

# On the human ability to discriminate audio ambiances from similar locations of an urban environment

Dani Korpi<sup>1</sup>, Toni Heittola<sup>1</sup>, Timo Partala<sup>2</sup>, Antti Eronen<sup>3</sup>, Annamaria Mesaros<sup>1</sup>, and Tuomas Virtanen<sup>1</sup>

<sup>1</sup>Department of Signal Processing, Tampere University of Technology, *dani.korpi@tut.fi*

<sup>2</sup>Human-Centered Technology, Tampere University of Technology

<sup>3</sup>Nokia Research Center

## Abstract

When developing advanced location-based systems augmented with audio ambiances, it would be cost-effective to use a few representative samples from typical environments for describing a larger number of similar locations. The aim of this experiment was to study the human ability to discriminate audio ambiances recorded in similar locations of the same urban environment. A listening experiment consisting of material from three different environments and nine different locations was carried out with nineteen subjects to study the credibility of audio representations for certain environments which would diminish the need for collecting huge audio databases. The first goal was to study to what degree humans are able to recognize whether the recording has been made in an indicated location or in another similar location, when presented with the name of the place, location on a map, and the associated audio ambiance. The second goal was to study whether the ability to discriminate audio ambiances from different locations is affected by a visual cue, by presenting additional information in form of a photograph of the suggested location. The results indicate that audio ambiances from similar urban areas of the same city differ enough so that it is not acceptable to use a single recording as ambiance to represent different yet similar locations. Including an image was found to

increase the perceived credibility of all the audio samples in representing a certain location. The results suggest that developers of audio-augmented location-based systems should aim at using audio samples recorded on-site for each location in order to achieve a credible impression.

## 1 Introduction

In online location-based services, locations are usually depicted based on visual aspects only. Examples include the online services Nokia Maps 3D [14] and Google Street View [6] which provide 360° street-level views to numerous places on the globe. As it is now, few services exist, which utilize audio ambiances for depicting real locations. However, for example, in computer games and movies the audio is effectively used for conveying more information about the environment to the user. It is intuitively clear that in order to achieve an authentic virtual reality, also audio ambiances must be utilized.

An example of a service which builds heavily on everyday audio ambiances is the Urban Remix project initiated by Georgia Institute of Technology [19]. It is mostly based on artistic aspirations, and all the audio is crowdsourced from users carrying mobile phones. Different recordings are then automatically annotated by the location they have been recorded in, and they represent in a way the soundscape of the location. The recordings can be used, for example, in creating a remix consisting of several audio samples.

---

The final publication is available at [www.springerlink.com](http://link.springer.com/article/10.1007/s00779-012-0625-z) (<http://link.springer.com/article/10.1007/s00779-012-0625-z>)

An alternative for using recordings from the actual environments is to use artificially created soundscapes. Early methods to combine visual and audio descriptions for location suggest using abstract or natural sounds for representing additional information about the location [11, 12], while later methods suggest characteristic audio sequences for mapping environmental noise into city models [17].

From application development perspective and considering the cost of collecting audio ambiance data, it would be attractive to use as few audio recordings as possible for representing everyday audio environments. In [15], the authors analyze the human accuracy in classifying the everyday environment, such as *street*, *car*, or *restaurant*, based on audio ambiance. The authors report that, for 19 subjects, the average correct recognition rate was 70 % for 25 different environments. On the average, the recognition took 20 s for the subjects. Furthermore, the most commonly reported cue based on which the subjects performed the recognition were prominent identifiable sound events.

The results in this earlier work suggest that humans are good at recognizing environments based on audio only, and therefore, it is unlikely that one could successfully use audio samples from a different environment (e.g., *a street*) for depicting a location on a map (corresponding to a restaurant, for example). However, the question arises how accurately humans are able to perceive differences in the audio samples captured in different locations, such as two streets of a city, but which nevertheless belong to the same environment, the *street*. In our current paper, we address this research question.

