Machine Listening for Park Soundscape Quality Assessment

Michiel Boes, Karlo Filipan, Bert De Coensel, Dick Botteldooren

Department of Information Technology (INTEC)
Ghent University, Sint Pietersnieuwstraat 41, 9000 Gent, Belgium
michiel.boes@intec.ugent.be

Abstract

The increasing importance attributed to soundscape quality in urban design generates a need for a system for automatic quality assessment that could be used for example in monitoring. In this work, the possibility for using machine listening techniques for this purpose is explored. The outlined approach detects the presence of particular sounds in a human-inspired way, and therefore allows to draw conclusions about how soundscapes are perceived. The system proposed in this paper consists of a partly recurrent artificial neural network modified to incorporate human attention mechanisms. The network is trained on sounds recorded in typical urban parks in the city of Antwerp, and thus becomes an auditory object creation and classification system particularly tuned to this context. The system is used to analyze a continuous sound level recording in different parks, resulting in a prediction of sounds that will most likely be noticed by a park visitor. Finally, it is shown that these indicators for noticed sounds allow to construct more powerful models for soundscape quality as reported in a survey with park visitors than indicators that are more regularly used in soundscape research.

1 Introduction

Over the past years, soundscape quality has become an increasingly important factor in urban planning and design, and substantial research efforts have been spent on methods that quantify how people perceive soundscapes [1][2][3][4]. It has been established that, even though relationships can be found between soundscape perception and outdoor energy equivalent sound pressure levels (L_{dn} or L_{den}), these alone are not sufficient to predict outdoor soundscape perception [5][6]. Although unnoticed sounds may influence emotional response to sound, psychological and neurophysiological findings strongly emphasize the significance of selective auditory attention processes in human analysis of acoustic environments [7][8]. Indeed, in order for a sound to contribute to an overall soundscape appraisal it needs to be paid attention to and attributed a meaning to [9][10][11][12]. Human attention processes depend on a range of sound signal properties, not just the level, but also in no negligible degree on temporal and spectral content. Furthermore, these processes are influenced by the state of mind and expectations of the listener [7].

It is clear that the nature of the noticed sounds, their recognition by the listener, and the meaning the listener attributes to them will be of great importance to the influence they will have on the general appraisal of the sound environment. In particular, certain sounds will generally be associated with a positive soundscape quality, while others will be related to a negative quality. More

concretely, from the viewpoint of a park visitor, soundscape quality is found to be negatively related to the presence of mechanical sounds (e.g. road traffic noise) and positively related to the presence of nature sounds, while the relation between sounds from human activity and soundscape quality depends on context, expectations and personal preferences [6].

Questionnaire studies, in which a significant number of people need to be interviewed about their perception of the soundscape, are time consuming and require considerable human resources. Thus, a more automated approach to obtain a measure for the quality of the soundscape would be of significant interest. One possibility here is the use of crowdsourcing, i.e. getting the general public involved in the collection of perceptual soundscape data. Recent advances in mobile computing offer the opportunity to allow many people to participate in such measurement campaigns, and thus make it an appealing approach [13][14]. Another possibility is the use of statistical or computational models to find a relationship between measured acoustical parameters and perceived soundscape quality [15][16]. Taking this approach to the furthest extent would imply to start off from raw sound recordings and use ab initio machine audition, emulating human sound processing, to extract meaning and perception from it. Research efforts in this area have so far mainly been focused on specific sub-problems in controlled environments and the establishment of theoretical frameworks [17][18][19].

This paper presents an ab initio machine learning model to achieve attention-driven humanlike auditory environment perception. We incorporate well-established human attention mechanisms in a 3-layer Artificial Neural Network (ANN), of which the input layer is fed human-inspired sound features, simplified to be extractable from common sound level meters, one timestep at a time. The output layer contains neurons that represent different sounds, and of which the activation strength depends on how clearly the input sound is noticed by the model, based on the implemented attention mechanisms. In this context, a sound is defined as an auditory object, a sequence and combination of acoustic features that can be observed by the human listener and that frequently co-occur or occur in the same sequence in a specific context (in this case parks in Antwerp). Due to the limited number of output neurons, similar sounds will be mapped to the same (set) of neurons.

This model, together with the data collection used to train it, is described in the methodology section. Subsequently, in section 3 it is validated that the model and training procedure result in the identification of auditory objects that are meaningful to a human listener. For this purpose, a human listener identified a few classes of bird sounds in a recording made in parallel to the level recording used as an input to the model. By confirming that each class of birds results systematically in the activation of the same set of output neurons the hypothesis is validated. In section 4 the noticing of mechanical sounds, natural sounds, and human sounds as predicted by the ANN model, are used as indicators in a statistical model for soundscape quality reported in an extended questionnaire survey in 8 urban parks. This section validates that these indicators obtained by human inspired identification of noticed sounds and their classification outperform classical noise level indicators for this purpose. Finally, in section 5, conclusions are formulated.

2 Methodology

2.1 Model

The aim of the proposed model is firstly to combine acoustic features to sounds, i.e. auditory objects, and select those that would most probably be noticed by a human listener due to their saliency within a continuous sound stream. Secondly, these sounds are then grouped into meaningful categories such as mechanical sounds, bird vocalizations, etc. In humans, the formation of meaningful auditory objects is aided by mechanisms such as attention, inhibition of return, adaptation and habituation. These mechanisms have therefore been an important

source of inspiration in constructing the tailored recurrent artificial neural network as explained below.

