Audio Engineering Society

Convention Paper 5404

Presented at the 110th Convention
2001 May 12–15   Amsterdam, The Netherlands

Recognition of Everyday Auditory Scenes: Potentials, Latencies and Cues

Vesa T. K. Peltonen, Antti J. Eronen, Mikko P. Parviainen and Anssi P. Klapuri[1]
Signal Processing Laboratory, Tampere University of Technology
Tampere, FIN-33101, Finland

ABSTRACT

A listening test was conducted where the human abilities in recognizing everyday auditory scenes based on binaural recordings were studied. The accuracy, latency, and acoustic cues used by the subjects in the recognition process were analyzed. The average correct recognition rate for 19 subjects was 70% for 25 different scenes, and the average recognition time was 20 seconds. In most cases, the test subjects reported that the recognition was based on prominent identified sound events.

INTRODUCTION

Computational auditory scene analysis has been an active area of research in last few years [1, 2, 3]. It refers to the computational analysis of an acoustic environment, and the interpretation of distinct events in it. This paper addresses the problem of *computational auditory scene recognition* (CASR) which means the automatic recognition of an acoustic environment and does not necessarily involve analysis down to the level of distinct sound events. Applications of CASR include intelligent wearable devices and hearing aids that sense the environment and adjust the mode of operation accordingly.

Human abilities give a good idea of how accurately and fast an artificial system could potentially operate, and what applications would be realizable. Earlier work in psychoacoustics has mainly concerned human abilities in recognizing single, isolated sound events instead of complex sound mixtures from different environments [4]. The listening test described in this paper aims at finding answers to the following questions:

1. How reliably are humans able to recognize environments based on binaural recording of the auditory scene,

2. How long is the response time, and
3. What are the cues used in making the inferences.

CASR is related to several fields of research. Automatic noise classification is an active subject of research nowadays [5, 6]. It refers to stationary and semi-stationary background noise recognition. In [7], El-Maleh studied the classification of commonly encountered noises in mobile telephone environments (i.e. car, street, bus, babble and factory). Another related field is speech/music discrimination. In the research of speech/music discriminators, many acoustical features have been studied and their efficiency in signal classification has been examined [8, 9, 10].

The remainder of this paper is organized as follows. Firs, the acoustic measurements that were made in a number of everyday auditory scenes are described. After that, the stimuli and the setup of this two part listening test are presented. Then the procedure of the test, and the subjects that participated in the test are reviewed. Subsequently, test results concerning accuracy, latencies and cues are presented and analyzed.

[1] Email: {peltonen,eronen,partsi,klap}@cs.tut.fi

ACOUSTIC MEASUREMENTS

A binaural recording setup was utilized to capture tens of auditory scenes from different contexts. The binaural setup included a standard B&K 4128 head and torso simulator with an associated B&K Nexus amplifier. The head and torso simulator was attached to a stand and employed at an approximate height of a standing person. The data was recorded to a digital recorder in 16-bit and 48kHz sampling rate format.

The recordings were made in everyday domestic and business environments, including e.g. family homes, vehicles, business buildings, and different outside environments such as streets, market places and roads. The recordings took place in two middle-size cities in Finland during the summer 2000.

LISTENING TEST

Stimuli

A total of 25 different environments were included in the two-part listening test, and some scenes had multiple instances. In the first experiment, 34 samples of one minute in duration were used. In the second experiment, a subset of these environments was applied. The 20 samples used in this experiment were approximately three minutes in duration. These samples were picked from the same recordings as those in the first test; however, different temporal sections were selected and the samples were presented in a different order. The recording environments of the stimuli used in the both experiments are listed in Table 1 in the order in which they appeared in the first experiment. The subset of samples used in the second test is highlighted.

A total of 92 minutes of binaural data was selected and downsampled to a sampling rate of 44.1 kHz. The levels of the recordings were appropriately normalized before they were written on two audio-CDs.

Test setup

Most of the tests were performed in a listening room at Tampere University of Technology, Signal Processing Laboratory. A number of blind subjects were tested in the premises of Tampere association for blind and visually impaired. A few tests were also performed at the subjects' houses. In the listening room, the amount of background noise and other interfering elements was negligible. The other premises had some occasional background noise sources evident due to e.g. air conditioning, but care was taken to make the locations as silent as possible and to avoid any interfering movements and noise.

