Edinburgh, Scotland
EURONOISE 2009
October 26-28

Understanding a soundscape through its components

Maria E. Niessen[a]
Dirkjan Krijnders[b]
Tjeerd C. Andringa[c]
Department of Artificial Intelligence, University of Groningen
P.O. Box 407, 9700 AK Groningen, The Netherlands

ABSTRACT

Human perception of sound in real environments is complex fusion of many factors, which are investigated by divers research fields. Most approaches to assess and improve sonic environments (soundscapes) use a holistic approach. For example, in experimental psychology, subjective measurements usually involve the evaluation of a complete soundscape by human listeners, mostly through questionnaires. In contrast, psychoacoustic measurements try to capture low-level perceptual attributes in quantities, such as loudness. However, these two types of soundscape measurements are difficult to link other than with correlational measures. We propose a method inspired by cognitive research to improve our understanding of the link between acoustic events and human soundscape perception. Human listeners process sound as meaningful events. Therefore, we developed a model to identify components in a soundscape that are the basis of these meaningful events. First, we select structures from the sound signal that are likely to stem from a single source. Subsequently, we use a model of human memory to predict the location at which a sound is recorded, and to identify the most likely events that constitute the components in the sound given the location.

1. INTRODUCTION

Human perception of sound in real environments is a complex fusion of many factors. Therefore, many fields of research are involved in trying to understand these factors, ranging from psychoacoustics[1] to cognitive psychology[2] and sociology[3]. Most psychoacoustic studies on sound quality evaluation focus on measuring one-dimensional attributes of isolated sounds[4,1]. However, several studies have shown that perceptual attributes can only explain part of the full perception of soundscapes[5,6]. The judgment of a soundscape is largely dependent on the meaning that a listener gives to the sound[7]. For example, whether a listener enjoys music depends on his or her choice to hear it. At a concert, music will be appreciated even (or especially) at a high loudness level, while the tolerance for the music that a neighbor is playing at night will be much lower.

Zhang and Kang[8] distinguish four categories in which the different factors that influence human soundscape perception can be organized, namely sound, space, people, and environment. The categories of sound and space comprise the acoustical factors of the source, altered by transmission effects, such as background sounds and reverberation. These acoustical factors in soundscape perception have been tested in psychoacoustic studies[9] and in psychological studies. However, acoustical factors cannot be studied in isolation, because the judgment of a listener is not based solely on the properties of the sound, but is also affected by factors like the

[a] m.niessen@ai.rug.nl
[b] j.d.krijnders@ai.rug.nl
[c] t.andringa@ai.rug.nl

listener's memory and his cultural background. This interplay can be accounted for either by controlling the sounds and varying the condition, such as the cultural background[10], or by correlating psychoacoustic measures to listeners' judgments[11]. Other non-acoustical factors that affect the listener's judgment can include social, demographical, and behavioral factors[12], and environmental factors, such as temperature, wind and sunshine[8].

Many of these recent studies on the judgment of soundscapes use holistic measurements, because soundscape evaluation is a holistic perception, not a sum of the evaluation of the acoustic properties of sound sources. However, the soundscape that listeners evaluate is composed of different acoustic events. Depending on the complexity of the acoustic surrounding and the state of the listener, some or all of these events will be identified and processed as meaningful events[2]. Consequently, they will affect the complete soundscape evaluation. These meaningful events can be seen as a link between the holistic judgment of a soundscape and the acoustic events.

To get insight in soundscape perception, we will focus on this link between a soundscape and acoustic events. More specifically, we propose a method to automatically identify acoustic events based on signal-driven hypotheses, which are guided by knowledge of the environment. The short-term history of an acoustic event is used to predict the location where the event is recorded. Vice versa, the predicted location is used to form expectancies of events that follow. The hypotheses about events approach meaningful events, because they are learned from human annotations. Furthermore, we use a model of human memory to manage the hypotheses. By using an approximation of short-term memory, we include an important cognitive factor in the analysis of a soundscape. Although these event hypotheses are not similar, or even close to cognitive representations, they can be used to automatically analyze a soundscape in a more meaningful way than through acoustic properties alone. Therefore, this method provides a basis for modeling the factors that are important in soundscape perception.

In the following section we will describe the methods that we developed to segment and label components in a soundscape. In the third section we present a data set, which is used in an experiment to test the combined methods. Finally we will discuss the results of the experiment, and give an outlook on future work.

2. METHODS

To identify acoustic events in a continuous sound signal, we first select components from the sound signal that are likely to stem from a single source (section A). Subsequently, we use a model of human memory to select the most likely label for the events that constitute these components, based on a prediction of the location (section B). The methods are only briefly described here. For a detailed description we refer to Krijnders *et al.*[13] and Niessen *et al.*[14].

