ACOUSTICS 2012

Listening strategies in a soundscape annotation task

R. Van Der Linden and T. C. Andringa

University of Groningen, Nijenborgh 9, 9747 AG Groningen, Netherlands
vdlinden@ai.rug.nl

Listeners employ a variety of strategies while annotating environmental sounds. The soundscape annotation tool we developed can track and record the behaviour of a subject performing a semantic annotation task. These data, combined with the actual annotations, provide insight in the strategies the annotators use in the task. We performed an experiment in which 19 participants annotated two real-world soundscape recordings. Two conditions were tested: one in which the annotators worked under time pressure, and one in which the annotators had time to add detailed annotations and sound classes. We give a explanation of our findings in terms of annotator strategies.

1 Introduction

Different listeners can have a different experience of the same sonic environment [4]. *Bottom-up attention* is the mechanism that selects what part of the stimulus the listener decides is worth further analysis and top-down attention allows detailed analysis. We assume that identifying the strategies that listeners employ when deciding where to aim the focus of attention is a key to understanding a listener's sonic perception. We propose a software tool that enables researchers to track listener's choices in listening to a real-world audio signal. As a validation of this method we present the results from a soundscape annotation experiment. In this experiment we asked listeners to annotate an environmental sound recording. We show how these annotations can be analyzed and how annotations from multiple annotators can be combined to come up with a 'best' annotation set.
In an experiment we test the influence of time pressure on the annotations and annotation behaviour of the participants. We show how both the annotations and the behavioural user data can be used to gain insight in the annotation task.

1.1 Listening modes

Among the interpersonal variance in sensory experiences of listeners, Gaver [3] proposes that humans can employ two distinct listening modes: *musical listening* and *everyday listening*. The former refers to a way of listening that pays attention to the structure and physical characteristics of the sound (harmonic structure, frequencies etc.), the latter is the focus of this article; it refers to a listener reporting to hear 'a passing aircraft' or 'someone playing a violin'. In everyday listening sonic experiences of the soundscape are given in terms of sound sources or activities and events in the environment.

1.2 Attention in the annotation task

Auditory attention is the process that results in a selection of the sonic 'input signal' to be processed up to a conscious level. Rather than a listener ignoring the 'irrelevant' part of the input, evidence suggest that all available sonic information is processed up to a rudimentary level that allows attention selection. *Hearing* is a continuous process, it is bottom-up processing of all available auditory information, whereas *listening* refers to selectively attentending to and further processing of task-relevant parts of the stimulus [5]. In this task the participant is forced into the mode of *listening*.

2 Method

2.1 Annotation on a timeline

The annotation paradigm that is put forward in this article is one where a human listener annotates sound sources on a timeline, aided by a graphical representation of the spectrum, here a cochleogram [1].

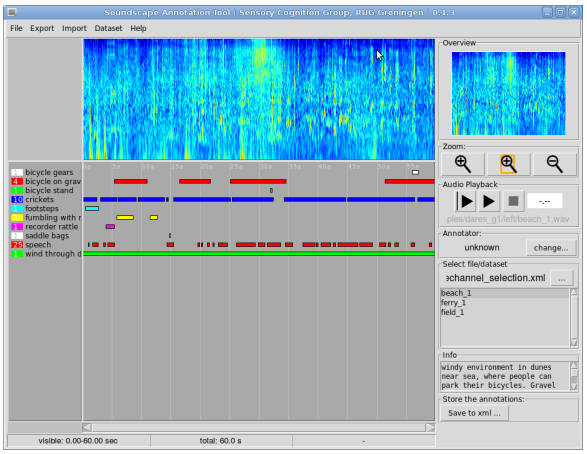

Figure 1: Screenshot of the annotation tool. The interface consists of three mayor parts: a cochleogram visualization of the loaded sound signal (top left), a window that presents the stored annotations sorted by class (bottom left) and a panel that provides controls for the audio playback (middle right).

Annotations here serve as an abstraction of sound events that refer to (the presence of) a sound source or combination of sound sources, and therefore are modeled as a short textual description (typically one or two words) of the source and an interval during which the sound event is audible.

2.2 Sound Annotation Tool

The annotation tool proposed in this article is an extended reimplementation of the Matlab-based annotation software used in [7].
A screenshot of the annotation tool described above is shown in figure 1. Annotations are made by pressing the shift button and dragging the mouse over the cochleogram. A popup dialog window allows the annotator to select a class or description for the annotation. Zooming functionality is provided to allow detailed inspection of the cochleogram.

