Modeling Human Core Affect Evaluation of Soundscape Quality with Physical Properties

Bachelor's Thesis

Ben Wolf, s1879227, b.j.wolf.1@student.rug.nl,
Tjeerd Andringa, t.c.andringa@rug.nl,
University of Groningen, Department of Artificial Intelligence

Abstract: The sounds around us form a sonic environment. The perceived sonic environment, or soundscape, affects our moods in yet unclear ways. The traditional model for soundscape quality is derived from sound pressure levels only and has a poor performance for classifying soundscape quality in terms of core affect. Core affect allows us to describe the shifts in our moods along two axes: pleasantness and eventfulness. It has been shown that soundscapes are evaluated quite uniformly in terms of core affect. Using physical properties – or features – of soundscapes, we set to model human core affect evaluation with an artificial neural network. The results show that our trained model offers insights in useful advances for automated classification.

1. Introduction

Our environment, and the way in which we perceive it, influences our emotional state. We all have experienced that cloudy or rainy weather can ruin your day and a sunny morning can brighten it. Intuitively, one would attribute these percepts to what we see, but another key element in experiencing our environment has to do with the perception of one's auditory environment, i.e. one's soundscape.

The phenomenon, soundscape, is a term first brought to life by Murray Schafer (1977) and can be described as the perceived auditory environment. Just like a landscape is composed of many landmarks, a soundscape can be composed of many different sources of sound. These can be natural sources, like chirping birds or artificial sources like music, speech and mechanical sounds.

Traditional research into the effects of soundscapes has primarily been focusing on the negative aspects of soundscapes in our environment: sound pollution. A commonly used determiner is the L_{den} sound pressure level scale as used by the European Union and is calculated with regard to the sound pressure levels during the day, evening and night. However, researchers have shown (Halpern et al, 1986) that L_{den} or any equivalent guidelines using sound pressure levels are not sufficient indicators of how one perceives a soundscape. A recording of forest sounds or a plane passing over can be perceived equally loud, but affect ones mood in different ways.

In order to cope with these different effects of soundscapes, Russel's (2003) theory of Core Affect provides a scientific approach to these mood changes. Core affect refers to an integral blend of feelings of valence (ranging from pleasure to displeasure) and arousal (ranging from eventful to uneventful), which are considered to reflect the most basic dimensions characterizing emotional feelings (Kuppens et al, 2012).

We can use core affect to describe moods in a similar way we describe colors. All colors are grouped in discrete categories: red, blue, yellow etc. The underlying physics of colors is described by wavelengths, on a continuous scale, ranging from ultraviolet to infrared. Likewise, our moods are described as distinct and discrete (Kuppens et al, 2012). For instance excited and bored could be perceived as the discrete categories, where Core Affect is an continuous underlying aspect of mood. Yik (2011) states, however that simply labelling core affect is not enough. Emotions, like anger and fear, could have identical states of Core Affect. Yik explains that both the core affect state and the shifts in core affect contribute to our labelled mood.

Still, the structure of Core Affect is seen most clearly in all kinds of studies in which moods are recorded via questionnaires or self-reports. The

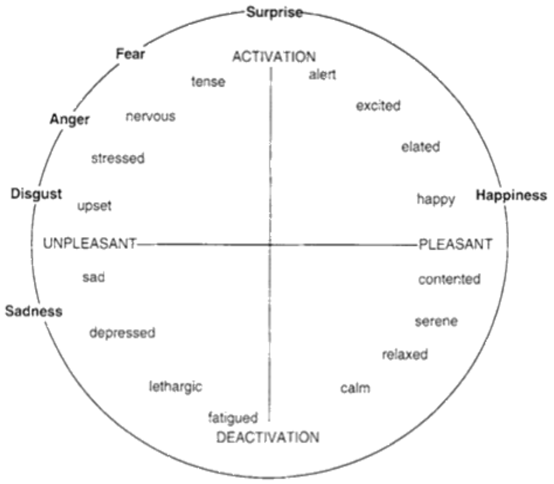

Figure 1: the Core Affect circumplex structure as proposed by Russel. The inner circle is a schematic mapping of words. The outer circle denotes a few prototypic moods. (Russell & Barrett, 1999).

descriptive structure derived from such studies can be classified into the Core Affect circumplex (Figure1).