Initially, the input sound is converted into a series of features, with a time resolution of 0.125s, inspired by human peripheral hearing. Multiple descriptor values are used, calculated in the same way as in Oldoni et. al. [20]: 4 values describing sound intensity at different frequency ranges, 6 describing spectral contrast at different frequency ranges and 6 describing temporal contrast at different time ranges. A spectral resolution of 0.5 Bark (the scale reaching from 0 to 24 Bark) is used, thus resulting in $(4+6+6) \times 24/0.5 = 768$ values per timestep. These features and time resolution are chosen to balance detailed human-mimicking processing on one hand and limited measurement hardware and computational resources on the other, as they can easily be approximated by 1/3 octave bands measured using standard sound level recording equipment. More advanced features typically used in speech recognition or bird song recognition such as MFCC would require dedicated sensor nodes or continuous recording for monitoring and have therefore not been used. The features are then used as excitation values to the 768 artificial neurons in the first layer of the model, after which the 3-layered neural network, the structure of which is shown in figure 1, processes the information.

The neural network builds on previous work by the same authors, and many of its mechanics are the same as described in detail in [21][22], but for clarity the essential elements and differences are described in this paragraph. The network consists of a first layer, called the input layer of 768 neurons as mentioned before. This layer has excitatory connections to a hidden, middle layer, consisting of 1000 neurons, which in turn has excitatory connections to the last, output layer with 400 neurons. This number of neurons is significantly lower than the number of neurons found in modern deep-learning networks and obviously only a fraction of the number of neurons found in biological brains. They were determined by trial and error as a balance between computational cost of training and accuracy. The output layer has excitatory feedback connections to the middle layer, with a time delay of one timestep, making the excitation pattern of the middle layer dependent on both the current input layer activation and the output layer activation on the previous timestep. Excitation of a neuron is calculated as the sum of the exciting inputs weighed by their respective neural connection weights, after which a normalization and saturation procedure is applied, as described in [22]. Final activation of the neuron is then calculated by means of a biologically inspired competitive selection procedure as will be explained in more detail below. Note that the inclusion of a difference of Gaussians filter on the neural activation pattern in the competitive selection procedure implements a form of lateral excitation, as seen in self organizing maps (SOMs) [27].

Learning of the connection weights is done following the Hebb principle: "cells that fire together, wire together". In the current implementation, connection weights are adapted both by learning (strengthening or weakening specific connections in order to create patterns) and by forgetting (random convergence of connection weights towards a set base level), while a dynamic equilibrium between these two effects determines final connection weights. In the untrained network, all connection weights are initialized at random values in a small interval around a base level (in this work the interval $[0.7, 0.9]$ and base level of 0.8 are used). During training, these weights are then modified by both the learning and forgetting mechanisms, while limiting their values to the $[0, 1]$ interval. For detailed analysis and mathematical details of the implementation of these mechanisms, we refer to [22].

In most theories on human attention (visual as well as auditory), the interplay between bottom-up, saliency-based and top-down, voluntary mechanisms, combined with a competitive selection process plays a central role [23][24]. On the one hand, the bottom-up mechanism enhances the response to conspicuous and salient sounds, whereas on the other hand, the top-down mechanism introduces a bias towards sounds that are most relevant for the listener's cur-

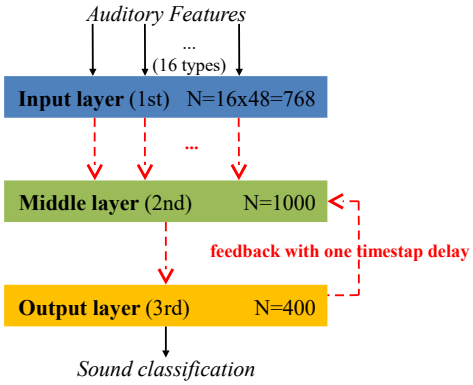

Figure 1: An overview of the structure and connectivity of the neural network model. The dashed red arrows represent excitatory connections between layers.

rent goal-directed behavior. Taking into account the effects of both mechanisms, competitive selection will decide which sounds will finally be consciously noticed by the listener. In addition, often, the concept of inhibition-of-return is introduced, which explains why people do not direct their attention permanently to a single sound [8]. In previous work of the same authors [25], these mechanisms have been implemented explicitly in a functional model of auditory attention. However, in the model proposed here, these mechanisms emerge naturally from the biologically inspired implementation of the 3-layered artificial neural network:

- Bottom-up, saliency-driven attention is implemented by the choice of input features to the model, encoding intensity, spectral and temporal contrast similarly to the features for sound saliency calculation [20][26]. The intra-layer normalization and saturation (implemented exactly as in [22]), combined with the competitive selection procedure (more details provided below) are mathematically very similar to the method used in these references, and thus the activation in the input layer reflects saliency values as calculated there. As connection weights to further layers are initially all around their base levels, activation in these layers will also reflect saliency. As learning proceeds, as described in the third paragraph of this section, however, differentiation between the different neurons grows, and activation will reflect not only saliency but also the degree of pattern recognition that has been learned based on frequent co-occurrence or sequential occurrence of these features. Because connection weights are limited to the $[0, 1]$ interval though, increasing saliency will still result in increasing neural activation. Thus, a more subtle measure for saliency which is not a single number saliency descriptor is achieved.

- Voluntary attention to sounds that are most relevant for the listener's goal directed behavior can obviously not be included in this model. However, top-down mechanisms are also responsible for sustained attention. Once the onset of an auditory object is detected, the probability that this object wins the competitive selection increases. This sustained attention also surpasses short periods of silence such as those present in bird song. The delayed feedback excitation from the third to the second layer of the model assure that this form of top-down, sustained attention emerges.