While the cochleogram in principle allows the annotator to make annotations solely based in the spectrogram, participants were instructed to listen first and to use the visual support as an aid, not as primary source of information. The role of the visual support is however unclear and needs further investigation because the annotation task as formulated here is not purely auditory and requires balancing between listening and adding annotations in the graphical interface. In addition time pressure leads to strategic behavior.

Together these choices (providing visual aid, multitasking, time pressure) lead to an annotation paradigm that aims at collecting information on what can be heard in a recording, in contrast to registering and modeling human audition in a listening task. While pursuing the former goal makes the

annotations suitable for machine learning purposes, the latter requires a different setup.

2.3 Stimuli

To allow comparison with earlier results a recording from [8] was selected. This recording was divided into two equal parts to compare two conditions: a single and double duration of the time available for annotation. The recording was made in a quiet environment in the rural town of Assen, located in the north of the Netherlands. The recording setup consisted of a single-channel omni-directional microphone placed on a tripod connected to a digital audio
recorder. The data was recorded at a sampling rate of 48 kHz and 24 bits per second, stored in PCM Wave format. The recording setup was located in the front yard of a three story villa, a few meters away from a lively road where pedestrians, cyclists and other kinds of traffic pass.

2.4 Participants

Subjects were recruited among student population of the University of Groningen, ranging in age between 18 and 26 year old. They were not trained in annotating soundscapes or otherwise trained to analyze environmental audio recordings. Computer experience ranged from moderate to experienced user.
Self-reported normal hearing was a prerequisite to take part in the experiment. Participation in this experiment was voluntary; about half of the subjects received a small financial compensation or study credits for a courser on Perception.

2.5 Instructions

Before starting the experiment the participants received instructions in pairs on the functionality of the annotation software. Each major element of the GUI was introduced and a short demonstration of the annotation process was provided. Proper functioning of the software and hardware was ensured before starting the experiment.

Three participants did not have Dutch as their native language (but Chinese, Spanish and German), the instructions therefore were provided to them in English. One native Dutch participant also received instructions in English.

The task provided to the participants was to report all sound events in the recording. Choice of classes was free, but limited to 20. Participants were instructed to divide the available time over the whole recording. When arriving to the end of the recording before the experiment was finished, participants were allowed to start from the beginning of the recording.

2.6 Conditions

In this experiment we study the effect of time pressure on participant's quatantitative and qualitative performance in the annotation task. Therefore two conditions were created:

1. **Single annotation time:** the time provided to the subjects to annotate the recording is equal to the duration of the recording; this implies that only the most prominent sound sources can be annotated and pausing the playback of audio inherently prevents the subject from working through the whole recording.
2. **Double annotation time:** the trial time in which the annotation had to take place was set to twice the duration of the recording itself. This enables the subject to listen to interesting portions of the recording twice and to provide more detail in the annotations.

A training effect may be observed for naive subjects; no training phase was applied, therefore the two conditions are applied to both recordings in two conditions. The training effect on annotations or strategies is not measured here; this would require two more conditions.

2.7 Annotation performance:

In many applications it is necessary to compare annotations to the best known set of annotations. What the 'best' annotation set for a particular purpose is depends on the application of the annotations; when studying attentive processes in the auditory domain the 'best' annotation set may be very different from that suitable for training an automatic sound recognition algorithm.

2.7.1 Calculating the 'mean' annotation of a group of annotators

When a ground truth (or *golden standard*) is absent the 'mean' annotation set of a group can be used as ground truth. The following steps were taken to obtain this 'mean' annotation set: Firstly, annotations were mapped (by hand) onto a small set of sound sources classes which we will denote as 'common' classes. Secondly, the time span of the recording was segmented in blocks of 60 seconds. For each block and each class, the number of annotators that made an annotation of that class in that block is counted. This results in a vector that indicates how many annotators noticed and annotated a sound of this class. A threshold can then be applied to the vector values, resulting in a new annotation set that can be seen as the annotations most annotators agree on. [6] chooses a threshold of 30 percent, meaning that when 30 percent of the annotators agree on the presence of a sound source, this is adopted as the consensus annotation for a group of annotators. This threshold value is also adopted for the current experiment. The choice of this threshold is arbitrary; each application of the annotation paradigm may require its own threshold setting.