Lindborgs (2010) attempt to classify soundscapes in terms of mood states for adults, have yielded promising results. Using principal component analysis, Lindborg found that these mood states can be fitted on a 2-dimensional plane, Mass and Variability Focus, which are closely related to Core Affect. An automated attempt to model subjective soundscape quality (Yu, 2009), proved that neural networks are able to predict subjective measures. Broers (2011) and Kangur (2011) present some novel features that could be used in automated classification of Core Affect.

This research focusses on creating a model, using physical features of soundscapes, which can predict Core Affect evaluation. We set to find out which features of soundscapes have predictive value with regards to the Core Affect evaluation and whether such a model is able to successfully predict Core Affect evaluation.

2. Methods

In this research, one dataset of tagged indoor sound fragments is used to train and test a two layer perceptron, using the leave-one-out method.

2.1. Dataset

The dataset consists of 56 indoor recordings with a mean length of 12 seconds. For each of these 56 soundscapes, a mean Core Affect score is obtained. 17 participants are asked via a questionnaire how much describers (as seen in the circumplex, Figure 1) match the soundscape they are listening to (Kangur, 2011). From this, the mean evaluation for each soundscape is calculated and fed to the network.

Some sound fragments are recorded at the Bernouilliborg on the University's campus, others are recorded at a facility caring for and supporting the mentally challenged. For a full description on the sound clips, see Appendix A.

2.2. Physical features

Loudness and the mean energy weighted frequency are simple features that can be derived quickly from soundscapes without a semantic translation. Other components found to be predictive are more complex (Broers, 2011). These include but are not limited to music, arguing people and normal speech. With the use of cochleograms, the characteristics of the different components were analyzed and translated into simple features.

Using the software package CPSP[1], created by the Auditory Cognition Group at the University of Groningen, we first convert the

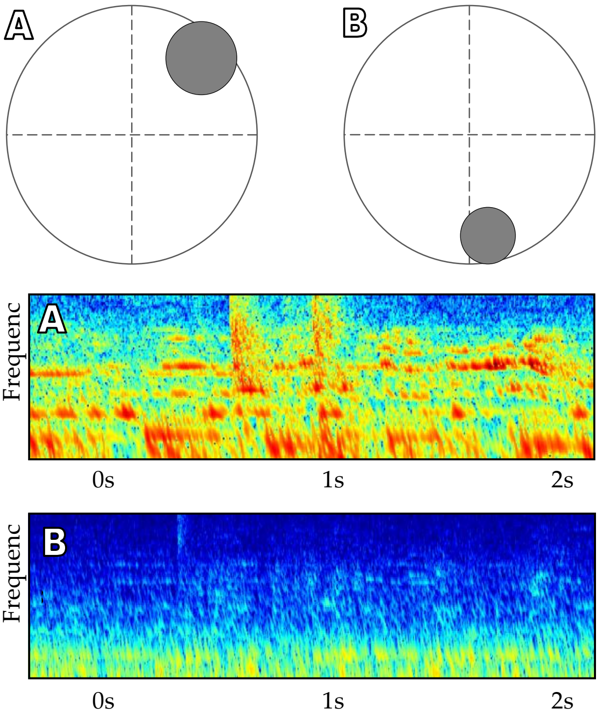

Figure 2: the Core Affect scores of two soundscapes, A and B, with an excerpt of 2 seconds of the normalized cochleogram (60 dB)

[1] http://www.ai.rug.nl/acg/cpsp/

sound fragments into cochleograms. The cochleogram indicates the distribution of source energy as function of time and frequency and therefore gives us more information about the sound fragment.

Then we normalize the input with a dynamic range of 60 dB, so that the overall loudness of a soundscape can't influence other features. Some features are described using these 60 dB levels, for instance the 22nd dB level refers to the 22nd dB level of the normalized dynamic range [0-60]. Features can be calculated from both the normalized and original cochleograms.

For instance, two sound fragments are shown in Figure 2. The first fragment, A, is from a soundscape rated as both pleasant and active and has a lot of high energy. In A, music is playing (seen in the rhythmic lower frequencies), someone slaps two times on a table (the two vertical structures) and some people are talking as seen by the harmonic complexes. The second fragment, B, is recorded in an empty hall where someone walks by, it has no interesting physical features and is rated as not active.

