

SUBJECTIVE EVALUATION OF AURALIZATION OF PHYSICS-BASED ROOM ACOUSTICS MODELING

Tapio Lokki

Helsinki University of Technology
Telecommunications Software
and Multimedia Laboratory
P.O.Box 5400, FIN-02015 HUT, Finland
Tapio.Lokki@hut.fi

Hanna Järveläinen

Helsinki University of Technology
Laboratory of Acoustics
and Audio Signal Processing
P.O.Box 3000, FIN-02015 HUT, Finland
Hanna.Jarvelainen@hut.fi

ABSTRACT

This paper describes the results of subjective evaluation of auralization by listening tests. The task was to compare real-head recorded and auralized sound samples. The evaluation process as well as the creation of soundtracks are briefly reviewed. The listening test procedure is presented along with the analyzed results of a case study. They show that with a simple room geometry (a lecture room) reliable and natural sounding auralization is possible with physics-based room acoustic modeling. However, there are still some modeling problems which are discussed as well as the guidelines for future work in evaluation.

1. INTRODUCTION

Room acoustics modeling and auralization, especially in real-time systems, are divided into two different approaches, namely perceptual and physical modeling. In perceptual modeling the aim is to find a set of perceptually relevant parameters by which room acoustics can be rendered. These parameters include source presence, envelopment, room presence, late reverberance, etc. [1]. By contrast in physical modeling the behavior of sound waves in a modeled space is simulated and sound rendering is done using parameters obtained from physics-based modeling. The most popular room acoustic simulation method used with auralization is image-source method [2, 3]. In this paper the scope is in the auralization of physics-based room acoustics modeling and especially in the subjective evaluation of auralization.

The paper is organized as follows. First, we discuss briefly the evaluation process including real-head recordings and auralization methods. Then we describe applied listening test methods and present the results of two listening tests. Finally, the problems in evaluation of dynamic auralizations are discussed and possible future work guidelines are suggested.

2. THE EVALUATION PROCESS

The evaluation of naturalness of auralization is done by comparing real-head recordings and auralized room acoustics simulation results (see Fig. 1). The real-head recordings were used as reference signals.

2.1. Real-Head Recordings

The reference soundtracks were prepared by playing and then recording the anechoic sound samples in a room using the real-head recording technique [4, 5] (see, e.g., [6] more about the fundamentals of binaural recording technique). The studied space was a lecture room (dimensions 12 m x 7.3 m x 3 m) presented in Fig. 2. The anechoic stimuli were played with a CD player and reproduced with a small active loudspeaker (Genelec 1029A). Small electret microphones were placed at the entrances of the ear canals and connected to a DAT recorder. The contents of the DAT tape were then transmitted to the computer, edited and equalized for headphone listening, if required (see leftmost column in Fig. 1).

2.2. Auralization of Room Acoustics Simulation Results

In room acoustics simulation the goal was to create a totally artificial virtual auditory environment, in other words no measured impulse responses were used. This means that the sound source characteristics, sound propagation in a room as well as listener modeling were done by digital signal processing. The DIVA auralization system [7], which enables both static and dynamic auralization, was used. A detailed description of the system is presented in [8].

In DIVA auralization the modeling of room acoustics is divided into three parts, the modeling of direct sound, early reflections, and late reverberation. The direct sound and early reflections are modeled with image source method, in this case up to the third order reflections. In the studied space (Fig. 2) this means 30-50 reflections which are coming to the listener in the time window of 50 ms after direct sound. With image source method the following parameters for each reflection, at each time moment (update rate was 100 Hz), are calculated:

- reflection number
- distance from listener
- azimuth and elevation angle in respect of listener
- orientation (azimuth and elevation angles) of sound source
- material filter parameters (from which materials each reflection is reflected)

These parameters are used in auralization process which is realized with a signal processing structure presented in Fig. 3. The signal processing blocks in Fig. 3 contain the following filters:

- $S_d(z)$ is a diffuse field filter of sound source directivity.