A CD player with adjustable volume control and high-quality headphones were used as test equipment. The listening level was adjusted so that the subjects felt comfortable. If desired, the volume was adjusted again during the test. A supervising person controlled the test and wrote down the answers.

Subjects

A total of 19 normally hearing subjects participated the test, five of which were blind. Blind people were recruited in order to test the hypothesis that blind people would be more experienced in analyzing environments by listening only. Both male and female subjects between 22 and 58 years of age were involved. All except two subjects were Finns. Some of the subjects were involved in audio engineering, but none had significant experience in listening tests.

Test Procedure

A two-part listening test was conducted. The aim of the first test was to find out how accurately and how fast the environment could be recognized. In this part, 34 samples of one minute in duration were used. The subjects were instructed to try to recognize the scene as fast as possible. The subjects were *not* provided with a list of correct scenes, but they were told that the scenes include everyday places, such as vehicles, public and private buildings and outside

	Description		Description
1	**Traffic, 80 km/h route**	18	Street traffic
2	**Railway station**	19	**Wind (water tower)**
3	**Church, concert**	20	**Traffic, trams passing**
4	**Car & Radio, 80km/h**	21	Supermarket
5	**Train**	22	Traffic, 80 km/h route
6	**Street café**	23	**Market place**
7	**Subway station**	24	Restaurant
8	**Amusement park**	25	Children playing, home
9	Car & speech, 40km/h	26	**Nature**
10	**Library**	27	Department store
11	**Street traffic**	28	**People indoors**
12	**Construction site**	29	**Pub**
13	**Supermarket**	30	Traffic, trams passing
14	**Restaurant**	31	Office
15	Presentation	32	Traffic, trams passing
16	Car, 40km/h	33	Replayed music
17	**Kitchen**	34	Bathroom

Table 1. List of the contexts used in the test

environments. If the subjects could not give any exact answer, they were asked to tell whether the place was public or private and inside or outside. The answers and the time to give them were written down by a supervising person, letting the subjects to concentrate on the audition. The response times were measured using the display of the CD player at a precision of one second.

Besides studying the recognition accuracy, the second test aimed at discovering the cues used by human listeners in auditory environment recognition. In this test, 20 samples of three minutes in duration were used. The subjects were allowed to make guesses and refine their answers while listening to the sample. In addition to that, they were asked to tell what information they based their guesses on. Again, the answers and the response times were written down by a supervising person. The subjects were allowed to listen to the whole sample if they desired. After completing the test, the subjects were not asked whether they were acquainted with the scenes, assuming that they are familiar with these everyday auditory environments.

RESULTS

Analysis of the Results

The answers from both tests were processed and the following information was collected for each subject and scene pair:
- The claimed scene
- The response time

Additionally, the cues mentioned by the subjects were collected in the second test. None of the subjects or the answers was discarded in the analysis process.

Accuracy

In Figure 1, the confusion matrix for the first test is shown. The rows of the matrix list the presented scenes and the columns describe the subjects' responses. The values in the matrix are recognition percentages. The overall recognition rate for the first test was 66% and the recognition accuracy for individual environment ranged from 0% to 100%.

Confusions between scenes were relatively rare. The highest off-diagonal value is reached when "*street café*" is recognized as a "*supermarket*" (58%). This may be due to the fact that sounds of a cash register and jingling sounds of change are heard clearly in that recording. The erroneous answers were mostly environments, which do not appear in the test set. An extreme case is the scene "*replayed music*" which was recognized as a concert in 63% of the cases. This context was a recording of classical music replayed from a CD player in a silent listening room.
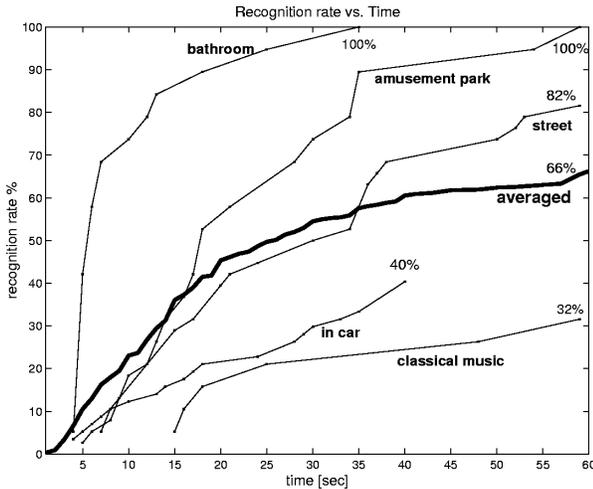
Figure 2. Averaged latencies in the first test