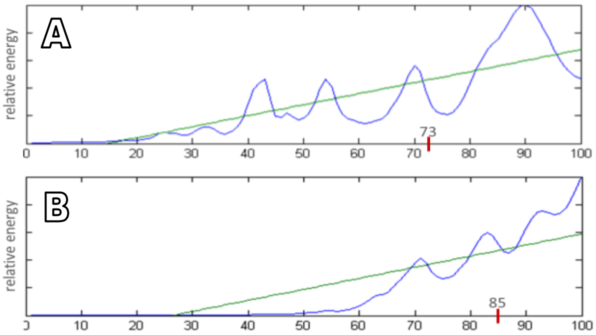

Figure 3: the relative energy compared with the frequency. The high frequency bands denote a low pitch and the low frequency bands denote a high pitch. The centroid of relative energy for A is frequency band 73 and the centroid for B is 85 (feature 5). The ratio that surpass the fitted line is a measure for variability (feature 8).

The first fragment seems to have relatively more energy in its higher frequencies. When we look at the centroid of the relative energy in terms of frequency bands (Figure 3) , the first fragment has a significant lower centroid of

Name	Description
1. mdB	The mean dB value after normalizing the cochleogram. A high value could indicate an active environment. *mean(CG) / dynamic range*
2. pdB	The peak of the dB levels before normalizing. *difference(mdB , original peak dB) / dynamic range*
3. rmdB	The relative mean dB before normalizing, compared to the loudest soundscape. *difference (original mean dB, loudest original dB of dataset) / dynamic range*
4. g31dB	Ratio dB > relative 31st dB level. Anything louder than the 31st dB is considered foreground material. *sum(CG > (max(CG) – 29)) / size(CG)*
5. l22dB	Ratio dB < relative 22nd dB level. Anything with a lower dB than the 22nd dB is background sound. *sum(CG < (max(CG) – 38)) / size(CG)*
6. cFreqdB	Centroid frequency band of the relative dB. Denotes the typical frequency of a soundscape. *sum(summed energy per frequency band × frequency band) / (n frequency bands × total energy)*
7. cFreqP	Centroid frequency band for pulses. High pulses could indicate speech. *sum(summed pulse filtered energy per frequency band × frequency band) / (n frequency bands × total energy)*
8. gFreqdB	A measure for variability. The ratio of frequencies that surpass the fitted frequency-energy line as seen in figure 3 *sum(energy of frequency bands > fitted line for energy over frequency) / size (*$CG_{frequency}$*)*
9. dCTimedB	The relative time between the centroid of energy over time and the half of the fragment. *difference(mean(summed energy per time unit × time unit) / length (*CG_{time}*) , .5)*
10. gTime	The ratio of time units that surpass the fitted time-energy line. *sum(energy of time units > fitted line for energy over time) / size(CG)*
11. TimeP	Where the most pulse components are located over time. *sum(pulse filtered energy of time units × time unit) / (n time units × total energy)*
12. Tslope	The slope of the fitted time-energy line. A measure for decreasing or increasing energy. *slope × n time units / dynrange*

Table 1: description of the features and how to calculate them. CG stands for the Cochleogram structure, with time and frequency as its dimensions and energy (dB) for values.

relative energy. This is a simple feature that can be fed to the neural network. The other used features are described in Table 1.

2.3. Neural network

The type of neural network used in this experiment is a two layered perceptron. The network consists of $n + 1$ input nodes, where n is the number of features and 1 the bias input. It furthermore has one hidden layer with m hidden nodes and a bias node, so $m + 1$. The hidden layer connects to two output nodes, one for pleasantness and one for eventfulness.

Every input node (and the input bias node) is connected to every hidden node, using – what in the model is called – *input weights*. The value of these weights range from -1 to 1 and are stored in a matrix.

Every hidden node (including the 2nd bias node) is connected to the two output nodes via the *hidden weights*. These *hidden weights* are also stored in a matrix.

On the very first epoch, the networks input is the translation of the first sound clip into twelve features. These values of these inputs range from 0 to 1. For every hidden node, the summed weighted activation is calculated by multiplying the (initial random) *input weights* matrix with the input vector. This yields the *hidden activation.* With a sigmoid function, the *hidden activation* is normalized within the range -1 to 1 and is now called *hidden output*. The *hidden weights* matrix is then multiplied with the *hidden output* vector to yield the *output activation.* The normalized *output activation* is called the *output output* and are in fact the coordinates of the prediction of the model, e.g. (-1, -1) is both uneventful and unpleasant whereas (1, 1) is considered eventful and pleasant.