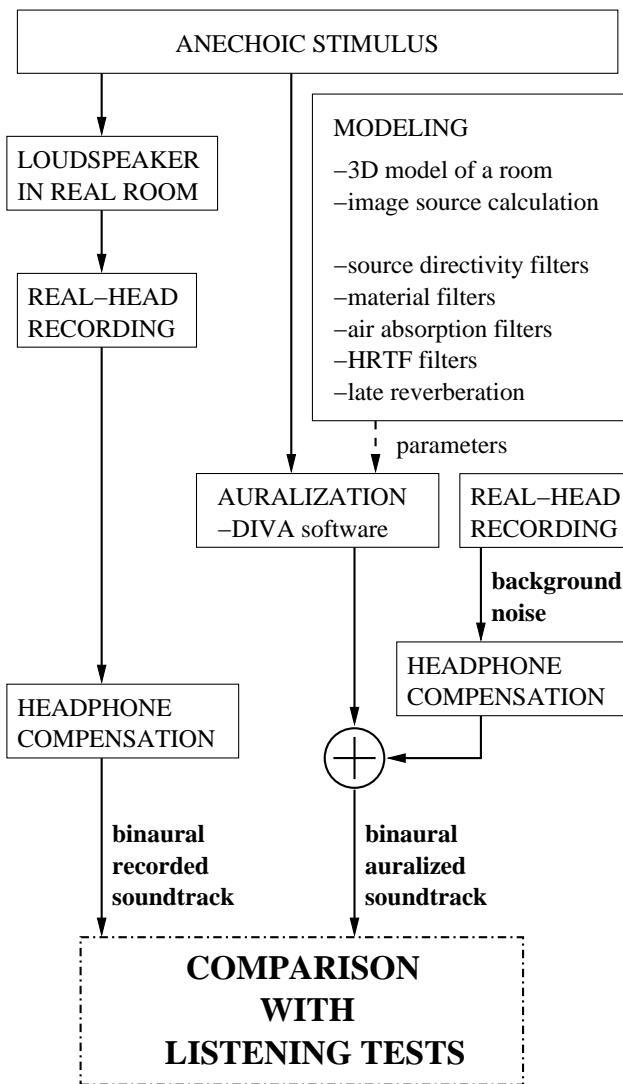


Figure 1: The creation procedure of the soundtracks. No separate headphone compensation for auralized sounds are done because the compensation is embedded in HRTF filters.

- $F_d(z)$ is a diffuse field filter of HRTF filters.
- $T_{0..N}(z)$ contain the sound source directivity filter, distance dependent gain, air absorption filter and material filter (not for direct sound).
- $F_{0..N}(z)$ contain directional filtering realized with separated ITD and minimum-phase filters for HRTF.
- Late reverberation unit is a parameterized late reverberation algorithm [9].

The signal processing system contains dozens of digital filters and other parameters which can be varied. Because the goal was to create as natural sounding virtual auditory environment as possible, we had to compromise computational efficiency and optimal filter orders.