The perception of audio has been investigated in numerous publications. In [9], a listening experiment was carried out in order to determine how certain discontinuities in synthesized speech are perceived, with the scope of improving existing speech synthesis algorithms. In [16], test subjects had to rate the similarity of several audio samples consisting of different types of sonar echoes to study perceptually significant signal features that should be useful for automatic classification. An investigation of the perceptual structure of everyday sounds was performed in [1] by asking the test subjects to group everyday sounds based on similar auditory characteristics. In [7], it was determined how well humans are able to recognize spoken digits under noisy conditions. This was done in order to im-

prove automatic recognition of speech in noisy environments. These examples indicate that using a listening experiment to extract information regarding the limits of human auditory perception and cognition is well justified in different contexts. Furthermore, this information can be used to develop automated audio signal processing algorithms.

Studies related to cross-modal perception are abundant in various disciplines, such as psychoacoustics, psychology, and computer-human interaction. For the technology field, the most important cross-modal combination is the audio-visual one [3, 18], and the psychological studies reveal that perception of visual quality is increased when presenting visual information together with high-quality auditory information. The influence of the sound is positive when the two perception modes are in synchrony regarding the information that is presented, or when the visual information leaves an ambiguity that can be solved by the sound [5, 21]. Humans are also able to create a visual image based on audio information they receive. The study presented in [10] asked subjects to select a depiction of bars that corresponds in size to the characteristics of different sounds of two wooden or metal bars struck together. The subjects were able to select a depiction of the bars that correctly corresponded to the approximate physical properties of the two bars.

The experiments in [8, 20] are related to the perception of urban environments. In [20], the test subjects were asked to rate the pleasantness of different urban sound environments while presented with pictures of different levels of urbanization. It was observed that the more urban environment the picture depicted, the less pleasant the sound environment was rated, concluding that the visual cues indeed affected the auditory perception of the test subjects. In [8], test subjects were asked to walk a certain route in a city. After that, they were asked to describe prominent sounds and rate the acoustical comfort of the locations along the path. The results of the experiment indicated that the most significant factors affecting the overall impression of the soundscape were odors and acoustic comfort.

The objective of the current study was to investigate whether several locations within one environment could be credibly represented by audio samples from a single location and whether visual cues affect the perceived credibility of the audio sam-

ples. By an environment, we mean a set of similar locations, such as the parks or streets in a city. A location is a particular street or a park in the corresponding environment.

We conducted an experiment studying two central research questions. First, we studied how easily the subjects, who are familiar with the locations, notice whether an audio sample is recorded in the indicated location or another similar location. Second, we studied how an additional visual cue, more specifically a photograph of the location, affects the recognition of the location based on audio ambiances.

## 2 Methods

Nineteen naïve test subjects participated voluntarily in the test. The age span of the subjects was between 20 and 40, and they included 16 males and three females. The test subjects were familiar with the locations under study. Seventeen of them were currently also inhabitants of the same city where the recordings were made. All the subjects reported normal hearing. The subjects participated in the experiment remotely using a web browser, and they were instructed to carry out the experiment with proper equipment, that is, headphones with sufficient audio quality. The exact purpose of the experiment was not revealed to the subjects.

### 2.1 Acoustic material

The material used in the experiment was collected solely for this purpose. In the experiment, we utilized audio samples from three different environments: *street*, *marketplace*, and *park*. Limiting the number of environments to three was necessary in order to keep the duration of the test reasonable. These three environments can be considered to be a representative subset of everyday urban outdoor environments. For each environment category, we selected three distinct physical locations inside a city. The locations used in the experiment, and their corresponding environments are listed in Table 1. A short description of each location is also included.

All of the audio was recorded during a single working day in the late May in the city of Tampere, Finland. Additionally, the recordings from

the different environments were done under as similar conditions as possible. It was made sure that the relative traffic levels on the streets were as similar as possible by doing the recordings outside the busy traffic hours. In parks, the recordings were done in the afternoon after people had got home from work and possibly gone out to relax. The recordings from the marketplaces were done under typical conditions. In some cases, it meant doing the recordings in the morning when there were market stalls present (the marketplace *Tammelantori*) and in some other cases, in the afternoon when there were people walking by (the marketplaces *Laukontori* and *Keskustori*).

One aspect to note was that on the day of the recording the weather was sunny, but the temperature was quite low. Thus, there were not as many people in the parks or in certain marketplaces as on the warmer days. Because the temperatures are usually relatively low in Finland, these weather conditions can be considered very ordinary. On a warmer day, there would have been much more people in those locations, and the auditory scene would have been somewhat noisier. For this reason, the audio samples can be considered to represent their recording locations rather well.