Recognition rate (%)	First test	Second test
Overall	66	78
Seeing subjects	68	78
Blind subjects	62	78
Best	85	90
Worst	47	60

Table 2. Recognition distribution between the subjects

The overall recognition rate for the second test was 78%. In this case, the recognition accuracy for individual environment ranged from 16% to 100%. In the first test, the three best recognized scenes were *bathroom* (100%), *amusement park* (100%), and *presentation* (95%). On the other hand, the three most difficult scenes were "*library*" (0%), *street café* (0%) and *replayed music* (32%). The best recognized scenes for the second test were *kitchen* (100%), *railway station* (100%) and *wind* (water tower, 100%). The worst cases were *library* (16%), *market place* (42%) and *in train* (68%).

Some of the reasons for confusions seem obvious. One of them was misleading sound events. For example, the scene "*market place*", which was recorded in Helsinki at the market of Hakaniemi near the

sea, was frequently recognized as a harbor. In that recording, screams of seagulls can be heard very loudly, which is likely to have misled 26% of the subjects. The scene "*trams*" was a recording of trams passing by at a distance of one meter. However, there were other sound sources as well, like cars and buses. For some subjects the sound event was unfamiliar and they suggested a train as an answer. In the first test, this scene was often confused (see Fig. 1). In the second test, there was less confusion; however, the scene was recognized as a street in 84% of the cases, which can be treated as a correct answer. Another possible source of errors is a lack of prominent sound events. The recording from the library was very quiet with very few loud sound events. The sound events were such as footsteps, beeps of a bar code reader and noises caused by handling of books. Some of the subjects described the scene as "there is nothing going on".

Differences in the recognition abilities between subjects are presented in Table 2. In contrast to our initial assumption that blind subjects would be likely to perform better than seeing subjects, no significant differences can be noticed between the recognition rate of these two groups. However, due to the limited number of subjects, a conclusion cannot be made. In both tests, the best recognition performance was achieved by a young female subject, whereas the worst recognition rate belongs to an older male subject.

Figure 1. Confusion matrix for the first test

Class	Subclass	Examples
Human	many speaking	
	one speaking	
	other human noise	screaming
	understood content	understood speech
	human activity	steps
Vehicles		cars, planes, bikes passing or starting
Continuous noise		car engine, rail noise
Natural	organic	animals
	inorganic	wind, water
Spatial information		reverberation
Prominent event	transients	dishes, coins, crash
	noise	paper bag, machine tools
	replayed content	announcements, radio
	live performance	live music, lecturing

Table 3. Categorization of the cues

Latencies

In Figure 2, average latencies of the recognition process are shown for a couple of scenes in the first test. The bold line in the figure is the latency averaged over all the scenes. On the average, successful recognition took 20 seconds in this test.

The latency curve was calculated as a cumulative sum of the correct answers at a given time. A point on the curve may represent more than one correct answer. This is because the response times were measured at a precision of one second.

A sharp increase in some curves was often caused by a recognized prominent sound event at that particular instant. A good example is the bathroom environment, where a toilet flushing sound is heard about five seconds from the beginning. Actually, none of the subjects recognized the scene before this particular sound event.

The average identification time for the second test was 46 seconds. The purpose of the second test was not to examine the response times. Therefore, the subjects were not hastened to give their answers which, in turn, resulted in longer response times.

Cues

In the second test, the subjects were asked to describe what were the cues they used in the recognition process. It turned out that the cues were most often described in terms of familiar sound sources or events. These answers were analyzed and categorized as shown in Table 3.

In Table 4, the categorized cues reported by subjects are shown. Values in the table are percentages, which indicate how many times each cue was mentioned in a successful recognition. The percentages are averaged over all the scenes and weighted by number of correct identifications. The only difference, which can be noticed between seeing and blind subjects, is that the blind subjects mentioned spatial information more often as a cue than seeing subjects (27% and 12% correspondingly). Some of expressions they used were "this is a big place", "there is a stone flooring" and "it is close to a wall".

On the average, in 47% of the cases transient sounds were reported as a cue (a subclass of "*prominent event*", see Table 3). "*Many speaking*" was reported as a cue in 24% and "*human activity*" in 21% of the cases. In contrast, "*one speaking*" was reported only in 2% and "*understood content*" in 5% of the cases.