After calculating this prediction, a simple back propagation algorithm calculates the relative error of each node – using the weights – and adjusts the weights accordingly. The maximum error per example is 4, 2 in the pleasantness dimension and 2 in the eventfulness dimension. When all sound clips are presented to the network in this way, the first epoch is over.

A two layer perceptron has some parameters that can influence the networks performance. These are: the number of hidden nodes and the learning rate (how much the relative error affects the weights).

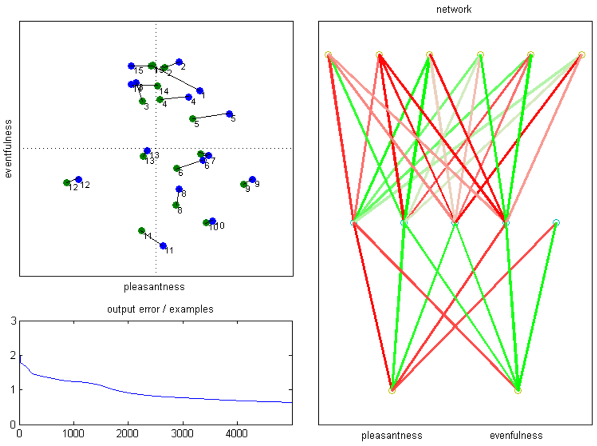

Figure 4: neural network visualization. The first graph shows the predicted (blue) and desired output (green) of the network. The bottom graph shows the summed error per example. The right graph is a visualization of an example network, with five input nodes , four hidden nodes, two outputs and both the input bias and hidden bias. The green lines represent positive weights and the red negative weights. The intensity of the color reflects the strength of the weight.

To obtain the right values for these parameters, a parameter sweep is done with a subset of the dataset. The number of nodes ranged from 4 to 10 and the learning rate varied from 0.5 to 1.0, with increments of 0.1. For each combination, the number of epochs is stored in which the network stops learning, i.e. the decrease in error is smaller than $\varepsilon = 0.001$ per prediction. The outcome of this parameter sweep is presented in the results section.

In the actual experiment, the network is set to stop learning when the difference in error between epochs is lower than some value ε. When the network is set to stop with $\varepsilon = 0$, the model could be 'over fitted' and unable to cope with new inputs. Choosing a small value for ε however prevents this from happening. In the experiment, $\varepsilon = 0.01$ *per example* or *0.55* is chosen, based on the learning curve (Figure 4) of the network.

2.4. Experiment

Using the leave-one-out method, 56 models are derived from training 56 neural networks with 55 examples. Every created model has to predict the Core Affect score of the one sound fragment that was left out of the training phase.

To determine which features have predictive value with regards to Core Affect scores, the features do not only have to explain some of the variance in the data, but these features must also consistently indicate a certain value of pleasantness and eventfulness. In this case, regardless of the model, the same feature should predominantly point in de same general direction when it comes to Core Affect evaluation. Should the selected features map consistently in these models, one can assume that these features have predictive value with regards to Core Affect.

Mapping the first feature with a model is done by feeding the model a vector with the first value being 1 and the rest 0. The prediction of the model is on this input in fact the location of the feature on the Core Affect circumplex. This is done for every feature and every model.

To determine whether the network is able to successfully predict the Core Affect scores, the ratio of correctly predicted scores is calculated. The Core Affect scores have a relatively high standard deviation in both pleasantness and eventfulness (Kangur, 2011; Lindborg, 2010), therefore a prediction is considered correct when *error < 0.5*.

3. Results

3.1. Parameter sweep

The network is trained on 13 examples with the learning rate and number of hidden nodes changing. Table 2 shows that the parameters influence the number of epochs required to train a model in an expected way.

	.5	.6	.7	.8	.9	1.0
4	6834	6253	5634	4031	3673	3412
5	6453	6021	5382	3852	3415	3115
6	5793	5012	4723	3135	3049	2842
7	5012	4214	3143	2715	2732	2681
8	4846	3816	3001	*2689*	2704	2666
9	4613	3431	2834	2652	2611	3412
10	4782	3150	2841	2643	2639	2524

Table 2: parameter sweep over the *learning rate (horizontal)* and *number of hidden nodes (vertical)*.