For each location selected to be included in the experiment (see below), three recordings of 3 min in duration were made. The equipment used consisted of an Edirol R-09 portable WAV/mp3 recorder and Soundman OKM II Klassik A3 stereo microphones worn in the ears. The acquired signals were recorded using a sampling rate of 44.1 kHz and saved as stereo audio files using the resolution of 24 bits/sample. The input gain of the recorder was set as high as possible, but it was made sure that there was no clipping in any of the recordings.

In the experiment, we used randomly selected audio samples from the 3-min recordings. The length of the audio samples was chosen to be 10 s, and the selection of test samples was done by randomly selecting a starting point from an audiofile and extracting a 10-s-long audio sample from that position. Two test samples were selected from each recording, making a total of six samples from each location. The audio samples were encoded into stereo MPEG-2 Audio Layer III format at a bitrate of 128 kbit/s in order to decrease the file size and to ensure fluent playback. The compression of the audiofiles was not assumed to have an effect on

Table 1: A short description of all the recording locations. Also the environment that the location represents is indicated

Environment	Location	Description
<i>street</i>	<i>Hämeenkatu</i>	Busy street in the city center, cobblestones
	<i>Teekkarinkatu</i>	One of the main streets in the biggest suburb
	<i>Hatanpään valtatie</i>	Quite a busy street near the city center
<i>marketplace</i>	<i>Laukontori</i>	A marketlace near a harbour
	<i>Tammelantori</i>	A busy marketplace
	<i>Keskustori</i>	Relatively busy square in the city center
<i>park</i>	<i>Sorsapuisto</i>	A park near the city center with a pond and some birdcages
	<i>Kauppi</i>	A natural park with sporting areas
	<i>Pyynikki</i>	A popular natural park with a beach

the results of the experiment as none of the aspects studied was considered to be dependent on having high audio quality. The total amount of audiofiles used in the experiment was 55 (one for every task). Since each task was allocated a different test audio sample, the subjects were not able to learn the samples.

## 2.2 Tasks

The experiment consisted of 54 tasks for the subjects, plus one familiarization task. Each task comprised an audio sample, a statement, a rating scale, a miniature map of the location, and in half of the tasks a picture of the location. All the material for a task was presented on a single page. On each task page, the subjects rated a statement that was of the form: *"This audio sample sounds like it has been recorded at X"*. The letter *X* denotes a physical location (e.g., a certain street in the city of Tampere).

In every task, there was a miniature map indicating the location on a map. The exact location the recording had (allegedly) been made at was marked on the miniature map with a red cross. The approximate time the recording had been made was also presented textually (*morning, noon, afternoon* or *evening*). The subjects were asked to give their ratings on a nine point Likert scale ranging from 1 (disagree) to 9 (agree). The value of one on the scale corresponded to complete disagreement while the value nine corresponded to complete agreement with the statement. In order to select a certain rating, the subject had to click a ticker corresponding to the value he had chosen. There was also a text

field where the subject could write an explanation for the selected rating, but filling the field was not mandatory. The exact layout of a single task page can be seen in Fig. 1.

The effect of visual cues was measured by presenting every task with and without a picture. The list of tasks for each location is presented in Table 2. After the actual experiment tasks, the subjects were also asked to rate how familiar they were with the nine locations used in the experiment. This was done in order to study whether the familiarity with a location affected the ratings. The subjects were shown all the information regarding the location in question (map and picture) and they were asked: *How well do you know X?* The ratings were again given on a nine point Likert scale with the value one corresponding to *Not at all* and the value nine corresponding to *Very well*. The locations were presented in random order. The location familiarity questions were at the end of the experiment in order not to give any information regarding the locations before the actual experiment.

## 2.3 Procedure


The experiment was implemented as a dynamic web page in order to enable flexible execution. This made it possible for the subjects to do the experiment whenever they wished. This also guaranteed that the listening environment corresponded as well as possible to the use environment of the possible final applications (i.e., a user in the proximity of his personal computer) [2]. The subjects were encouraged to use headphones and to choose the rating according to their opinion, not based on whether

## Location Recognizability Test

Test progress:

Read the statement and look at the picture(s). There is always a map visible which points the location of the place that the statement is about. There might also be a picture of the place. The approximate time the audiosample has been recorded is given as well. After having seen all the information, you can listen to the audiosample by clicking the play-button. After that choose your answer (1-9) to the statement. You can also write a verbal reason for your rating in the text field (mention why it sounded or did not sound right). It is not necessary to finish the test at once because you can continue the test by going to the start page again and writing your email-address to the login field.