3 Annotator behaviour

During the annotation sessions the following numerical data were extracted or calculated:
Audio player actions: The number of times the audio player was started, stopped or paused the audio player.
Zoom actions: Subjects were free to zoom the cochleogram representation in and out (on the time axis, the frequency axis was fixed to allow identification of sonic information in the frequency plane).
The number of annotations created by the subject. Not all annotations persisted during the experiment, some were deleted: the number of deleted annotations is also mentioned.

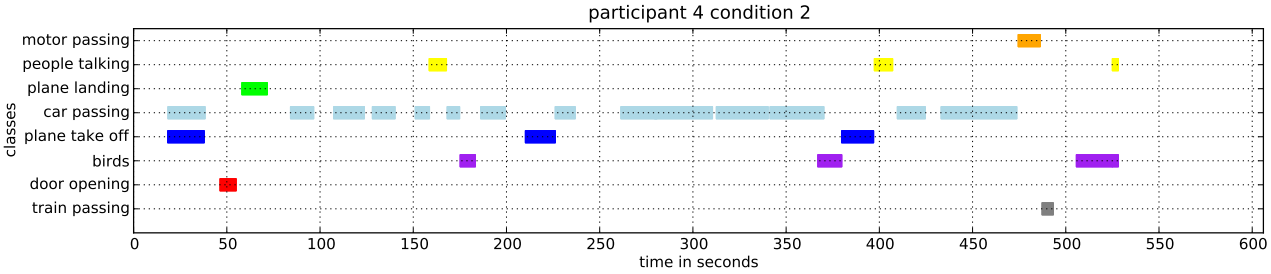

Figure 2: A selection (first 180 seconds) of semantic annotations that were produced by one participant, indicating the presence of birds, cars, people and the event of a door opening. The annotator mistakenly heard a plane - this was actually the sound of a car. The classes were chosen by the participant. The order of the classes reflects the order in which they were added by the participant. This participant used a visual strategy to first add two annotations near the end of the file.

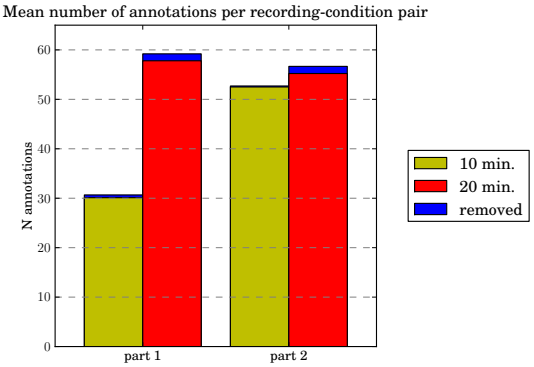

Figure 3: Graphical presentation of the data in table 1. The blue bars represent the mean number of annotations that were removed during the trail. Note that the first and the fourth bar correspond to condition A, the second and third bar to condition B. The number of deleted annotations was extracted from the log file, the other data is calculated from the annotation files.

Table 1: Mean annotation counts for each condition (20 minutes, 10 minutes) and recording (Part 1, Part 2) pair.

	Part 1		**Part 2**	
	10 min	20 min	10 min	20 min
Persistent annot.	31	60	53	57
Removed annot.	0.5	1.5	0.0	1.0

4 Results

4.1 Annotations

Annotations were collected for each of the trials. As an example, a fragment of the annotations is shown in figure 2.

Table 3 presents the annotation counts for each condition. Figure 3 shows a difference between the two conditions and between the two annotation sessions (and the corresponding recordings). These results are discussed in the next section.

4.2 Behavioral data: event frequencies

Two indicative statistics were calculated for each trial and compared between conditions: the number of *play intervals* and the number of *zoom actions*. The means are presented in table 2.

Table 2: User data from the experiment for each recording and condition: the mean number of *play intervals* that indicates the number of fragments in which the annotator listened to the recording, and mean number of *zoom actions*, number of times the annotator changed the cochleogram view by zooming in, zooming out, or shifting the window to the left or right.

Condition/Recording		**Play intervals**	**Zoom actions**
Part 1	Short (10 min)	22	56
	Long (20 min)	81	64
Part 2	Short (10 min)	14	42
	Long (20 min)	71	111

4.3 Inter-annotator consistency

We define annotator consistency as the fraction of annotators that agrees on the existence of a sound source for a certain point in time (under the segmentation described above). This measure is calculated per class for the second annotation trial of each participant (performed on the second half of the recording) to reduce the influence of the training effect on the analysis. As an example, these plots for two classes are plotted in figure 4.