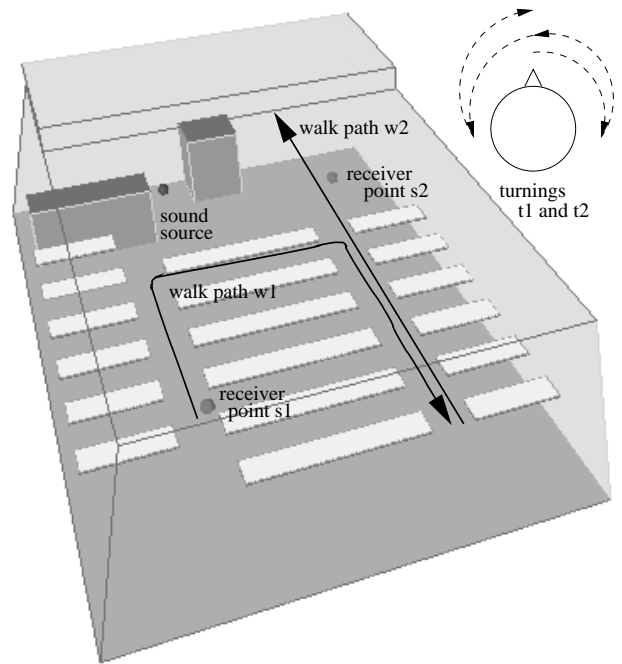


Figure 2: The 3D model of the studied lecture room. In addition, different rendering cases are depicted. Head turnings ($t1$ and $t2$) are applied in static receiver points $s1$ and $s2$.

2.3. Discussion About the Evaluation Process

All attributes which did not depend on auralization were held constant in both cases, if possible. To achieve this, the same stimulus was used in each case. In addition, errors in sound source and listener modeling were kept minimal. The radiation characteristics of the sound source were measured and radiation filters were designed to fit to the measured frequency responses. The applied HRTFs were measured from the same person who did the real-head recordings. Real-head recordings without any stimulus were also made to capture the background noise which was then added to the auralized soundtracks. Both soundtracks were compensated for the frequency response of the headphones as presented in Fig. 1.

3. LISTENING TESTS

To find out how realistic our auralization sounded, a listening test was conducted. Both recorded and auralized soundtracks, with durations from 10 to 20 seconds, were played to the listener who could switch between them. The test was an AB comparison test with four different stimuli and six different cases. The stimuli were *clarinet* (cla), *guitar* (gui), *continuous drumming with snare drum* (dru), and *female singing* (voi). A short sonogram of each stimulus is presented in Fig. 4. It can be seen that drum was transient-like wideband signal while clarinet and singing are tonal signal containing no transients. The rendering cases, as illustrated in Fig. 2, were *two static ones* ($s1$ and $s2$), *two turnings* ($t1$ and $t2$), and *two walk-paths* ($w1$ and $w2$). A turning means that in both static listening points the head is turned to both sides. With four stimuli and six cases, the total amount of pairs to compare was 24.

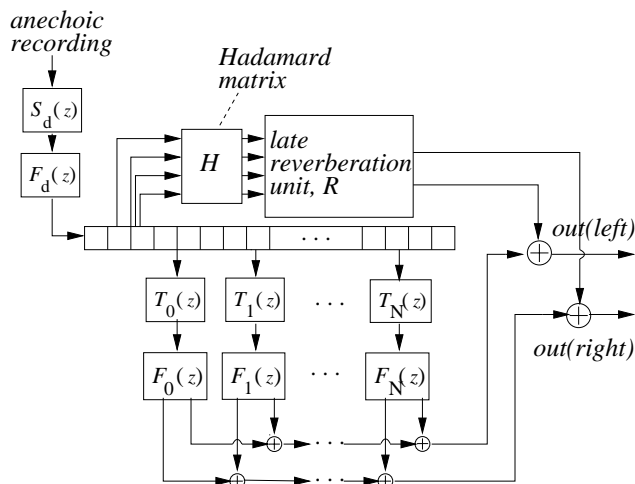


Figure 3: The DIVA auralization signal processing structure.

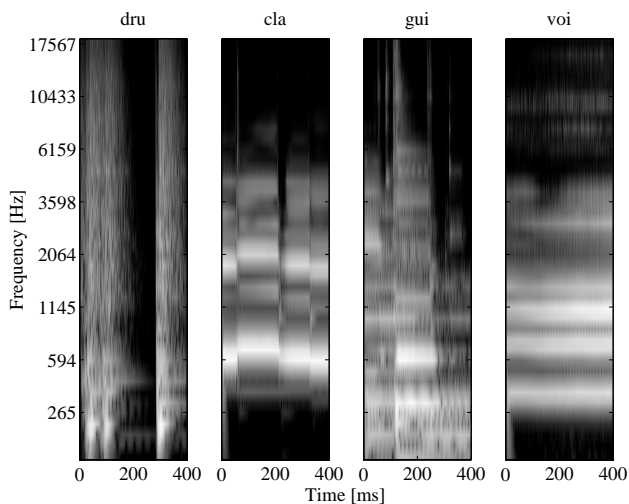


Figure 4: A short sonogram of each anechoic stimulus.