- Competitive selection is incorporated as an intra-layer excitation-inhibition mechanism making a biologically plausible selection amongst the neurons within each layer. This is implemented by an iterative procedure in which the neural activation pattern of the layer is transformed by self-excitation and inhibition by neighbors (implemented by convoluting the activation pattern with a difference of Gaussians filter) in addition to a default inhibition, similarly to the implementation in [8] and [26]. Formally, this transformation is given by $p \leftarrow \max(0, p + \alpha p * DoG - \beta)$, in which p is the neural layer activation pattern, DoG is the difference of Gaussians filter and α and β describe the relative strengths of each of the contributions. This method results in only the most strongly activated neurons retaining positive values, and thus implements competitive selection between the neurons, in a way which

is in line with the way saliency is calculated in [20][26]. The values of α and β determine the selectivity of the model, and can be adapted to the desired amounts of selected sounds (default values of $\alpha = 1.0$ and $\beta = 0.5$ were used in this work).

- Inhibition-of-return is also included, represented by a neural excitation reducing mechanism as a consequence of continuous stimulation of the neuron, mimicking the gradual depletion of neurotransmitters in real neurons. The concentration of neurotransmitters over time is modelled as $c(t + \Delta t) = \min[1, c(t) + \Delta t\,(\rho - c(t)A(t))]$, in which $c(t)$ is the concentration at time t, Δt is the model's time step, ρ is the concentration regeneration rate and $A(t)$ is the neuron activation strength. In order to calculate the effective activation of the neuron, it is first multiplied by its respective neurotransmitter concentration c. When the neuron is persistently activated, c will decrease over time, and consequently the effective activation of the neuron will decrease, thus effectively implementing inhibition-of-return.

The artificial neural network is trained unsupervisedly: there is no teacher that assigns for example a label to the sounds. This results in the neural connection weights being trained in order to group sounds based on only feature co-occurrence and temporal consistency, or, in other words, feature sequential occurrence. Training on co-occurence resulting in clustering is a direct extension of the self organizing map (SOM) [27] that has been used in our previous work [20]. The temporal consistency is introduced by the feedback loop between the middle and output layer, while grouping based on feature co-occurrence happens mainly between the first two layers. It should be noted that the time constants implemented by this feedback are of the order of 125 msec or longer. Differentiation between sounds based on faster amplitude modulations are captured by the temporal contrast features. The neural activation in the output layer can then be interpreted as a clustering of the input sound, in which each neuron represents a category of sounds.

During the analysis phase, the activation of the output neurons can be interpreted as the degree to which the corresponding sound is likely to be noticed by a park visitor. It does not give any label to this sound and only states that this sound has been observed before and is now present again. In case no neurons in the output layer are activated (which in a typical urban park environment happens most of the time), this signifies that the sound is not being noticed. Note that the categories that are represented by the different output neurons are not predetermined or manually chosen, but determined in an unsupervised way during the learning phase.

2.2 Measurements

In order to train and test the model proposed above, a dataset of sound level recordings and perceptual assesments, obtained in 8 different urban parks in Antwerp, was used. Figure 2 shows the location of the 8 parks: Rivierenhof, Stadspark, Nachtegalenpark, Te Boelaerpark, Bisschoppenhof, Park Sorghvliedt, Park Den Brandt and Domein Hertoghe. The soundscape study was performed during 22 days in August and September 2013. Continuous, mobile sound level recordings were made by three sound level recording devices carried by three different researchers performing random walks through the parks, in order to obtain a sufficient coverage of the soundscape of the entire park. More than 380 hours of sound level recordings were collected, thus about 48 hours per park, divided over the three mobile recording devices. For one of the parks, Rivierenhof, sound recordings conducted simultaneously with the level recordings will be used for recognizing bird songs in section 3.

Concurrently with these sound level recordings, a face-to-face questionnaire study was conducted amongst the park visitors in order to obtain their opinion about the overall park environment and more specifically their assessment of its soundscape. The questionnaire contained 22 questions, including a number of personal background questions (gender, age, roads used to get to the park, reasons to visit the park, etc.) and a number of questions asking for

the visitor's perception of the park on an auditory, visual and general level. As this work focuses on the auditory perception, a selection of relevant questions was made: "How did you experience the sound environment today?", with possible replies on a 9-point bipolar scale between "pleasant" and "unpleasant", "To what degree did you hear these sounds during your current park visit?", with possible replies on a 5-point unipolar scale between "not at all" and "very often" for the sound categories of "human sounds", "natural sounds" and "mechanical sounds". A total of 660 questionnaires was filled in, divided over the 8 parks resulting in approximately 80 questionnaires per park.

To compare the results derived from the sound recordings with those from the questionnaires, the results are grouped per day and per park, as some parks did not have enough visitors and thus not enough filled in questionnaires in order to get meaningful information on intra-day patterns. It was also decided not to follow individual park visitors with the sound measurement equipment, even though this could enable analysis on a visitor by visitor basis, because this approach would likely introduce a bias in the results as the park visitor would be more attentive to sound when being aware of the presence of sound measurement equipment.

Figure 2: A map of the city of Antwerp with the investigated parks.

3 Sound extraction validation

A main hypothesis underlying the approach for soundscape assessment is that the proposed ANN, trained in an unsupervised way, will select and cluster auditory objects in a meaningful way. That is, the sounds as defined by the network on the basis of co-occurrence and sequential occurrence of (salient) features, correspond to 'sounds' in common understanding of people. To validate this hypothesis, the model is applied to a selected period of the sound recordings, and the outcome from its analysis is compared to the labeling of bird sounds by an attentive listener in the same data. As the presence of bird sounds is generally seen as a strongly positive element in a park soundscape [6], it is an interesting and valuable benchmark. In order to achieve this, one attentive listener listened to two full days of recordings (twice 8 hours on 3 microphones, so a total of 48 hours) in Rivierenhof park. A user interface was created in which the listener could press one key at the start of a bird sound, and another one at the end, with the additional possibility to relisten and correct if necessary. Afterwards, the same listener went through all selected bird sound recordings and labeled them according to bird family (geese, pigeons, gulls, jackdaws, ducks, crows and songbirds). This way, 2129 bird sounds were selected and labeled, the duration of which ranged from around half a second (short shouts) to as long as five seconds (full songs).

