inter.noise 2010
noise and sustainability
13-16 June 2010 . Lisbon . Portugal
www.internoise2010.org

Differences between annotating a soundscape live and annotating behind a screen

Johannes D. Krijnders;Tjeerd C. Andringa

Affiliation: INCAS³; University of Groningen
e-mail: dirkjankrijnders@incas3.eu; t.c.andringa@ai.rug.nl

Abstract
What sensory information do we use when we are asked to listen? To answer this question we asked participants to annotate real world city sounds in two conditions. In the first two conditions participants were present and annotated during recording. In the first condition the participants could see the environment. In the second annotation condition, participants sat behind a screen that blocked their view. The first condition corresponds to a normal situation for humans.

Keywords: Annotation, Soundscape.

1 Introduction

What people hear when they listen is an open question. For example do you hear an aircraft, “a quasi-harmonic tone lasting approximately 3 seconds with smooth variations in the fundamental frequency and the overall amplitude” [1] or 70 dB? The person in the street will probably answer the first, someone trained to describe the structure of sound will answer the second and the noise legislation is concerned with the third. Also if you here an metallic cracking sound with some noise, could you recognize it as a bicycle? Maybe in the Netherlands, but in the USA, and what if you don’t know were the recording was made?
The case of the aircraft describes the answers the question on the level of the relevant description. Gaver [1] termed the first answer level the “everyday” listening mode and the second the “musical” listening mode. The examples of Gaver ask specifically about the sound source. Guastavino et al.[2] researched three different settings and asked their participant afterwards to describe the sound. In the resulting discourse they find that people mention mainly sound sources when talking about the foreground sounds and more about themselves if they talk about the background. They also show big differences in the way people report about the background depending on the situation they heard the sound. Three ways were tested, field study, stereophonic reproduction in the lab and ambisonic reproduction in the lab. The stereo reproduction resulted in a strong decrease in the amount of subject-centered discourse.

We are interested in the process of annotating sound events[3]. When annotating sounds we are interested in the “everyday” description of events. Both to learn about how people evaluate sound events and to find out which parameters influence the perception of these sound events and by how much. This last information may help to improve and evaluate automatic sound recognition systems.

Figure 1 - The participants annotating at location two. Behind the blue screen are the participants in the condition without sight.

2 Experimental setup

The experiment was performed with two conditions on two locations in Assen (60.000 inhabitants in the north of the Netherlands).

2.1 Location

This pilot study was performed in the city of Assen on two locations. Location one was in a park situation facing a tarmac road with a bicycle path. Opposite the road was a large recreation center (ice-skating rink, pool, indoor sports facilities) and a construction site. Location two was in a built area facing a cobblestone road. Traffic on the road was about equal at both locations. Both locations were chosen such that most sounds would come from the front, in view of the participants.

2.2 Participants and task

The participants were students of the course “sound recognition” given at the university of Groningen, all male, between 20 and 25 years of age. For both conditions there were 8

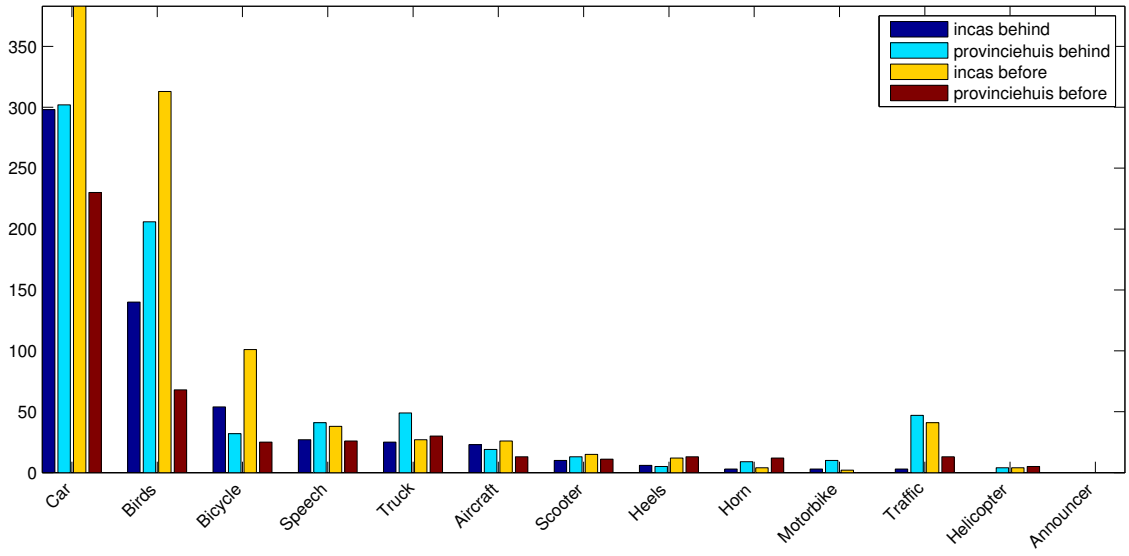

Figure 2 - Standard categories and average number of annotations made per condition and location.

participants, of which 5 dutch speaking. For the analysis only the dutch speaking participants were considered.
The task given was to listen and write down the heard sound sources at their estimated maximum. The participants were supplied with a list of sound source classes and abbreviations to standardize and facilitate the annotation. For the annotations timelines at which the sources could be marked were supplied. A timer presented on a laptop screen for all to see was used to synchronize all annotations. Participants were instructed to use the supplied sound source classes, but were allowed to use their own classes if the supplied classes did not fit.