In general, the more hidden nodes and the higher the learning curve, the faster the network can train. For the experiment, the learning rate was set on .8 and the number of hidden nodes on 7, since the number of epochs seems to stall there, while calculation time increases with more nodes.

3.2. Features

When comparing the features placed in the 56 models, some features are indeed consistently mapped by the models (Figure 5).

Feature 2 seems to be an indicator for unpleasant and eventful soundscapes, otherwise described as chaotic.

High values for features 1, 7 and 10 indicate unpleasantness. Feature 8, 9, 11 and 12 seem to indicate uneventful soundscapes. Features 3, 4 and 5 are not mapped consistently enough to be general indicators for Core Affect evaluation.

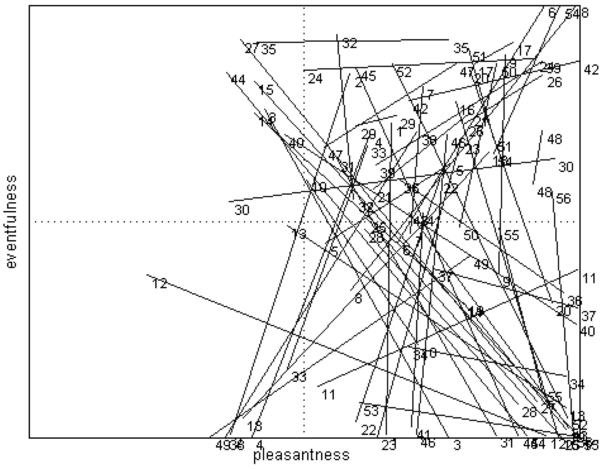

Figure 6: a plot of the Core Affect predictions connected to their true value.

3.3. Performance

Figure 6 shows each prediction of the model connected to the actual Core Affect evaluation. The numbers refer to the fragment let out during training and, therefore the soundscape evaluation that has to be predicted.

Although some predictions are poor, 44% of the prediction lie within *error < 0.5.* And, as Figure 7 shows, the errors are too large in order to be called an predictive model.