This **audiosample** sounds like it has been recorded at **Keskustori**.

 **Play sample**

disagree          agree  
1 2 3 4 5 6 7 8 9

Here you can write an explanation for your rating (not obligatory):

A verbal answer is written here.

Next question →

### Extra information

Recording time: Afternoon



Picture



Map

Figure 1: Screen capture of a task page

Table 2: The structure of the tasks regarding one location (marked by  $X$ )

Task #	Picture from the location "X"	Audio sample used in the task
1	Yes	From location "X"
2	Yes	From the second location of the environment
3	Yes	From the third location of the environment
4	No	From location "X"
5	No	From the second location of the environment
6	No	From the third location of the environment

they felt it was the right answer or not. The latter was done in order to ensure that the subjects concentrated on the features of the audio sample instead of guessing the location.

The different tasks regarding one location are constructed in the way presented in Table 2. All the possible combinations were covered within each environment: Every audio sample was in turn claimed to have been recorded at each of the three locations, and all the tasks were performed both with and without a picture. There were in total nine locations and six tasks per location; therefore the total amount of tasks was  $9 \times 6 = 54$ . The order of the tasks was fully randomized separately for every subject. Thus, the subjects' possible decrease in concentration toward the end of the experiment should not have affected the results.

The experiment was divided into three main stages: the familiarization stage, the actual experiment, and the location familiarity questionnaire. The familiarization stage consisted of one example task. The subject was told to closely investigate all the information in the familiarization task so that he would be able to do the actual experiment in a proper way. The audio sample used in the familiar-

ization task was correct, meaning that it was from the location mentioned in the statement. During the familiarization task, the subject was textually instructed on all the necessary components used in the experiment. These included the scale to be used (1–9) and the text field for qualitative comments. The subject was also told about the picture and the map that would be visible in the task pages.

After the familiarization question, the actual experiment began and tasks were presented on the screen one by one, together with an audio sample. In half of the tasks, a picture of the location was also shown on the screen. A progress bar indicating how much of the experiment had been completed was also visible. For every statement, there was a numerical rating scale using which the desired rating was given by selecting an item. The subject could choose his rating at any time and possibly write an explanation for it in the text field. In order to listen to the audio sample, the subject had to click a play button. Therefore, it was ensured that the subjects had enough time to investigate the statement and all the other information available before concentrating on the audio. It was allowed to listen to the audio sample several times. This ensured that possible random distractions could be compensated by merely listening the audio sample again. Whenever the subject was ready, he could click a button to move on to the next task.

## 2.4 Data analysis

As a result of the experiment, we obtained a data set where all possible location combinations inside each environment obtained a rating on the nine point scale indicating the degree with which the subjects considered that the recording was made in the suggested location. There were two variables in the data set whose effect was studied: the location in which the audio sample was recorded (correct/similar), and whether a picture was presented along with the task or not.

The first null hypothesis ( $H_0$ ) was as follows:

*The subjects are not able to recognize whether the audio sample is from the correct location or not in the test conditions used.* (1)

If the null hypothesis is true, it would mean that there was no significant difference in the rat-

ings regarding whether the audio sample was really recorded in the suggested location compared with the ratings where the audio sample was in fact recorded in a different location. The second null hypothesis ( $H_0$ ) was as follows:

$$\begin{aligned} & \textit{The inclusion of a picture describing the} \\ & \textit{location has no effect on the subject's} \\ & \textit{ratings in the test conditions used.} \end{aligned} \quad (2)$$

The data were first analyzed with Matlab using two-way repeated measures analysis of variance (ANOVA). The two variables were within-subjects variables, and they formed a two by two factorial design (two variables with two levels: correct/similar audio sample, picture/no picture).

### 3 Results

The test subjects rated all the tasks on a numerical scale. These ratings constitute the main data for the analysis. The subjects were also asked to write freeform comments in a text field whenever they wished to do so. These qualitative comments give some insight into the reasons behind certain trends in the ratings. The results also revealed that the subjects' familiarity with the locations did not affect their ratings.