The ANN on the other hand was trained in an unsupervised way on the full measurement dataset of 380 hours as described previously. The input sound is fed into the network consecutively, in the same way as a human listener would listen to the recordings. The implemented attention mechanisms and SOM-like lateral excitation result in attention-fuelled competitive learning, in which a certain degree of plasticity remains, thanks to the inclusion of the "forgetting" mechanism as mentioned before.

As the input sounds in the context of this work are all of a similar nature (park sounds), an equilibrium in the connection weights is reached eventually when no completely new sounds are presented to the ANN anymore. In the ANN used in this work, 95% of the connection weight change compared to their initial values happened within the first 100 hours of training. After training, the model was run on the level recordings of the two days in Rivierenhof park for which the synchronised sound recordings had been analyzed by the human listener. The attentive reader will notice that this validation is done on a subset of the training data. In a classical machine learning context, validation checks whether a complex model is applicable in different contexts or it over-fits the data it is presented with, and thus, validation should be done on a set independent of the training set in place and in time. Yet, the goal here is to validate that the automatic construction of auditory objects matches the sounds that a human listener would identify. The model will specialize on park sounds in a specific region where the training set is collected obviously, but so would a human listener living in one particular continent or area with a certain degree of organization. By means of the attention and gating mechanisms implemented in the model described above, the artificial neurons in the output layer are not activated continuously, but rather in well delimited short timeframes, thus selecting noticed sound events that are likely to be noticed, and at the same time classifying them, depending on which neuron in the output layer is activated. Note that this is achieved without any supervision or any interaction with the model, and not only bird sounds are selected, but a whole range of sounds.

In order to quantify the perfomance of the model, these two, completely independent selections and categorizations of sounds from the same pool need to be compared. Two important factors were evaluated, the first being the attentiveness, i.e. the amount of bird sounds the model actually selects. The second factor is the correctness, or the accuracy of the model's ability to categorize all the sound events it detects. Since the first factor is not a property that can simply be described in terms of 'correct' or 'incorrect', as attentiveness varies between different listeners and their mood and activity at the time of listening, it is represented by the percentage of bird sounds that are paid attention to. In order to obtain this percentage, a bird sound is considered selected by the model if the overlap between the bird sound time interval as determined by the human listener and a neural network selected sound interval is sufficiently high (in this case an overlap of 50% was used). The second factor is an exact property that can be quantified by its false/true positives/negatives, and in this work it is represented by a Receiver Operating Characteristic (ROC) or ROC curve. This curve shows the True Positive Rate (TPR), the number of true positives divided by the total number of positives, of a binary classifier as a function of the False Positive Rate (FPR), the number of false positives divided by the total number of negatives, for a range of threshold values θ. In order to determine which neurons of the network represent positives, i.e. "birds", the fraction of selected sounds for each of the neurons that correspond to a human selected bird sound (correspondence is defined as above with a minimum time interval overlap of 50%) is compared to the threshold value θ. In case it exceeds the threshold, this neuron is considered to represent bird sounds, and thus a positive, and vice versa. This selection of bird sound neurons is done with the use of the data of the first measurement day. Next, the TPR and FPR are calculated on the data of the second measurement day, calculating the TPR as the total number of selected sounds that correspond to human selected bird sounds that are categorized in bird sound neurons, divided by the total number of selected sounds attributed to these neurons, and calculating the FPR as the total number of selected sounds that correspond to human selected bird sounds that are categorized in non-bird sounds neurons, divided by the total number of selected sounds attributed to these neurons. Thus, the ideal point on the ROC curve is clearly at a TPR of 1 and a FPR of 0, while a random classifier would result in points on the diagonal where TPR=FPR.

First, the model is evaluated with its default parameters, resulting in 21.5 percent of the labeled bird

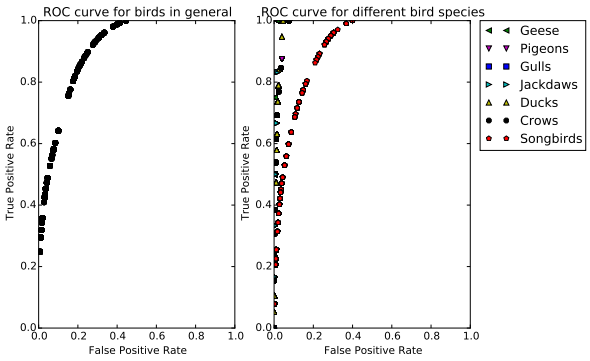

Figure 3: ROC curves for the bird sounds selected by the ANN.

sounds being selected by the model. The resulting ROC curve can be seen in figure 3, calculated for bird sounds in general (left) as well as for each of the different bird families separately (right). Because the majority of birds in most of the parks are songbirds, the ROC curve for bird sounds in general is very similar to the one for just songbirds, as can be seen in figure 3. It can also be seen that songbird recognition performs worse than the other kinds of birds. Closer inspection and listening to false positives reveals that this is mainly due to a certain degree of confusion by the model between songbirds and playing children. The other bird families have more distinct sounds and are not as easily confused with other sound sources present in the park, which is reflected in their ROC curves.

By adapting the value of β in the competitive selection as described in 2, the model can be set to be more or less selective to sound input, just like the attention human listeners attribute to sound can change depending on the environment and the current occupation of the listener. Changing β to 0.1 compared to the default value of 0.5 and thus making the model more attentive to sounds in general, a percentage of 47.9 of noticed bird sounds is reached. The ROC curve in this case moves to the situation as shown in figure 4. It can be seen that the categorization quality of the model is slightly reduced in this case, as also less salient and thus more difficult to categorize sounds are selected by the model, which results in more mistakes in the categorization.