3.1. Method

The test was conducted using the GuineaPig2 system [10]. The graphical user interface is depicted in Fig. 5.

The whole test set contained 32 pairs, played in random order. In addition to the 24 test pairs, eight pseudo pairs were used where both samples were exactly the same. This way the answers were checked for reliability and a possible bias. All subjects were trained with four extra pairs which were listened before the test under surveillance of the test supervisor. This way it was confirmed that everyone understood the tasks.

3.2. Tested variables

The listeners were asked to rate the samples according to *sound source localization*, *externalization*, *sense of space* and *timbre*. The answering scale was from very different to very similar (see Fig. 5). Each answer corresponded to an integer from 1 to 5, score 1 being for “very different”.

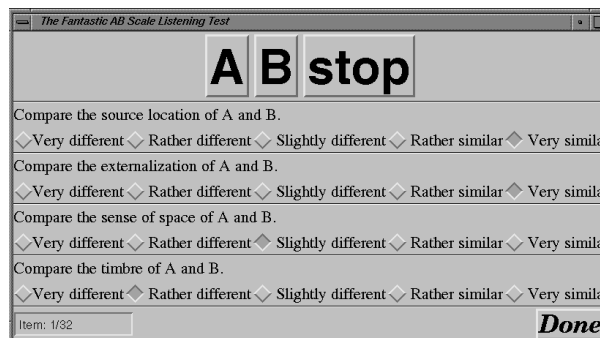


Figure 5: The graphical user interface used in listening tests.

3.3. Test 1

Twelve subjects participated in the first listening test. All of them reported normal hearing. The subjects were not aware how the soundtracks were created and what was the position of sound source. They were only aware that the task was to compare two soundtracks and answer to all questions.

3.4. Results of test 1

The initial analysis considers the reliability of the subjects and the test. In each test set, there were eight pseudo pairs (four recorded and four auralized) in which the sound samples were exactly the same. All subjects found these pseudo pairs equal (median was “very similar”), only a few outliers “slightly different” or “rather similar” were found. This shows that the subjects were reliable.

An analysis of variance (ANOVA) model was employed for the analysis of ratings with each question. Figure 6 shows the results with each stimulus. The overall result was promising. All stimuli except the drum gave “rather similar” or “slightly different” result as a median value to all questions (see Fig. 6). The stimulus drum was not rated as well as the others. This was expected because drum sounds, being very wide-band transient signals, give no excuses with modeling errors. Especially, *timbre* and *sense of space* were clearly judged “rather different”. In all questions ANOVA showed that the difference between drum and other stimuli was significant ($p \ll .01$). Also in *externalization* ($p = .004$) and *sense of space* ($p = .004$) clarinet was rated significantly better than other stimuli. *Timbre* ratings for guitar were significantly lower ($p \ll .01$) than for clarinet and voice.

In addition to the analysis of each stimulus the different cases were analyzed, see Fig. 7. In *sense of space* and *timbre* no significant differences were found between cases. Actually, the subjects had answered to both questions that the recorded and auralized soundtracks were “slightly different”, although a few “very different” and “very similar” judgments were found. In *source location* the ratings for s2 were significantly worse than the others on the $\alpha = 0.05$ level ($p = .015$). Also in *externalization* case s1 was judged significantly ($p = .007$) worse than the others. These two last findings are in line with Wenzel’s statement [11] that the dynamic auralization (when head movements are included) is considered more reliable in source location and externalization.

Finally, we applied two-way ANOVA to find possible interactions between rendering cases and stimuli. No such interaction was found which means that all stimuli gave equal results in different rendering cases.