A. Sound Processing

The spectrogram of the sound signal is segmented on the basis of the local spectro-temporal properties. Segments are likely to stem from a single source when they are based on local properties. For example, local energy maxima that resemble tones and are developing smoothly in time are likely to stem from the same source. The robustness and reliability of these segments, called signal components, are improved with grouping principles from auditory scene analysis, such as common onset, common offset and common frequency development[15,16]. The strategy to combine local signal properties and grouping principles allows to select qualitatively different types of groups, namely tones and harmonic complexes, pulses, and broadband events. A physical description of these groups is used to classify and label them as sound events with a *k*-nearest neighbor (*k*-NN) classifier.

B. Dynamic Network Model

The segmented groups, described in the previous section, are labeled according to the information in the sound signal. However, the sound signal can be distorted or masked by transmission effects such as background sounds and reverberation, resulting in a low confidence of certain group labels. Furthermore, some acoustic events have a distinct meaning, but similar sound structures, such as screaming and laughing. To resolve these confusions, we propose a method that incorporates knowledge of (part of) the context, based on the short-term history of the events.

This method, which is inspired by cognitive research[17,18], constructs a dynamic network that keeps track of both signal-driven groups and knowledge of the context. The nodes of the dynamic network represent information about sound events at different levels of complexity, and the connections between them represent the probability that these pieces of information belong together. Each node holds an activation value. A hypothesis (node) is more likely to be correct when its activation value is higher than its competitors. Whenever new signal-driven information becomes available, the network is updated by adding nodes, which represent new pieces of information, and removing nodes whose activation is below a threshold. Subsequently, the activation of the new nodes spreads through the network. Furthermore, new nodes are used to form expectancies of future sound events. If a signal-driven group matches an expected event, it is more likely to be correct.

An example of a network configuration is depicted in figure 1. The nodes at the lowest level represent segmented groups, at the middle level they correspond to event hypotheses, and at the highest level to hypotheses about locations. Nodes at the different levels are connected with some strength, denoted by weight *w*. These weights can be used to infer probable locations of sound events. For example, birds are heard more often in a park than at a road intersection. The strength between the node that represents a location and the nodes that represent the individual sound events is calculated according to a term-weighting approach used in automatic document retrieval[19]. In this method the importance of a term (word or phrase) in a document is determined by multiplying its frequency in the document with the inverse frequency it occurs in other documents. Hence, the term is important for a document if it occurs often in that document and infrequently in other documents. Analogously, if a sound event *E* is encountered often at location *L*, and little at other locations, it is an important indicator for location *L*. Accordingly, the strength between the sound event *E* and the location *L* is (scaled between 0 and 1):

$$w_{E,L} = tf \cdot \frac{10\log(N) - 10\log(n)}{10\log(N)} \text{ with } tf = \frac{T_{E,L}}{T_E}, \quad (1)$$

where *N* is the total number of locations, *n* is the number of locations at which *E* occurs, and *tf* is the term frequency, calculated as the fraction of $T_{E,L}$, the total duration of occurrences of *E* at *L*, and T_E, the total duration of occurrences of *E* in a training set.

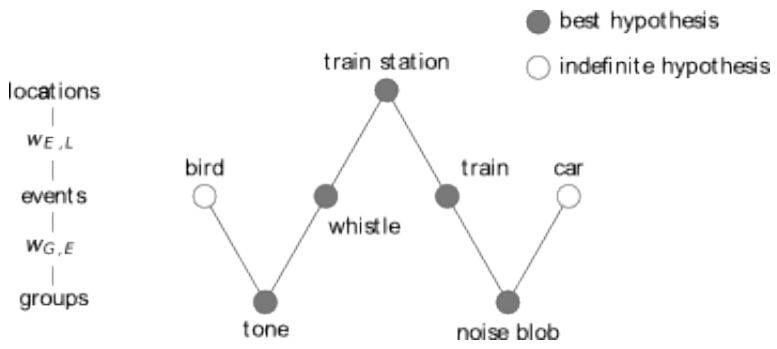

Figure 1: Example of the dynamic network model.

3. EXPERIMENT

We present an experiment to demonstrate that the proposed methods can be used to identify events in a soundscape given a predicted location. First, we describe the data set that is used in the experiment. Next, the setup of the experiment is explained, and in the last part we present the results of the experiment.