Literature values for bird sound detection rates in background vary widely, depending on the method used to quantify the quality of the detection, the experimental setup, the relative strength and type of background sound, the species of the birds, etcetera, thus making a comparison very difficult. To give an idea, Papadopoulos et. al. [28] report AUCs (Area Under Curve, the total area under the ROC curve) of over 0.9 for 10 out of 15 species, but as low as 0.56 for some. Potamitis et. al. [29] on the other hand focus on just two species of birds, and evaluates by means of a precision and a recall percentage, instead of a ROC curve. They report precision values between 71% and 88% and recall values between 77% and 92%. Even though these values are not directly comparable to the values obtained in this work, because of the aforementioned reasons, it can be stated that the quality of the current model is roughly comparable, even though it is not explicitly designed for the purpose of detecting and classifying bird sounds only, unlike the other techniques.

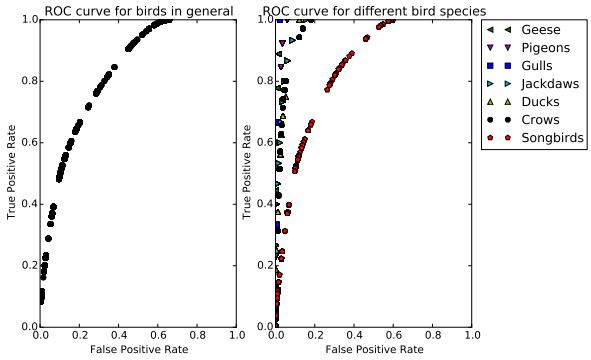

Figure 4: ROC curves for the bird sounds selected by the ANN, with $\beta = 0.1$ in the ANN competitive selection.

4 Application in soundscape appraisal

Several studies have shown that the frequency of hearing mechanical, natural, and human sounds is a strong predictor of soundscape quality [5][6]. The proposed ANN allows to identify the sounds that a park visitor would most likely notice. Hence, in this section it is investigated whether the calculated percentage of the time that these sounds are noticed are good indicators for soundscape quality. For this, the sound events selected and categorized by the ANN model into the different output neurons need to be labeled. From the complete 380 hours of recordings, 7292 sounds were automatically selected, with an average duration of 0.86s (thus amounting to less than 0.5% of the time), divided over the 400 output neurons of the ANN. For each neuron, a small random sample was taken from the sounds selected by the ANN, and based on these sounds a human listener could assign to each of these neurons one of the three classes as used by Nilsson et. al. [6]: natural sounds (mainly birds, but also the flow of water and wind), mechanical sounds (mainly traffic around the park, but also some construction sounds) and human sounds (people talking, restaurant sounds). In case of doubt (sounds belonging to different classes in one neuron), no category was assigned, but this was only the case for a small minority of neurons ($< 5\%$). In this work, a fourth class was added, containing all the sounds related to the execution of the measurements, such as sounds caused by the movement of the backpacks containing the mobile measurement devices or occasional voices of the researchers executing the measurements. Although all reasonable efforts were made to stay quiet during the measurements, the proximity of these sound sources to the microphones caused these sounds to be relatively salient, and thus causes the model to detect them fairly easily. As these sounds are also very distinct and different from most other sounds heard in a city park, the model is well tuned to these sounds because it learned them very well in the training phase by virtue of their saliency, they are effectively categorized apart from other sound sources. This allows to easily eliminate these non-relevant and contaminating sounds from the measurements.

The applied method allows for the calculation of the number of sound events per hour noticed by the model in each of these classes per park and measurement day, denoted by H_{ANN}, N_{ANN} and M_{ANN} for human, natural and mechanical sounds respectively. These can then be related to the responses given by the park visitors in the questionnaire to the question "To what degree did you hear these sounds during your current park visit?" and "How did you experience the sound environment today?". The mean of the responses of the park visitors is calculated per park and measurement day, denoted by H_Q, N_Q, M_Q and Q_Q for human sounds, natural sounds, mechanical sounds and soundscape quality respectively. Regressions were created for each of the four questionnaire results using an ordinary least-square method [30], in which all three ANN results were included in forward selection, with the model selecting by the highest F-value.

The regression coefficients and adjusted R^2 and F-statistics are given in table 1 (p-values for all independent variables are < 0.05 and the F-values are well above their critical value for 5% significance, as this is 3.522 in the case of 2 used variables, and 4.351 in the case of 3 used variables), while the regression is visualized by plotting the actual questionnaire values as a function of the predicted values by the model for the different parameters (human sounds, natural sounds, mechanical sounds and soundscape quality) in figure 5. For the prevalence of human sounds as reported by the questionnaire respondents, the only significant predictor was found to be the prevalence of human sounds resulting from the ANN, with a positive regression coefficient, as expected. For both the reported natural and mechanical sound prevalence, the ANN predicted natural and mechanical sound prevalences were both found to be relevant predictors, with calculated mechanical sound prevalence having a negative regression coefficient for percieved natural sound prevalence and vice versa. Lastly, also for the reported soundscape quality, only

the ANN predicted natural and mechanical sound prevalences were found to be significant predictors, with a positive regression coefficient for the natural sound prevalence and a negative one for mechanical sound prevalence, which is in line with the results found by Nilsson et. al. [6].

For comparison, the same method was applied to correlate the questionnaire answers per park and measurement day to classic acoustic indicators per park and measurement day. The indicators that were used are A-weighted percentile levels ($L_{A10}, L_{A50}, L_{A90}$), the A-weighted equivalent level (L_{Aeq}), the difference between A-weighted and C-weighted equivalent levels ($L_{Ceq} - L_{Aeq}$), the 50-percentile Zwicker's loudness (N_{50}) [31], the 50-percentile Von Bismarck's sharpness (S_{50}) [32], the spectral center of gravity (COG), the music-likeness (ML) [33] and the number of sound events (NCN) [33]. The resulting regression coefficents and adjusted R^2 and F-statistics are given in table 2 (p-values for all independent variables are < 0.05), while the visualization of the regression is given in figure 6. For mechanical sounds the only significant predictor was found to be sharpness. Furthermore, center of gravity was found to be representative of natural sounds perception demonstrating that the spectral information was a relevant predictor for these types of sounds as well. The extracted model for human sounds, on the other hand, includes multiple indicators showing that the perception of these sounds was difficult to characterize with a single indicator. Finally the only significant predictor for soundscape quality is found to be sharpness, with a regression coefficient which has an opposite sign than the one for mechanical sounds, implying that less mechanical sounds result in better soundscape quality.