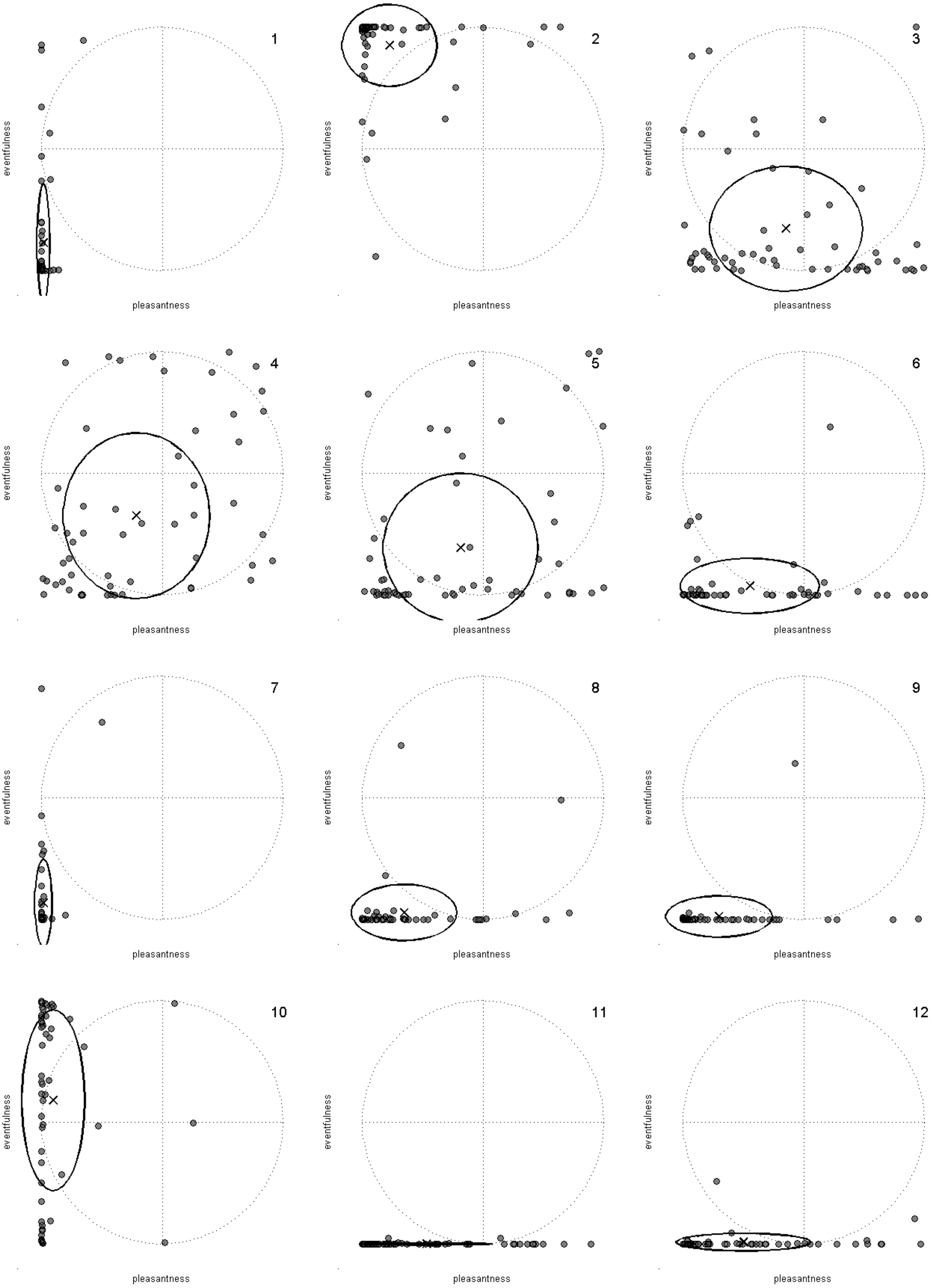

Figure 5: these twelve plots show the mapping of the features for each trained model. For each feature the 56 mappings, the mean Core Affect score (×) and the standard deviation in both pleasantness and eventfulness (the ellipse) are plotted.

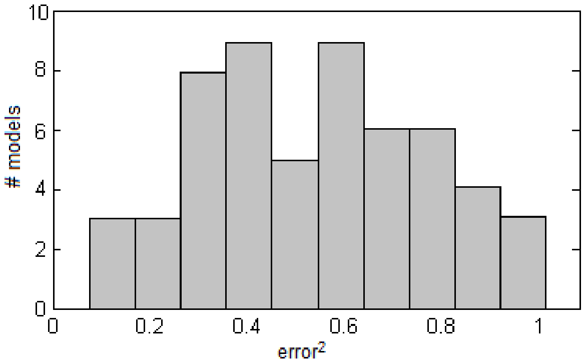

Figure 7: histogram of the mean squared error of the 56 trained models.

4. Discussion

It seems that the current tagged data in combination with a two layer perceptron neural network and features is not able to yield a reliable predictive model for Core Affect scores. The performance of the model is not high enough to be able to predict the Core Affect scores of new sound fragments.

When looking at the different models the neural network has come up with, some features are agreed upon in one dimension and some even in two. These are promising results, for it shows that although the network is not able to predict evaluations above chance level at this time, the features have a predictive power, albeit poor.

Regarding the dataset, most sound fragments are considered both pleasant and eventful, as in real life, but this poses a threat regarding the performance of a neural network; the so called sample bias. To train a neural network optimally, the examples – or input – should be equally distributed along both axes. In this case, however, all but few examples are pleasant. This diminishes the potential performance.

Besides the sample bias a fundamental flaw could arise. As Yik (2011) states, our emotions are explained by our moods and the shift in our moods. Therefore the shifts in Core Affect evaluation during the sound fragment could prove more informative than the current used 'overall evaluation'.

With the help of annotated material with both Core Affect and the shifts over time (as used by Doesburg, 2013), a model capable of dealing with the temporal effects of mood change and therefore Core Affect could yield better results than this static model. The information about the shifts in Core Affect is lost in the current approach.