#### 3.1 Quantitative results

The mean ratings are presented in Table 3, grouped by the state of the variables. It is not possible to draw conclusions based on mean ratings only, but they provide some further insight into the results of the statistical analysis. It can, for example, be observed that the audio samples from the correct location were rated the highest and that showing a picture also seemed to increase the ratings.

The mean ratings of the location familiarity questions are presented in Table 4. Most of the subjects were familiar enough with all the locations, and hence, a full analysis about the effect of the familiarity was not viable. However, based on a visual inspection, it seems that familiarity with a location did not have a significant effect on the ratings. This can be seen from Fig. 2. The graph depicts the average rating with respect to the subjects' familiarity with a corresponding location. It can be

Table 3: Mean values of the recognition ratings for the environments, grouped by the state of the variables

Environment	Audio sample		Picture	
	Correct location	Similar location	Yes	No
<i>all</i>	7.3	5.7	6.5	5.9
<i>street</i>	7.1	5.1	6.1	5.5
<i>marketplace</i>	7.0	5.8	6.7	5.9
<i>park</i>	7.5	6.2	6.9	6.4

Table 4: Mean values of the subjects' familiarity ratings with the locations used in the experiment

Environment	Location	Mean value of familiarity
<i>street</i>	<i>Hämeenkatu</i>	8.3
	<i>Teekkarinkatu</i>	6.6
	<i>Hatanpään valtatie</i>	7.4
<i>marketplace</i>	<i>Laukontori</i>	6.2
	<i>Tammelantori</i>	5.4
	<i>Keskustori</i>	7.9
<i>park</i>	<i>Sorsapuisto</i>	5.6
	<i>Kauppi</i>	5.0
	<i>Pyynikki</i>	6.8

seen that the graph is nearly flat, indicating that when all the data are taken into account, the subjects' familiarity with a location does not correlate with their ratings. The fluctuations of the average ratings occurring at the lower familiarity values are most likely due to small amount of data (76 % of the familiarity ratings were 6 or higher). Investigating the effect of the location familiarity more closely is, however, beyond the scope of this paper.

Finally, the results of the ANOVA are presented in Table 5. The  $F$  values and corresponding significance levels are shown for the two investigated variables alongside with their interaction effect (the correlation between the effects of the two variables). These results confirm the observations made based on the mean ratings. From Table 5 it can be seen that all the aforementioned differences in the ratings were indeed statistically significant. Hence, it can be concluded that the subjects perceived the

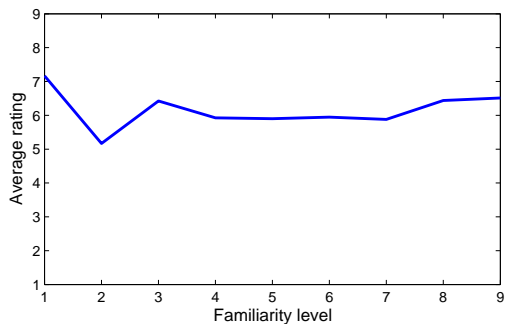


Figure 2: The effect of the familiarity to the subjects’ ratings. The *horizontal axis* represents the subjects’ familiarity with a location, and the *vertical axis* represents the average of the ratings corresponding to a certain level of familiarity

audio sample from the correct location the most credible and that a picture increased the credibility of an audio sample.

### 3.2 Qualitative comments overview

The subjects’ estimation about the real recording location of the audio sample was in most cases quite close to reality. Even though the locations were chosen so that they were from similar urban environments, each of them still had at least a vague distinct feature. Furthermore, it seems that the subjects were able to detect that feature and they could usually differentiate between the locations.

#### Street environment

According to the results, the most easily recognizable locations were in the *street* environment. There were two busy streets from the city center (*Hämeenkatu* and *Hatanpään valtatie*), and one of the main streets from the biggest suburb (*Teekkarinkatu*). It seems that the test subjects usually had quite firm knowledge regarding where the audio sample could have been recorded and where it could not have been.

From the qualitative comments, it was observed that the knowledge of there being cobblestones in *Hämeenkatu* was quite significant in terms of recognizing audio samples recorded there. If there were not any cobblestones, then it was not

*Hämeenkatu* and vice versa. When a recording from *Hämeenkatu* was claimed to represent other streets, the cobblestones revealed in most cases that the recording was incorrect. During the recordings, there was also a street musician in *Hämeenkatu* and many people rightly deduced based on the music that those recordings were indeed from *Hämeenkatu*.