5 Discussion and Conclusions

5.1 Annotation counts

Table 1 and figure 3 present average annotation counts per annotation session (on both parts of the recording). In the first session the annotators from both conditions produce on average the same number of annotations per (trial) minute; however in the second session the number of annotations for each condition is on average comparable.

We explain these findings by hypotising that the time constraints in first session determine how annotators perform in the second session: annotators that are under time pressure in the first session do not 'relax' in the second session, but on average produce about the same number of annotations per minute. In contrast, annotators who are under more time pressure in the second session aim to produce the same level of detail as they did in the first session. The training effect (annotators becoming acquainted with the task) may also have influenced these results.

Figure 4: Annotator consistency plots for annotations of two classes, namely *car* and *people*. These plots were derived from the second trial.

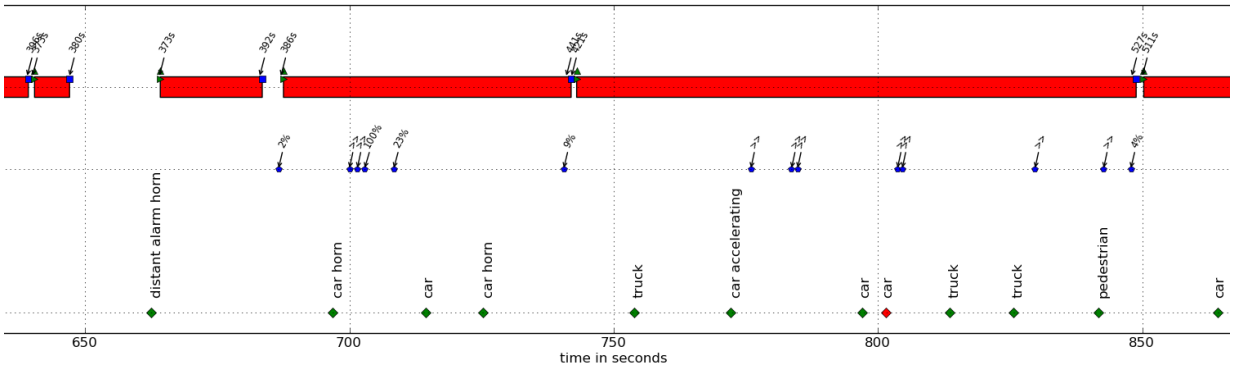

Figure 5: Example fragment of a timeline that was derived from the log file of one run on a time axis; units are seconds. The red line on top indicates the time intervals that the audio player was running; the textual description represents the starting point of the audio playback, in seconds. The blue dots in the middle indicate zoom actions, the textual description refers to the zoom level (on the time-axis) set by the action. The diamond-shaped icons at the bottom indicate the addition of an annotation.

Figure 3 also shows the average number of deleted annotations; in the short annotation session almost no annotations are removed, probably because the for this the annotator has to stop the recording and decides that this cost is not worth the effort.

5.2 Annotation consistency

The plots allow for detailed assessment of the interplay between sound sources; this article does not provide the space for a thorough analysis of all results. As an example, three major observations in the current experiment are reported:

Traffic sounds are prominent to listeners: For the class that was annotated most, is 'car'. The plot for 'car' shows clear boundaries for the presence of car sounds. The curve hits the 1.0 level several times in both conditions, indicating that all annotators have total consensus on the presence of a car (or cars) for that time point. Annotators also agree fully on the absence of cars for certain time regions.

Salient sound sources result in short peaks: From the plots for classes that denote sudden sounds' (plots not shown here) one can observe that salient, sudden sounds result in sharp, high peaks in the plots. This effect shows for the class 'bang' that holds all class descriptions that correspond to sudden, explosive acoustic events, such as 'bang', 'door' and 'car door closes'. Apparently these events are salient enough to appear in the annotation sets of most participants.

