aware digital audio workstations in the production of game sound, film sound, and virtual reality audio environments. We imagine a further integration of such technology in autonomous sound design systems embedded in game engines responding to players' cues to evoke truly personalized contextual experiences.

6. REFERENCES

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

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

[3] M. Thorogood and P. Pasquier, "Impress: A Machine Learning Approach to Soundscape Affect Classification for a Music Performance Environment," in Proc. Int. Conf. on New Interfaces for Musical Expression (NIME2013), 2013, pp. 256–260.

[4] J. Fan, M. Thorogood, and P. Pasquier, "Emo-Soundscapes: A Dataset for Soundscape Emotion Recognition," in Proc. Int. Conf. on Affective Computing and Intelligent Interaction (ACII2017), 2017.

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

[6] J. A. Russell, A. Weiss and G. A. Mendelssohn, "Affect Grid: A Single-Item Scale of Pleasure and Arousal," in J. Personality and Soc. Psych., 1989, vol. 57, no. 3, pp. 493–502.

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

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

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

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

[11] 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.

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

[13] 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.