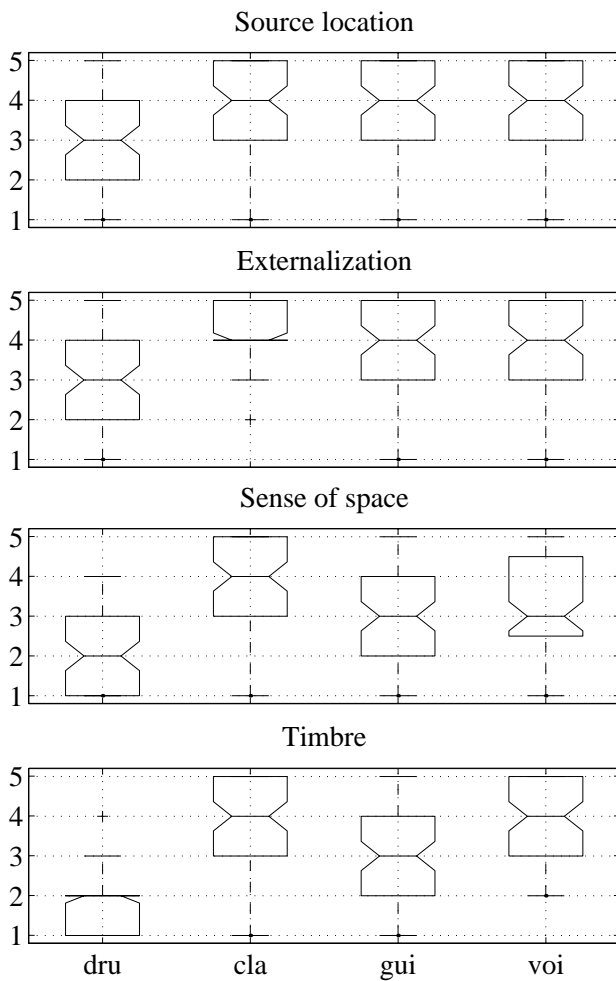


Figure 6: The results of the first listening test with each stimulus. The boxplot depicts the median and the 25%/75% percentiles. On vertical axis numbers are: 1=very different, 2=rather different, 3=slightly different, 4=rather similar, and 5=very similar.

3.5. Test 2

In test 1 the subjects did not know beforehand how the soundtracks were created. The general comment after the test 1 was that all pairs sounded very similar, except the drumming in which there was a clearly audible difference in timbre (or sense of space). In addition, some of them had problems to realize if the sound source was moving or if they were moving themselves. Some subjects also reported that *source location* was hard to define, because both soundtracks were located inside the head.

The comments of Pellegrini [12] as well as the verbal feedback from subjects forced us to make a second listening test with the same soundtracks. This time, before the test, we carefully explained to the subjects the idea of the test and we also showed the rendered videos of the walk-paths. Thus, subjects had knowledge of the modeled space and the position of sound source as well as how they were moving in this modeled space. Finally, the second test was accomplished by six subjects who also participated to the first test earlier.