2.3 Conditions

Two conditions were tested in the field, one where the participants had full view of the scene, the other where participant were behind a wind screen blocking their sight of the scene while minimizing the differences in auditory circumstances. Everything, except the view, was equal in both conditions.

3 Results

Figure 1 shows the number of occurrences of the standard classes in both conditions and both locations. The number of occurrences does not point to a consistent effect of the wind screen. Two more analyses were done, one based on time series generated by the annotations and one based on the added classes.

3.1 Time series

To compare the time series resulting for the annotations we replace each annotation with a Gaussian with a standard deviation of 1.5 seconds and a maximum of one. This to compensate for small time differences in the annotation. To create a ground truth or golden standard the sum of the time series is thresholded at 30% of the participants, corresponding to the intervals that at least 30% of the annotators agree. An example of resulting ground truth can be seen in figure 3 as the black line. For a single annotator the same procedure is used resulting in the gray line in figure 3.

Table 1 Standard classes and their abbreviations as given to the annotators. The third to sixth column indicate the average f-measure and standard deviation of the agreement between a single annotator and a ground truth made of the other annotators in the same condition.

Abbreviation	Class	Location one		Location two	
		no visual	visual	no visual	visual
Ca	Car	0.43 ± 0.04	0.43 ± 0.03	0.31 ± 0.03	0.31 ± 0.04
Bi	Birds	0.15 ± 0.03	0.30 ± 0.06	0.16 ± 0.06	0.02 ± 0.02
Sc	Scooter	0.07 ± 0.02	0.05 ± 0.03	0.03 ± 0.02	0.09 ± 0.02
By	Bicycle	0.07 ± 0.02	0.09 ± 0.05	0.00 ± 0.00	0.02 ± 0.01
Sp	Speech	0.08 ± 0.03	0.11 ± 0.04	0.06 ± 0.03	0.07 ± 0.03
Ai	Aircraft	0.11 ± 0.02	0.04 ± 0.05	0.08 ± 0.06	0.07 ± 0.04
He	Helicopter	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Ho	Horn	0.03 ± 0.03	0.00 ± 0.00	0.00 ± 0.00	0.08 ± 0.01
Tr	Truck	0.10 ± 0.06	0.06 ± 0.06	0.01 ± 0.01	0.02 ± 0.02
An	Announcer	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Mo	Motorbike	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Tf	Traffic	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.01	0.00 ± 0.00
He	Heels	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.01

3.1.1 Metric

The F-measure(F) is used as a performance metric [4]. It is calculated as the harmonic mean of precision(P) and recall(R). Precision is a measure for the fraction of time an annotator was correct, and recall is a measure for the fraction of detections that a annotator correctly annotated.

$$P = \frac{TP}{TP+FP} \tag{1}$$

$$R = \frac{TP}{TP+FN} \tag{2}$$

$$F = 2\frac{P*R}{P+R} \tag{3}$$

where TP is the true positive rate, FP is the false positive rate, and FN is the false negative rate.

Table 2 Indicates the average f-measure and standard deviation of the agreement between a single annotator and a ground truth made of the other annotators in the other condition.

Class	Location one	Location two
Car	0.49 ± 0.14	0.36 ± 0.09
Birds	0.19 ± 0.12	0.13 ± 0.12
Scooter	0.56 ± 0.29	0.30 ± 0.31
Bicycle	0.13 ± 0.08	0.00 ± 0.00
Speech	0.40 ± 0.22	0.24 ± 0.21
Aircraft	0.25 ± 0.21	0.18 ± 0.16
Helicopter	0.00 ± 0.00	0.00 ± 0.00
Horn	0.18 ± 0.29	0.19 ± 0.16
Truck	0.18 ± 0.17	0.06 ± 0.09
Announcer	0.00 ± 0.00	0.00 ± 0.00
Motorbike	0.00 ± 0.00	0.00 ± 0.00
Traffic	0.00 ± 0.00	0.01 ± 0.03
Heels	0.00 ± 0.00	0.05 ± 0.08