A. Data

The data set was collected under different weather conditions on a number of days in March 2009 in the town of Assen (65,000 inhabitants, in the north of the Netherlands). The recordings were made by six groups of three students as part of a master course on sound recognition. Each group made recordings of three minutes at six different locations: a railway station platform, a pedestrian crossing with traffic lights, a small park-like square, a pedestrian shopping area, the edge of a forest near a cemetery, and a walk between two of the positions. Recordings were made using M-Audio Microtrack-II recorders with the supplied stereo microphone at 48 kHz and 24 bits stereo. This data, with annotations, will be made available on http://daresounds.org.

All the recordings were annotated by two students separately. These two annotations were merged, such that equal labels did not overlap, but became one instance. We examined the resulting merged annotations, and adjusted them when necessary. However, we did not introduce new annotations. (An exception was made for the annotations of one group, which we had to complete because they were too meager.) We ensured that the names of events were uniform across all the files to prevent the dynamic network model from learning annotators rather than locations. The total of 44 audio files, with an average duration of 3,5 minutes, were annotated for 54 different classes. However, half of these classes were annotated less than 5 times, while just a few classes comprised most of the annotations. In table 1 a few examples of annotated classes are given, ranked according to their frequency in the complete data set.

Table 1: Examples of annotated classes and their occurrences.

Class	Total number of occurrences	Sum of duration of occurrences
Bird	238	17 min
Bike	30	2 min 20 sec
Rooster	16	43 sec
Horn	8	11 sec
Shopping bag	1	7 sec

B. Setup

The annotations of sound recordings were used to train both the *k*-NN classifier, for the labeling of the signal-driven groups, and the knowledge of the dynamic network. For the *k*-NN classifier, all 44 audio files were processed with the signal-driven method described in section 2A. The segmented groups with the highest score that overlapped with an annotation were given that annotation as a label. Groups that did not overlap in time with an annotation were labeled as noise. All other groups were discarded. From these processed files, 44 file pairs were generated. Each file pair consisted of a file used for training, for which the labeled groups from 43 files were used, and a test file, which contained all the groups from the one file that was left out, resulting in a leave-one-out set. Additionally, the annotations of the training file of each file pair were used to train the weights in the dynamic network model (see section 2B).

In the test phase, the groups in the test file are used as input for the dynamic network (see figure 1). Subsequently, the possible classes that the group can represent are initiated as event hypotheses in the network. The weight between the group and the event hypotheses is the probability of each class given by the *k*-NN classifier. If the event cannot be classified and is labeled as noise, the weight is set to the prior probability that the event occurs. Based on these events, the network forms a hypothesis of the location, which in turn initiates expectancies of certain sound events that might follow. The results of this combined approach are the mostly likely events that explain the segmented groups, given the identified sound events and their predicted location.

The most likely events according to the *k*-NN classifier and the combined model are compared to the annotations through the *F*-measure. The *F*-measure is used in information retrieval to test the effectiveness of the performance of a system[20], for example a search engine. The *F*-measure is computed as the harmonic mean between the recall, which represents whether relevant results are retrieved, and the precision, which represents whether irrelevant results are not retrieved. Applied to the results of automatic sound identification, precision is a measure for the fraction of time the identifications are correct, and recall is a measure for the fraction of identifications that are made out of the amount that should have made.

C. Results

The success of the dynamic network model as it is applied in this study is dependent on whether the location prediction is correct. The location predictions of the test files are listed in table 2. The number of test files at each location is in parentheses behind the location name. The location predictions of the 7 test files of recordings during walking are not included, because they cannot be assigned to a single location. The top 1 indicates how many location predictions are correct on average for a specific location (the spread in standard deviations is given in parentheses). The model has an activation or confidence value for all the location hypotheses. Therefore, if the best prediction is not correct, the second best might be. The top 2 and 3 specify whether the correct location is among the second or third best predictions.

Table 2: Results (average and spread) of location predictions.

Location	Top 1	Top 2	Top 3
City center (7)	0.01 (0.02)	0.02 (0.02)	0.03 (0.05)
Graveyard (7)	0.01 (0.01)	0.01 (0.02)	0.17 (0.14)
Museum (8)	0.24 (0.17)	0.77 (0.29)	0.89 (0.17)
Traffic lights (7)	0.06 (0.06)	0.20 (0.20)	0.67 (0.32)
Train station (8)	0.75 (0.18)	0.92 (0.08)	0.94 (0.06)

Only two locations can be predicted well, the train station and the museum, because some of the sounds the model can identify are very specific for one of these two locations, such as train sounds for the train station. In contrast, many of the other sounds the model can identify well, such as cars and speech, are generic, and can be heard at any of the locations. Therefore, the location prediction is not reliable in many test files.