When comparing these regression models, based on classic acoustic indicators, to the ones based on the ANN output, it is clear that the adjusted R^2 is higher for the ANN based models, thus indicating that a larger proportion of the variance in the questionnaire responses is predicted by the ANN based models than by the classic indicators based models,

	H_Q	N_Q	M_Q	Q_Q
C	1.0976	3.0154	2.5396	6.6500
H_{ANN}	0.1761			
N_{ANN}		0.1786	-0.1685	0.2739
M_{ANN}		-0.1693	0.1913	-0.1726
Adj. R^2	0.662	0.621	0.774	0.598
F-value	42.18	18.19	36.88	16.63

Table 1: Linear regression models for human, natural and mechanical sounds and soundscape quality as reported by park visitors in the questionnaire as a function of human, natural and mechanical sounds as detected by the ANN (C denotes the intercept). Only regression coefficients for relevant contributors, as determined by forward selection based on F-value, are shown. In addition, adjusted R^2 and F-statistics are given for each model.

	H_Q	N_Q	M_Q	Q_Q
C	-3.7768	1.3182	9.8262	-0.8368
L_{A90}	0.1000			
ML	2.6954			
COG		0.0039		
S_{50}			-5.3652	6.1567
Adj. R^2	0.486	0.341	0.645	0.493
F-value	10.91	11.88	39.22	21.38

Table 2: Linear regression models for human, natural and mechanical sounds and soundscape quality as reported by park visitors in the questionnaire as a function of classic acoustic indicators (C denotes the intercept). Only regression coefficients for relevant contributors, as determined by forward selection based on F-value, are shown. In addition, adjusted R^2 and F-statistics are given for each model.

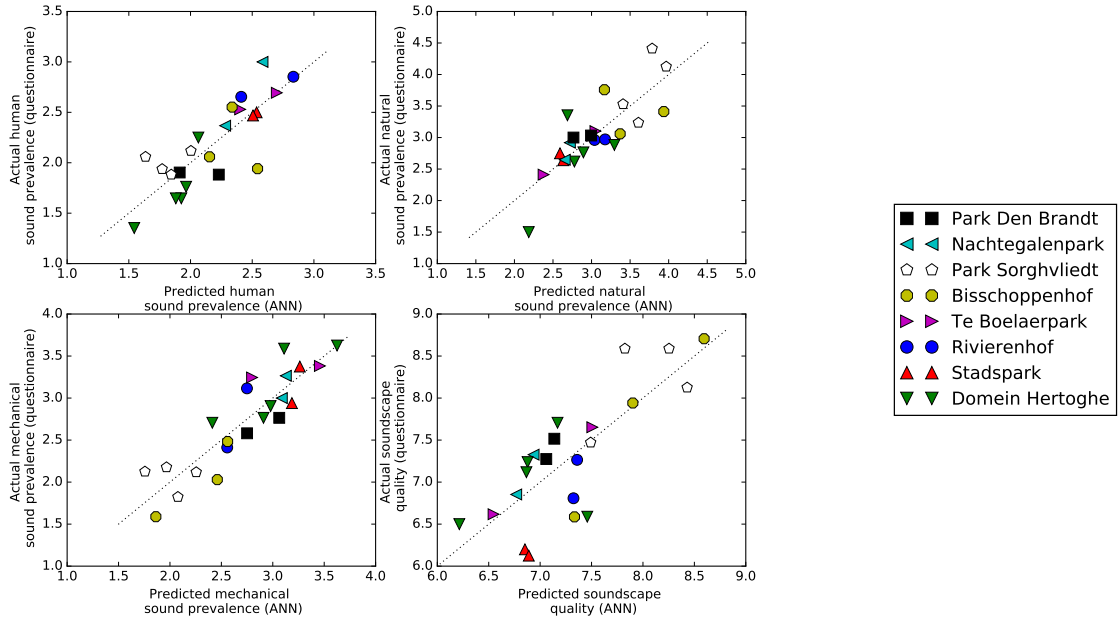

Figure 5: Actual values as a function of predicted values by the regression models based on ANN results given in table 1