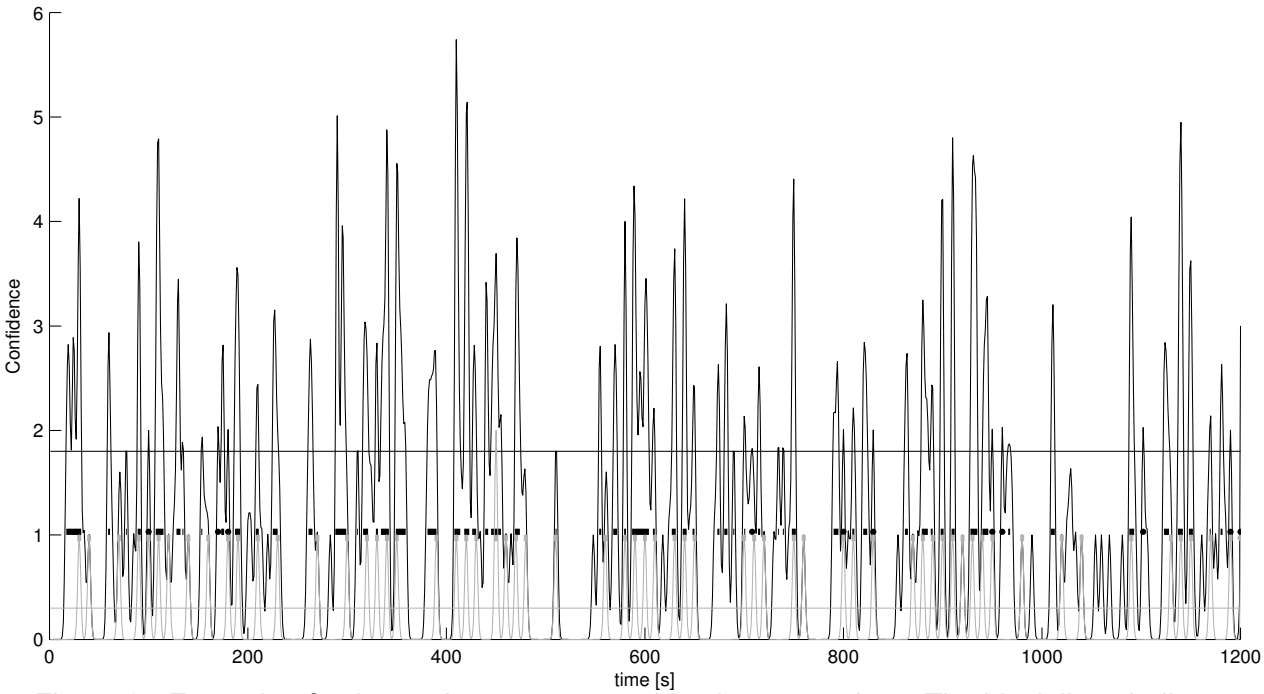

Figure 3 - Example of a timeseries as generated by the annotations. The black lines indicate the summed annotations of all annotators in one condition and the thresshold used (30%). The gray lines indicate the timeseries generated by a single annotator and the thresshold used (30%). The stripes around y = 1, indicate the resulting binary time-series that is used for for the F-measure. This time-series is for the class “car”.

3.1.2 Inter-annotator agreement

To establish a base-line of agreement between annotators we calculate the F-measure between a single annotator and the ground truth made from all other annotators in the same condition. The results can be seen in table 1.
The F-measures are not high, only the often occurring class “Car”, which also has a clear maximum as a high(er) F-measure. The standard deviations are low, meaning that the differences in disagreement are low.

3.1.3 Inter-condition agreement

To compare the conditions the agreement of each annotator with the ground truth made from all other annotators in the other condition. was calculated. The results can be seen in table 2. The F-measures are comparable to the F-measures found for the inter-annotator agreement. This entails that blocking the sight of the annotators did not influence the annotation a lot. Note however the the standard deviations are bigger then in the inter-annotator agreement meaning that the individual annotators had different agreement with the other condition.

3.2 Free classes

The annotators were allow to add classes when they deemed that necessary. To compare the groups in both conditions we compared to number of classes added by the annotators. Both the number of classes added as well as the total number of occurrences can be seen in table 3. The differences here seem to indicate that in the visual condition annotators add less classes, but annotate more of them. This could have to do with the ambiguity of sounds in the no-visual condition. However the differences are too close too call.

Table 3 - The number of added classes and their number of occurrences in all conditions.

Location	# new classes		Occurrences	
	no visual	visual	no visual	visual
One	18	17	34	45
Two	25	22	55	71

4 Conclusions and future work

We found that limiting peoples sight did not influence their capacity to identify sounds when they were present at the scene. The number of classes added and annotated could indicate that the sounds were more ambiguous in the no visual condition.
Future work will include repeating this experiment and increasing the number of conditions, including lab studies. Also the annotating of sound events will be supplemented by a questionnaire to analyze the discourse used.

References

[1] Gaver. What in the World Do We Hear?: An Ecological Approach to Auditory Event Perception. *Ecological Psychology*, vol 5 (1), 1993, pp. 1-29.

[2] Guastavino and Katz. Perceptual evaluation of multi-dimensional spatial audio reproduction. *The Journal of the Acoustical Society of America,* Vol 116 (2), 2004, pp. 1105-1115.

[3] Krijnders et al., Annotating Soundscapes, *Internoise 2009*, “In CD-ROM”

[4] Hripcsak and Rothschild. Agreement, the F-Measure, and Reliability in Information Retrieval. *Journal of the American Medical Informatics Association,* Vol. 12 (3) pp. 296-298.