Human sounds attract attention: Most annotators clearly distinguish and annotate human sounds, slightly more in the long session than in the 10 minutes session, but still present in the latter. Apparently there were some occasions of human speech close to the recording setup, as the plots for both recordings show clear peaks. When comparing the plots of 'car' and 'people speaking', the figure suggested a reciprocal relation between these two classes: when 'people speaking' are annotated, the annotation counts for 'car' are low, but not zero, which indicates that there was a car audible for that time period. We hypothesize that 'people soeaking' attract (auditory) attention in such a way that car sounds can be missed completely by most annotators.

Furthermore the plots reveal some interference of different classes in the long annotation session that is absent under the 10 minutes time constraint, which suggests that annotators are more specific in the 20 minutes session than in the shorter session.

5.3 Discussion of annotation consistency method

The method described in section 4.3 provides a reasonable 'summary' of the annotations from different participants. There are however some potential problems with this approach. For example, the choice of 'common' classes is arbitrary; different common classes result in quite different plots. Applying a (hierarchical) ontology may solve this. Also, annotations may overlap: when an annotator adds two annotations ascribed to two classes and both these classes are mapped to the same 'common' class, one annotator attributes double to the sum of annotation. This can be corrected by checking the overlap in annotations from one participant, but this results in loss of information. Some annotators added annotations for very long time regions indicating that the sources was 'always present'; in this approach, this causes the 'signal' to never become zero and the whole plot to shift upwards. It is not desired, but deleting these long annotations causes loss of information.

5.4 Plotting annotator behaviour

Figure 5 presents a fragment of the user data that was collected for one annotation session for one participant, shown here as an example. From these data and corresponding plots observations can be made on the annotation strategies that annotators employ. We list some observations that can be made from the plot:

1. **Deviating annotator behaviour can be recognized:** In circumstances where the participant cannot be continuously monitored, the statistics and plots allow for (automatic) recognition of anomalies in the experiment. When no action was registered for a longer period, this may be an indication of problems with the software or the task.
2. **Annotation styles become visible:** As the example plot shows, annotators exhibit multitasking behaviour; they add annotations while the audio is still playing. This is time-efficient, but directs participant's attention the popup window instead of the sound. The data may provide insight in (the effect of) these strategies. This allows researchers to steer the annotator's behaviour in the desired direction.

5.5 Future research: annotation methods

We have shown that the software tool and sound annotation method proposed in this article yield valuable annotations (also in real-time). Furthermore we argue that the behavioural data described in this article provides detailed insight in the strategies annotators employ.

Future research will include extensive assessment of the behavioural data. A variety of annotation methods will be compared to provide a set of annotation methods from which the most appropriate method can be picked for a certain application.

5.6 Potential application: Soundscape reseach

The results show that the proposed method - analysing listener's behaviour through an annotation task - provides insight in the strategies humans employ to abstract meaning from a sonic environment. This makes the annotation task and associated tool a feasible candidate for studying the role of listening strategies in soundscape research, which is the study of the relation between a listener and it's sonic environment [2]. This application will studied in future research.

Acknowledgments

This research was partly funded by the Province of Drenthe, the Netherlands.

References

[1] T.C. Andringa. *Continuity preserving signal processing*. PhD thesis, 2002. University of Groningen, available from http://irs.ub.rug.nl/ppn/237110156.

[2] D. Botteldooren, C. Lavandier, A. Preis, D. Dubois, I. Aspuru, C. Guastavino, L. Brown, M. Nilsson, and T.C. Andringa. Understanding urban and natural soundscapes.

[3] W.W. Gaver. How do we hear in the world? Explorations in ecological acoustics. *Ecological psychology*, 5(4):285–313, 1993.

[4] C. Guastavino, B. Katz, J.D. Polack, D.J. Levitin, and D. Dubois. Ecological validity of soundscape reproduction. *Acta Acustica united with Acustica*, 91(2):333–341, 2005.

[5] S. Harding, M. Cooke, and P. Konig. Auditory gist perception: an alternative to attentional selection of auditory streams? *Lecture Notes in Computer Science*, 4840:399, 2007.

[6] J.D. Krijnders. Differences between annotating a soundscape live and annotating a recording. Internoise 2010, Lisbon, Portugal, 2010.

[7] J.D. Krijnders and T.C. Andringa. Soundscape annotation and environmental source recognition experiments in assen (nl). Internoise 2009, Ottawa, Canada, 2009.

[8] JD Krijnders, ME Niessen, and TC Andringa. Sound event recognition through expectancy-based evaluation ofsignal-driven hypotheses. *Pattern Recognition Letters*, 31(12):1552–1559, 2010.