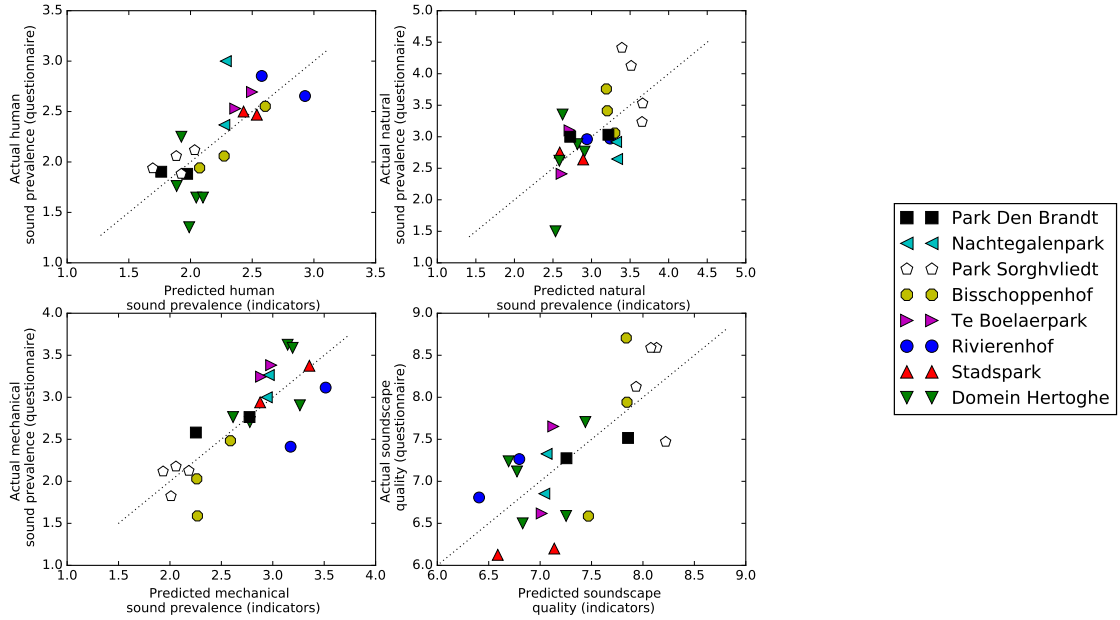

Figure 6: Actual values as a function of predicted values by the regression models based on acoustic indicators given in table 2

while F-values are comparable in magnitude (higher for the ANN based models for human and natural sounds, the other way around for mechanical sounds and soundscape quality). Thus, the ANN provides an improvement on classical acoustic methods.

Note that the broader applicability of the derived models is limited by two factors. Firstly, the average expectation pattern of the listeners determines which sounds will be more often noticed, and whether they will be percieved more positively or negatively [34]. Thus adaptation of the regressions will be required in case this is different, as the questionnaire results on which these are based will not be generally valid any more. In case of urban parks in the same city, it can be assumed that the expectation patterns are similar, but in order to assess urban soundscapes outside of parks, for instance, the average expectations will clearly be different. Secondly, the corpus of sounds that are present in the environment also needs to be similar to the corpus of sounds in the training set in order for the ANN output to be reliable. Again, in the case of urban parks with similar sound sources, the corpus of occurring sounds will be similar, but in order to assess office soundscapes, or park soundscapes with completely different fauna for instance, the ANN would need to be retrained on an appropriate training set of sounds, in order to be able to select and classify these. Thus, even though these two factors limit the general applicability of the derived models, the method to obtain them should remain valid. Therefore, benchmarking its accuracy in completely different environments forms an interesting topic of further research.

5 Conclusions

It was illustrated that machine listening techniques could be used to predict the categories of sounds that park visitors are likely to notice and that these indicators could be used to construct a model for soundscape quality. It was shown that the prediction of noticebility of different classes of sounds and soundscape quality appraisal by users of 8 parks in Antwerp was better or at least as good as a prediction based on classical sound level indicators. Yet the model has the advantage of explicitly including the mechanisms underlying perception of the sound environment. The machine listening system proposed in this work to achieve these results is a 3-layered artificial neural network adapted to take human attention mechanisms and inhibition-of-return into account, thus enabling the network to only process the information that receives attention. In addition to the comparison to a soundscape questionnaire filled in by park visitors, the ability of this model to select and classify auditory objects is validated by a comparison to an attentive human listener's labeling of different bird species' sounds in continuous park sound recordings. The machine listening model used in this work uses 1/2 Bark or 1/3 octave band average levels sampled at a 125 msec interval as raw input. Although this allows to use standard sound level meters to collect data, models relying on more detailed features extracted from continuous sound streaming or dedicated sensor nodes, will most likely outperform the model presented here. Likewise combining the innovations presented in this work with new ANN architectures and extreme learning such as deep networks could advance the application of machine listening in soundscape research.

Acknowledgement

Michiel Boes is a doctoral fellow of the Research Foundation-Flanders (FWO-Vlaanderen); the support of this organization is gratefully acknowledged. The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme *FP7/2007-2013/* under REA grant agreement n°290110, SONORUS "Urban Sound Planner" and from FWO-Vlaanderen under grant no. G0D5215N.

References

[1] J. Kang, "A systematic approach towards intentionally planning and designing soundscape in

urban open public spaces," in *Proceedings of Internoise 2007*, (Istanbul, Turkey), 2007.

[2] B. De Coensel, A. Bockstael, L. Dekoninck, D. Botteldooren, B. Schulte-Fortkamp, J. Kang, and M. E. Nilsson, "The soundscape approach for early stage urban planning: a case study," in *Proceedings of Internoise 2010*, (Lisbon, Portugal), 2010.

[3] W. J. Davies, M. D. Adams, N. S. Bruce, R. Cain, A. Carlyle, P. Cusack, D. A. Hall, K. I. Hume, A. Irwin, P. Jennings, M. Marselle, C. J. Plack, and J. Poxon, "Perception of soundscapes: An interdisciplinary approach," *Appl. Acoust.*, vol. 74, pp. 224–231, 2013.

[4] B. Shulte-Fortkamp and A. Fiebig, "Soundscape analysis in a residential area: An evaluation of noise and people's mind," *Acta Acustica united with Acustica*, vol. 92, pp. 875–880, 2006.

[5] H. M. E. Miedema and H. Vos, "Noise sensitivity and reactions to noise and other environmental conditions," *J. Acoust. Soc. Am.*, vol. 113, pp. 1492–1504, 2003.

[6] M. E. Nilsson and B. Berglund, "Soundscape quality in suburban green areas and city parks," *Acta Acustica united with Acustica*, vol. 92, pp. 903–911, 2006.

[7] M. Elhilali, J. Xiang, S. A. Shamma, and J. Z. Simon, "Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene," *PLoS Biol.*, vol. 7, p. e1000129, 2009.