*Teekkarinkatu* was recognized in most cases due to the knowledge of there being somewhat less traffic compared with the two streets from the city center. The incorrect recordings were usually slightly too busy to represent it in a credible way. Due to the same reason, the recordings from *Teekkarinkatu* were also too quiet to represent the other two streets.

#### Marketplace environment

In this environment, the locations were not as clearly distinguishable, but there were still some auditory features that could be used to recognize the recording location. The subjects knew that there should be a lot of babble noise in *Tammelantori* because of the marketstands so it was possible to use this feature in recognizing recordings from there.

*Keskustori* was in most cases recognized by the sound of the buses idling nearby. There also happened to be some birds audible in the recordings from *Tammelantori* and *Laukontori* and some subjects reckoned there would not be any birds in *Keskustori*. This helped these test subjects to correctly determine that those audio samples were not from *Keskustori*.

In the case of *Laukontori* the qualitative comments did not reveal any general trend regarding the features leading to correct recognition. Based on the ratings, *Laukontori* seems to have been rather challenging to recognize because in the tasks where a picture was included, audio samples from both *Laukontori* and *Keskustori* were rated almost equally credible.

#### Park environment

The locations in the *park* environment were also very close to each other in terms of auditory scenes, and there were not many distinct features. The sounds of football from *Kauppi* were recognized by



Table 5: The  $F$  values and significance levels obtained from the analysis of variance on the experiment data. The degree of freedom was 1 in all the cases

Environment	Correct / similar audio sample		Picture / no picture		Interaction	
	$F$ value	Significance level	$F$ value	Significance level	$F$ value	Significance level
<i>all</i>	122.39	$p < 0.001$	18.83	$p < 0.001$	0.57	$p > 0.1$
<i>street</i>	39.26	$p < 0.001$	5.26	$p < 0.05$	0.70	$p > 0.1$
<i>marketplace</i>	42.07	$p < 0.001$	18.97	$p < 0.001$	2.67	$p > 0.1$
<i>park</i>	54.96	$p < 0.001$	10.36	$p < 0.01$	0.78	$p > 0.1$

many subjects, but usually, it was not enough to make them certain about the location of the recording.

Also, the sounds of birds in the recordings from *Sorsapuisto* were of some help in the recognition as they are a characteristic feature of this park. Sounds of beach volley were also rightly connected to *Pyynikki*.

However, there were not any overwhelming trends in the recognition of any of the locations. In some cases, certain sounds were even attributed to the wrong location. For example, some subjects assumed that the sound of tennis belonged to *Kauppi* even though the audio was recorded in *Pyynikki*.

## 4 Discussion

According to the current results, the subjects were able to discriminate between the cases where the sample was from the presented location and the cases in which the sample was from a different but similar location. This was evidenced by the finding that tasks with audio samples recorded at the suggested location received significantly higher ratings than tasks with audio samples from the other similar locations. Thus, the first null hypothesis (1) can be rejected.

Actually, there were only 5 cases out of 18 where a similar audio sample got higher ratings than the correct one. The only location which could not be recognized at all was the *Sorsapuisto* park: When the suggested location was *Sorsapuisto*, audio samples from the park of *Pyynikki* were rated the most credible to represent the location both with and without a picture. This might be due to the fact that the distant car sounds audible in the recordings from *Pyynikki* were attributed as sounds of a

park in the center of the city according to the voluntary qualitative comments. It can also be observed from the qualitative comments that the subjects assumed the sounds of children playing to belong somewhere else than *Sorsapuisto* even though this was not the case.

It can be expected that the weather, season, time of the day, weekday, or some other factor apart from the type of the location may also have an effect on the credibility of the environment audio ambiance, but these aspects are left outside the current study. The audio ambiances in the chosen locations also include some time-varying properties (e.g., the market places are more busy during market hours), which may make them more easy to recognize at certain time. The reason for doing the recordings on the same day was to exclude the effect of most of these factors, and thus, ensure that they did not significantly affect the main results of the experiment.