The location prediction is based on the classified segmented groups, and used to select the most likely label for the group. Of all 54 annotated classes, 12 classes are identified (segmented and labeled) by the combined model (the segmentation algorithm, the *k*-NN classifier, and the dynamic network). These 12 classes are the classes that are mostly annotated. Hence, the *k*-NN classifier and the dynamic network model can learn them better than classes that occur infrequently. Table 3 shows the *F*-measure, precision and recall of the identifications made by the *k*-NN classifier (K) and by the dynamic network model (D) for the 12 classes. The number of test files (out of the total of 44) in which at least one instance of a class was found by either one of the models, is given in parentheses behind the class name. The *F*-measures that are 0 for both models are not included in the mean values in the table. The bottom row indicates the measures weighted for the number of test files. On average, the dynamic network model improves the *F*-measure, mostly through an increased recall, which means that more correct instances of annotations are found than with the *k*-NN classifier.

Table 3: F-measure, precision, and recall of *k*-NN classifier (K) and dynamic network model (D).

Sound class	*F*-measure (K / D)	Precision (K / D)	Recall (K / D)
Bird (6)	0.02 / 0.17	0.34 / 0.65	0.01 / 0.12
Braking train (3)	0.20 / 0.09	0.15 / 0.12	0.32 / 0.14
Bus (8)	0.10 / 0.23	0.22 / 0.19	0.09 / 0.41
Car (27)	0.45 / 0.36	0.62 / 0.53	0.43 / 0.35
Footsteps (12)	0.02 / 0.17	0.49 / 0.71	0.01 / 0.14
Passing train (2)	0.73 / 0.73	0.62 / 0.62	1 / 1
Pressure cleaning (1)	0 / 0.08	0 / 1	0 / 0.04
Speech (15)	0.03 / 0.18	0.52 / 0.33	0.02 / 0.16
Starting train (2)	0.22 / 0.13	1 / 0.50	0.12 / 0.08
Truck (3)	0.05 / 0.25	0.63 / 0.88	0.02 / 0.15
Truck stationary (1)	0.09 / 0	1 / 0	0.05 / 0
Wind (24)	0.10 / 0.19	0.48 / 0.34	0.08 / 0.25
Weighted average	**0.18 / 0.24 (+33%)**	**0.50 / 0.46 (-8%)**	**0.17 / 0.26 (+53%)**

4. CONCLUSIONS

In the previous section we have demonstrated that a model that combines both signal-driven algorithms and knowledge in the form of the predicted location, improves the identification of sound events in a real-world environment. The overall results might not seem impressive, but this is (partly) explained by the performance measure. The *F*-measure is based on the overlap of the annotations and the labeled groups. Therefore, it is dependent on both the annotations and the detection algorithm. Annotating sound is a complex process. The annotators did not only use information in the sound, but also knowledge of the environment, because they were present during the recordings. We cannot determine to what extend the annotations are based on the sound or on their knowledge. Some annotated sound events can even hardly be identified by a human listener that has to rely on the audio signal alone.

In contrast, the detection algorithm only relies on the sound signal. This signal is uncontrolled and thus very challenging for the algorithm that segments relevant parts. Furthermore, the

recordings contain a wide variety of sounds events, most of which occur only a few times in all the recordings. To be able to learn the patterns of a sound event, the *k*-NN classifier (or any other classifier) needs more examples than were available of most classes in the data set in this study.

These observations demonstrate that modeling the context is essential to achieve robust event identification in real-world environments. Indeed we have shown that context, in the form of location, substantially improves event identification, even though it is so far only based on acoustic information. Since the dynamic network model relies on the segmented groups, it cannot identify events that are not segmented. Additionally, the location is not predictive for many generic classes, such as speech and cars. However, the generic classes occur most often, and are best classified by the *k*-NN classifier. In other words, the sparse events are the events that are good predictors of a location, while these are the hardest events to learn, because they are sparse. Fortunately, the dynamic network model is not limited to process acoustic information. In another study we show that the dynamic network model can also be used to improve visual robot localization[21]. Because the model can receive input from different modalities, it can combine multiple modalities and factors in a single system that returns a single analysis. We plan to integrate information from multiple sources of knowledge so that the context is modeled more profoundly.

In summary, the combined model provides a new way to analyze soundscapes by identifying its components. Because these components are also based on knowledge of the context of the acoustic events, they are a first approximation of meaningful events. However, to improve the identification of these components in the complexity of real soundscapes, we require a combined development of segmentation algorithms and models that can include non-acoustical factors. Furthermore, we will study human perception in parallel, so we can validate the model for soundscape analysis. Vice versa, the development of a system to analyze a soundscape automatically might increase our understanding of human soundscape perception.