[8] B. De Coensel and D. Botteldooren, "A model of saliency-based auditory attention to environmental sound," in *Proceedings of 20th International Congress on Acoustics, ICA 2010*, (Sydney, Australia), 2010.

[9] L. R. Schomer, P. D. an Wagner, "On the contribution of noticeability of environmental sounds to noise annoyance," *Noise Control Eng. J.*, vol. 44, pp. 294–305, 1996.

[10] M. Sneddon, K. Pearsons, and S. Fidell, "Laboratory study of the noticeability and annoyance of low signal-to-noise ratio sounds," *Noise Control Eng. J.*, vol. 51, pp. 300–305, 2003.

[11] B. De Coensel, D. Botteldooren, T. De Muer, B. Berglund, M. E. Nilsson, and P. Lercher, "A model for the perception of environmental sound based on notice-events," *J. Acoust. Soc. Am.*, vol. 126, pp. 656–665, 2009.

[12] D. Dubois, C. Guastavino, and M. Raimbault, "A cognitive approach to urban soundscapes: Using verbal data to access everyday life auditory categories," *Acta Acustica united with Acustica*, vol. 92, pp. 865–874, 2006.

[13] N. Maisonneuve, M. Stevens, M. Niessen, and L. Steels, "Noisetube: Measuring and mapping noise pollution with mobile phones," in *Information Technologies in Environmental Engineering* (I. Athanasiadis, A. Rizzoli, P. Mitkas, and J. M. Gomez, eds.), pp. 215–228, Springer-Verlag, 2009.

[14] C. Mydlarz, I. Drumm, and T. Cox, "Application of novel techniques for the investigation of human relationships with soundscapes," in *Internoise*, (Osaka, Japan), 2011.

[15] L. Yu and J. Kang, "Modeling subjective evaluation of soundscape quality in urban open spaces: An artificial neural network approach," *J. Acoust. Soc. Am.*, vol. 126, pp. 1163–1174, 2009.

[16] D. Botteldooren and B. De Coensel, "Quality labels for the quiet rural soundscape," in *Internoise*, (Honolulu, Hawaii, USA), 2006.

[17] T. Adringa, "Audition: From sound to sounds," in *Machine Audition: Principles, Algorithms and Systems* (W. Wang, ed.), ch. 4, pp. 80–105, Information Science Reference, 2011.

[18] T. Adringa, J. Jolie, and L. Lanser, "How pleasant sounds promote and annoying sounds impede health: A cognitive approach," *Int. J. Environ. Res. Public Health*, vol. 10, pp. 1439–1461, 2013.

[19] J. Salamon, and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24(3), pp. 279–283, 2017.

[20] D. Oldoni, B. De Coensel, M. Boes, M. Rademaker, B. De Baets, T. Van Renterghem, and D. Botteldooren, "A computational model of auditory attention for use in soundscape research," *J. Acoust. Soc. Am.*, vol. 134, pp. 852–861, 2013.

[21] M. Boes, D. Oldoni, B. De Coensel, and D. Botteldooren, "A biologically inspired recurrent neural network for sound source recognition incorporating auditory attention," in *IEEE International Joint Conference on Neural Networks (IJCNN)*, (Dallas, TX, USA), 2013.

[22] M. Boes, D. Oldoni, B. De Coensel, and D. Botteldooren, "Long-term learning behavior in a recurrent neural network for sound recognition," in *IEEE International Joint Conference on Neural Networks (IJCNN)*, (Beijing, China), 2014.

[23] J. B. Fritz, M. Elhilali, S. V. David, and S. A. Shamma, "Auditory attention – focusing the searchlight on sound," *Curr. Opin. Neurobiol.*, vol. 17, pp. 437–455, 2007.

[24] E. I. Knudsen, "Fundamental components of attention," *Annu. Rev. Neurosci.*, vol. 30, pp. 57–78, 2007.

[25] M. Boes, D. Oldoni, B. De Coensel, and D. Botteldooren, "Attention-driven auditory stream segregation using a som coupled with an excitatory-inhibitory ann," in *IEEE International Joint Conference on Neural Networks (IJCNN)*, (Brisbane, Australia), 2012.

[26] O. Kalinli and S. Narayanan, "A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech," in *Proc. Interspeech 2007*, (Antwerp, Belgium), 2007.

[27] T. Kohonen, "The self-organizing map," *Neurocomputing*, vol. 21(1), pp. 1–6, 1998.

[28] T. Papadopoulos, S. Roberts, and K. Willis, "Detecting bird sound in unknown acoustic background using crowdsourced training data," *Big Data Sciences for Bioacoustic Environmental Survey*, vol. arXiv:1505.06443, 2015.

[29] I. Potamitis, S. Ntalampiras, O. Jahn, and K. Riede, "Automatic bird sound detection in long real-field recordings: Applications and tools," *Appl. Acoust.*, vol. 80, pp. 1–9, 2014.

[30] S. Seabold and J. Perktold, "Statsmodels: Econometric and statistical modeling with python," in *9th Python in Science Conference*, pp. 57–61, 2010.

[31] ISO, "Acoustics – method for calculating loudness level," ISO 532:1975, International Organization for Standardization, Geneva, Switzerland, 1975.

[32] G. von Bismarck, "Sharpness as an attribute of the timbre of steady state sounds," *Acta Acustica united with Acustica*, vol. 30(3), pp. 159–172, 1974.

[33] D. Botteldooren, B. De Coensel, and T. De Muer, "The temporal structure of urban soundscapes," *Journal of sound and vibration*, vol. 292(1), pp. 105–123, 2006.

[34] K. Filipan, M. Boes, B. De Coensel, C. Lavandier, P. Delaitre, H. Domitrović, and D. Botteldooren, "The personal viewpoint on the meaning of tranquility affects the appraisal of the urban park soundscape," In prep.