When drawing implications based on the current results, the role of subject selection has to be also discussed. In this experiment, the subjects were students or professionals of 20–40 years of age. By selecting these subjects, we aimed at conducting the test with subjects, who possess normal cognitive and perceptual abilities. The recognition rates might have been lower, for example, for older subjects with decreased long-term memory capacities. In addition, gender distribution could not be balanced in this study for practical reasons. There is evidence for different kinds of minor gender effects on a number of cognitive and perceptual aspects, for example, visual perception and spatial memory [4]. Overall, however, the current results were quite clear, and a different gender or age distribution would not probably have changed the main

results. The impact of those factors on location recognition based on audio ambiances is thus left to be explored in subsequent studies.

Nevertheless, according to the results, the effect of the picture is also very clear. When the analysis was done for all the data, the effect of the picture was statistically significant, with  $p < 0.001$ . Together with the findings that the interaction effects were not significant, this suggests that there was a statistically significant difference in the ratings independent of whether the audio sample was from the suggested location or not. From Table 3 it can be observed that a picture increased the grades in all environments. We can thus reject our second null hypothesis (2) and conclude that a picture has an effect in all the environments.

When looking at the  $p$  values for interaction effects from Table 5 it can be observed that none of them is below the significance level. Thus, it can be said that there was no significant interaction between the two variables. This means that the two variables had an effect regardless of what the state of the other variable was. Hence, the subjects were able to discriminate between those audio samples that were from the suggested location and those who were only from a similar location in an equal manner regardless of whether a picture was shown or not. Furthermore, as was already observed from Table 3, when a picture was provided the ratings given by the subjects significantly increased regardless of the audio sample.

The important observation is that when shown a picture, also the audio samples recorded in another similar location became more credible to the subjects. Although a picture did not affect the ability to discriminate between audio samples from different locations within one environment, it was able to increase the credibility of the audio samples recorded from another similar location as well as that of the audio samples recorded from the suggested location.

These results indicate that the visual cues actually have an effect on perception of audio from everyday environments. Similarly, as the perception of visual quality was increased when presented auditory information [3], the perceived credibility of audio seems to increase when presented additional visual information. However, unlike in [5, 21] where it was observed that auditory stimulus allowed the test subjects to perceive visual information more

precisely, here the effect was somewhat different. Now the presence of visual stimulus seemed to decrease the test subjects' attention toward the auditory stimulus, hence increasing the perceived credibility of all audio samples. An interesting research topic would be to further investigate the relationship between humans' visual and auditory perception. It is clear that there are still several factors whose effects are yet to be understood.

The results obtained from our experiment were remarkably explicit and consistent. Using native test subjects gives these results significant practical value since the results are applicable to the general public. The findings of this listening experiment are hence just as essential in terms of practical applications as in terms of research on auditory perception.

## 5 Conclusions

In our test with three everyday environments and three different locations per environment, the test subjects were able to recognize whether an audio sample was recorded at the suggested location or not. To the best of our knowledge, this is the first time this has been experimentally shown. This result indicates that the audio ambiances used for real locations need to be either from the exact same location or a location which sounds very similar, possibly a nearby location. Our results show that recordings from different streets or parks of the same city are not likely to provide credible enough audio ambiance. Hence, in order to represent a certain set of locations in terms of audio, one must have audio material from each location. This is in some sense an unpleasant result as making audio recordings in all locations obviously requires more resources than just making them at selected locations.

However, a more encouraging observation was that the inclusion of a picture depicting the suggested location increases the ratings of an audio sample regardless of whether the audio sample has actually been recorded in the suggested location or not. This indicates that a visual cue increases the credibility of an audio sample in representing a certain location. In the previous literature, this result has been a subject of hypothesizing [13]. From application developer's perspective, this re-

sult means that more careful selection of audio samples is needed when no visual information is available, whereas the audio samples can be somewhat less representative of the actual location when the users can also rely on visual information.

More research is needed to study what are the cues which are important when recognizing a certain location and which are the situations where audio ambiences from different locations could be credibly used to represent a location. This would likely require conducting a listening experiment with much more similar locations. In this way, it might be possible to determine the threshold at which the soundscapes of two locations are too similar for the subjects to be able to reliably distinguish them. This information — together with the findings of the current experiment — would be very valuable for the developers of future audio-augmented location-based systems. The results could be used as guidelines for creating credible audio impressions while keeping the number of different audio ambiences required as small as possible.

## Acknowledgements

Funding from the Live Mixed Reality 2011 project by the Finnish Funding Agency for Technology and Innovation (TEKES) and Nokia Research Center is gratefully acknowledged. The authors would also like to thank all the test subjects who participated in the listening experiment.