ACKNOWLEDGMENTS

This work is supported by the Foundation INCAS[3] (Assen, The Netherlands). Additionally, Maria Niessen's work is supported by SenterNovem (Dutch Companion project grant no. IS053013), and Dirkjan Krijnders' work is supported by The Netherlands Organization for Scientific Research under Grant 634.000.432 within the ToKeN2000 program. Finally, Maria Niessen thanks Nicolas Maisonneuve, Matthias Stevens, Peter Hanappe, and Luc Steels at Sony CSL (Paris) for their collaboration.

REFERENCES

1. H. Fastl, The psychoacoustics of sound-quality evaluation, *Acustica* **83**, pp. 754-764, (1997).

2. C. Guastavino, Categorization of environmental sounds, *Canadian Journal of Experimental Psychology* **61***,* pp. 54-63, (2007).

3. B. Schulte-Fortkamp and A. Fiebig, Soundscape analysis in a residential area: An evaluation of noise and people's mind, *Acta Acustica United with Acustica* **92***,* pp. 875-880, (2006).

4. J. Blauert and U. Jekosch, Sound-quality evaluation - A multi-layered problem, *Acustica* **83**, pp. 747-753, (1997).

5. J.A. Ballas, Common factors in the identification of an assortment of brief everyday sounds, *Journal of Experimental Psychology: Human Perception and Performance* **19***,* pp. 250-267, (1993).

6. E. Maris, P.J. Stallen, R. Vermunt and H. Steensma, Evaluating noise in social context: The effect of procedural unfairness on noise annoyance judgments, *Journal of the Acoustical Society of America* **122***,* pp. 3483-3494, (2007).

7. D. Dubois, C. Guastavino, V. Maffiolo and M. Raimbault, A cognitive approach to soundscape research, *The Journal of the Acoustical Society of America* **115***, 115*, pp. 2495, (2004).

8. M. Zhang and J. Kang, Towards the evaluation, description, and creation of soundscapes in urban open spaces, *Environment and Planning B-Planning & Design* **34***,* pp. 68-86, (2007).

9. K. Genuit and A. Fiebig, Psychoacoustics and its benefit for the soundscape approach, *Acta Acustica United with Acustica* **92**, pp. 952-958, (2006).

10. H. Hansen and R. Weber, Semantic evaluations of noise with tonal components in Japan, France, and Germany: A cross-cultural comparison, *Journal of the Acoustical Society of America* **125**, pp. 850-862, (2009).

11. M. Raimbault, C. Lavandier and M. Berengier, Ambient sound assessment of urban environments: field studies in two French cities, *Applied Acoustics,* **64**, pp. 1241-1256, (2003).

12. L. Yu and J. Kang, Effects of social, demographical and behavioral factors on the sound level evaluation in urban open spaces, *Journal of the Acoustical Society of America,* **123**, pp. 772-783, (2008).

13. J.D. Krijnders, M.E. Niessen and T.C. Andringa, Sound event identification through expectancy-based evaluation of signal-driven hypotheses, *Accepted for Pattern Recognition Letters*, (2010).

14. M.E. Niessen, L. Van Maanen and T.C. Andringa, Disambiguating sound through context, *International Journal on Semantic Computing* **2**, pp. 327-341, (2008).

15. A.S. Bregman, Auditory Scene Analysis: The Perceptual Organization of Sound, MIT Press, Cambridge MA 1990.

16. D.P. Ellis, Using knowledge to organize sound: The prediction-driven approach to computational auditory scene analysis and its application to speech/nonspeech mixtures, *Speech Communication* **27***,* pp. 281-298, (1999).

17. M.R. Quillian, Word concepts: A theory and simulation of some basic semantic capabilities, *Behavioral Science* **12**, pp. 410-430, (1969).

18. J.L. McClelland and D.E. Rumelhart, An interactive activation model of context effects in letter perception: I. An account of basic findings, *Psychological Review* **88**, pp. 375-407, (1981).

19. G. Salton and C. Buckley, Term-weighting approaches in automatic text retrieval, *Information Processing & Management* **24**, pp. 513-523, (1988).

20. C.J. Van Rijsbergen, Information Retrieval, Butterworths, London 1979, pp.112-140.

21. M.E. Niessen, G. Kootstra, S. De Jong, T.C. Andringa, Expectancy-based robot localization through context evaluation, to appear in *Proceedings of the International Conference on Artificial Intelligence*, Las Vegas, 2009.