In conclusion: we were able to determine some physical features of soundscapes that have predictive value with regards to the Core Affect evaluation. Some features proved to be generic features capable of indicating a certain area in the Core Affect circumplex. Such as feature 2, the height of the highest peak. It would seem that a high peak-value indicates a chaotic soundscape. Feature 7 shows that a lack of high pitched pulses, indicative for speech, predicts a boring soundscape; i.e. speech is a candidate feature for an active soundscape.

A model using a static two layer perceptron neural network in combination with these features is not able to correctly predict Core Affect evaluation.

5. References

Broers, E. (2011). Een eerste stap in de automatische evaluatie van soundscapes gebaseerd op binnenopnames. *University of Groningen, The Netherlands*

Doesburg, I. (2013). Using a Joystick to Express Your Opinion About a Sonic Environment. *University of Groningen, The Netherlands*

Halpern, D. L., Blake, R., & Hillenbrand, J. (1986). Psychoacoustics of a chilling sound. *Perception & psychophysics*, *39*(2), 77–80.

Kangur, A. (2011). Het categoriseren van geluidsomgevingen aan de hand van de gemoedstoestanden die worden opgeroepen. *University of Groningen, The Netherlands*

Kuppens, P., Champagne, D., & Tuerlinckx, F. (2012). The Dynamic Interplay between Appraisal and Core Affect in Daily Life. *Frontiers in Psychology*, *3*. doi:10.3389/fpsyg.2012.00380

Lindborg, P. (2012). Correlations Between Acoustic Features, Personality Traits and Perception of Soundscapes.

Russell, J. A., Ward, L. M., & Pratt, G. (1981). Affective Quality Attributed to Environments: A Factor Analytic Study. *Environment and Behavior*, *13*(3), 259–288. doi:10.1177/0013916581133001

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of personality and social psychology*, *76*(5), 805.

Schafer, R. M. (1977). *The Tuning of the World*. Knopf.

Yik, M., Russell, J. A., & Steiger, J. H. (2011). A 12-point circumplex structure of core affect. *Emotion*, *11*(4), 705–731. doi:10.1037/a0023980

Yu, L., & Kang, J. (2009). Modeling subjective evaluation of soundscape quality in urban open spaces: An artificial neural network approach. *The Journal of the Acoustical Society of America*, *126*(3), 1163. doi:10.1121/1.3183377

Appendix A : *description of soundscapes*

#	Description	#	Description
1	Music, some speech and sound of pots	**29**	Birds chirping, loud and positive reaction
2	Loud speech, laughter	**30**	Liquids spilling, machine running
3	Neutral speech, moaning, falling objects	**31**	Motivating speech, moaning
4	Singing, booing, stirring in a cup	**32**	Someone coughing, some pencils falling
5	Children's program (tv), violin piece	**33**	Question, joyful answer
6	Neutral question and explanation	**34**	Clearing table after dinner
7	Faint music, some neutral speech	**35**	Rocking chair, nature sounds
8	Rumbling a box with cutlery	**36**	Neutral explanation of upcoming events
9	Birds chirping,	**37**	Rolling of a coin
10	A gentle stream of water	**38**	Playing shuffleboard
11	Mostly quiet, shutting a door	**39**	Lunchtime, stacking coffee cups
12	Sound of a machine running	**40**	Sound of coffee machine
13	Incomprehensible human sounds	**41**	Forest sounds
14	Humming, neutral speech, verbal abuse	**42**	Beeping of electrical device
15	Shouting, people dishwashing	**43**	Faint moaning, normal conversation
16	Mostly quiet, some neutral speech	**44**	Moaning, shouting
17	Positive speech, laughter	**45**	Someone eating soup
18	Gentle conversation	**46**	Complaints and yawning
19	High heels, closing door	**47**	Faint noise on the background
20	Loud speech, mashing cutlery on plate	**48**	Placing and filling a plastic cup
21	Someone speaking in himself, laughter	**49**	A plastic bag
22	Electrical noise, many loud voices, claps	**50**	Someone running in a hall
23	Children's program (tv)	**51**	Singing and neutral speech
24	Some greetings: "hey, hi" etc.	**52**	Quiet
25	A spitting patient, shouting and music	**53**	Moving a chair, choking laughter
26	Two people playing a board game	**54**	Laughter, falling object, grinning
27	Two patients moaning and ringing bells	**55**	Scratching fabric
28	Opening package, phone notification	**56**	Faint crying