## References

- [1] Terri L. Bonebright. Perceptual structure of everyday sounds: A multidimensional scaling approach. In *Proceedings of the 7th International Conference on Auditory Display*, pages 73–78. ICAD, Laboratory of Acoustics and Audio Signal Processing and the Telecommunications Software and Multimedia Laboratory, Helsinki University, 1998.
- [2] Terri L. Bonebright, Nadine E. Miner, Timothy E. Goldsmith, and Thomas P. Caudell. Data collection and analysis techniques for evaluating the perceptual qualities of auditory stimuli. In *Proceedings of the ICAD '98*. ICAD, British Computer Society, 1998.
- [3] David Burr and David Alais. Combining visual and auditory information. *Progress in Brain Research*, 155:243–258, 2006.
- [4] J.C. Chrisler and D.M. McCreary. *Handbook of gender research in Psychology: Vol. 1. Gender research in general and experimental psychology*. Springer, New York, 2010.
- [5] Francesca Frassinetti, Nadia Bolognini, and Elisabetta Ladavas. Enhancement of visual perception by crossmodal visuo-auditory interaction. *Experimental Brain Research*, 147(3):332–343, 2002.
- [6] Google. Street view. Online. Accessed 2011-10-10.
- [7] P. D. Green, M. P. Cooke, and M. D. Crawford. Auditory scene analysis and hmm recognition of speech in noise. In *Proceedings of the ICASSP '95*, pages 401–404, 1995.
- [8] J. Y. Hong, P. J. Lee, and J. Y. Jeon. Evaluation of urban soundscape using soundwalking. In *Proceedings of 20th International Congress on Acoustics*, Sydney, 2010.
- [9] E. Klabbers and R. Veldhuis. Reducing audible spectral discontinuities. *IEEE Transactions on Speech and Audio Processing*, 9(1):39–51, 2001.
- [10] Stephen Lakatos, Stephen McAdams, and Ren Causs. The representation of auditory source characteristics: Simple geometric form. *Perception And Psychophysics*, 59(8):1180–1190, 1997.
- [11] Alan M. MacEachren and David Ruxton Fraser Taylor. *Visualization in Modern Cartography*. Pergamon, New York, 1994.
- [12] Ryan MacVeigh and R. Daniel Jacobson. Increasing the dimensionality of a geographic information system (gis) using auditory display. In Gary P. Scavone, editor, *Proceedings of the 13th International Conference on Auditory Display (ICAD2007)*, pages 530–535, Montreal, Canada, 2007. Schulich School of Music, McGill University.

- [13] Nadine E. Miner. *Creating wavelet-based models for real-time synthesis of perceptually convincing environmental sounds*. PhD thesis, University of New Mexico, 1998.
- [14] Nokia. Maps 3d. Online. Accessed 2011-10-10.
- [15] Vesa T. K. Peltonen, Antti J. Eronen, Mikko P. Parviainen, and Anssi P. Klapuri. Recognition of everyday auditory scenes: Potentials, latencies and cues. In *Proceedings of the 110th Audio Engineering Society Convention*, Amsterdam, 2001. Hall.
- [16] S. Philips, J. Pitton, and L. Atlas. Perceptual feature identification for active sonar echoes. In *Proceedings of the IEEE OCEANS Conference*, pages 1–6, 2006.
- [17] Jochen Schiewe and Anna-Lena Kornfeld. Framework and potential implementations of urban sound cartography. In *Proceedings of the 12th AGILE International Conference on Geographic Information Science*, Hannover, 2009.
- [18] Russell L. Storms. *Auditory-Visual Cross-Modal Perception Phenomena*. PhD thesis, Naval Postgraduate School, Monterey, California, 1998.
- [19] Urban Remix project. Online. Accessed 2011-10-10.
- [20] Stphanie Viollon, Catherine Lavandier, and Carolyn Drake. Influence of visual setting on sound ratings in an urban environment. *Applied Acoustics*, 63(5):493–511, 2002.
- [21] Jean Vroomen and Beatrice De Gelder. Sound enhances visual perception: Cross-modal effects of auditory organization on vision. *Journal of Experimental Psychology: Human Perception and Performance*, 26(5):1583–1590, 2000.