In Table 5, the cues used in different scenes are listed. Percentages in the table indicate how many times each category of cues was reported to been used. Only correctly identified scenes were counted.

	All subjects	Seeing	Blind
Human	41	42	40
Vehicles	29	28	31
Cont. noise	9	9	9
Natural	20	21	17
Spatial information	16	**12**	**27**
Prominent event	69	68	72

Table 4. Cues used by the subjects

	Human	Vehicles	Continuous noise	Natural	Spatial information	Prominent event
In train	42		95	5	11	58
Restaurant	55				13	95
Trams	100	100		21	5	26
Supermarket	58				11	95
In car			89	5	5	95
Kitchen	42			58	5	100
Road		100		16		
People indoors	84				58	21
Street	11	89		11	11	37
Market place	79	74		95	5	32
Subway station	16	89	5	5	5	11
Amusement park	89	11			5	74
Nature	68	63	11	79		21
Construction site	11				5	95
Church	11				47	89
Library	53				47	84
Railway station	26	11			47	89
Wind	47	26		95		68
Street café	68	89			16	84

Table 5. Cues used in different environment recognition

For example, 100% means that the cue was mentioned by all subjects that recognized the scene correctly.

CONCLUSIONS

A listening test was described which shows that humans are able to recognize everyday auditory scenes in 70% of cases on average. The confusions were between scenes that had same types of prominent sound events, and were usually not fatal from the point of view of computational auditory scene recognition applications. This can be seen from the confusion matrix in Figure 1.

Recognition latency was of the order 20 seconds. This suggests that an accurate automatic recognizer for similar material should utilize relatively long excerpts of input signals in making the inferences, it is, tens of seconds of audio. In addition, it would seem advantageous to focus the recognition process to distinct sound events. Analysis of the different cues used by humans provides directions in constructing the feature extractors. On the other hand, humans are so much oriented towards analyzing auditory scenes into distinct sound sources that it is not surprising that they described what they heard in term of sources. It is quite probable that amateur listeners are not able to provide low-level qualities of the sound scene, although these may have affected the recognition process.

ACKNOWLEDGMENTS

REFERENCES

[1] Bregman, A.S. (1990). "Auditory Scene Analysis: The Perceptual Organization of Sound". Cambridge, Massachusetts: MIT Press, 1990.

[2] Mellinger, D. K. (1991). "Event Formation and Separation in Musical Sounds". Ph.D. Thesis, Report No. STAN-M-77, Department of Music, Stanford University, CA, 1991.

[3] Ellis, D.P.W. (1996) "Prediction-driven computational auditory scene analysis". Ph.D. Thesis, Massachusetts Institute of Technology, 1996.

[4] Ballas, J.A. (1993) "Common Factors in the Identification of an Assortment of Brief Everyday Sounds" Journal of Experimental Psychology: Human Perception and Performance, 19(2), pp. 250-267.

[5] Gaunard, P.; Mubikangiey, C.G.; Couvreur, C.; Fontaine, V. (1998). "Automatic classification of environmental noise events by hidden Markov models". In proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 6, pp. 3609 – 3612.

[6] Dufournet, D; Jouenne, P.; Rozwadowski, A. (1998), "Automatic Noise Source Recognition". In Proceedings of the 16th International Congress on Acoustics and 135th Meeting Acoustical Society of America (ICA/ASA '98), Seattle, Washington, 1998.
     D., P. and A. presented 20-26 June 1998 - Seattle - ICASSP98

[7] El-Maleh, K.; Samouelian, A.; Kabal, P. (1999). "Frame level noise classification in mobile environments". In Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol.1, pp. 237 – 240.

[8] Saunders, J. (1996) "Real-time discrimination of broadcast speech/music". In Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol.2 pp. 993 – 996.

[9] Scheirer, E. D.; Slaney, M. (1997), "Construction and Evaluation of A Robust Multifeature Speech/Music Discriminator". In Proceedings of the 1997 IEEE Conference on Acoustics, Speech and Signal Processing, vol. 2, pp. 1331 – 1334.

[10] Carey, M.J.; Parris, E. S.; Lloyd-Thomas, H. "A comparison of features for speech, music discrimination". In Proceedings of the 1999 IEEE international Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 149 – 15.