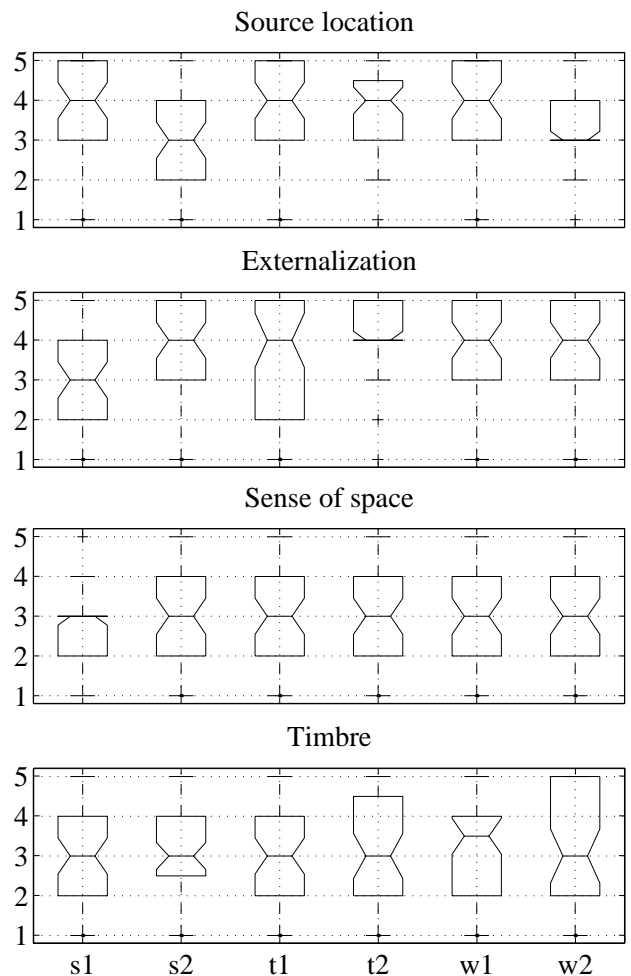


Figure 7: The results of the first listening test with each case (see Fig. 2). The boxplot depicts the median and the 25%/75% percentiles. On vertical axis numbers are: 1=very different, 2=rather different, 3=slightly different, 4=rather similar, and 5=very similar.

3.6. Results of test 2

Figure 8 shows the results of the second test with each stimulus. The plot compares the results of tests 1 and 2 for six subjects. It can be seen that the knowledge of sound source location, size of space as well as dynamic movements has raised the overall ratings. ANOVA model was also employed to these results (result of the test 1 tested against result of the test 2) and obtained *p* values are collected to Table 1. The values shows that the perceived difference between recorded and auralized soundtracks was smaller when subjects knew where the sound sources were and how they were moving in the space. The raised ratings for *source location* were expected, but it was surprising that also *timbre* judgments were raised significantly for three stimuli (dru, cla, and gui).

A similar comparison between test 1 and test 2 for rendering cases is presented in Fig. 9. The same trend is seen than for the different stimuli (Fig. 8); the knowledge of the size of the space and of the rendering cases complicated the perception of differ-

	dru	cla	gui	voi
Source location	0.0110	0.0109	-	0.0019
Externalization	-	-	-	0.0496
Sense of space	-	0.0347	-	0.0020
Timbre	0.0031	0.0194	0.0016	-

Table 1: The ANOVA results (p values) for the data depicted in Fig. 8 when result of test 1 is tested against result of test 2.

ences between two soundtracks. However, only in the rendering case s1 the ratings were significantly better in test 2 than in test 1 to all questions (highest p value were $p = .007$). This result show clearly that the knowledge of sound source location helps in spatial sound reproduction with headphones, especially in this rendering case s1 in which the sound source is in front of the listener. Other significant differences were with *timbre* in cases s2 ($p = .0079$) and t2 ($p = .08$).

3.7. Discussion

Asking all four questions at the same time might have affected the results. A few subjects claimed that they heard some difference but could not exactly say what this difference was. In such case the task to answer four detailed questions was too hard. In addition, the questions may not have been orthogonal, especially *timbre* and *sense of space* have been judged similarly. Also answers to *source location* and *externalization* seemed to correlate strongly.

Ratings for *timbre* in test 2 are interesting. The difference were considered significantly smaller when subjects were aware of the space and rendering cases. However, there was no changes in sound samples (they were exactly the same samples than in test 1). This fact gives us reason to suspect that in *timbre* judgments some spatial characteristics were also listened.

Some subjects claimed that it was difficult to judge differences in localization if both samples were localized inside the head. Also externalization was sometimes misunderstood, because some subjects said that they judged distance along with externalization.

The listening tests as well as the objective analysis of binaural impulse responses [13] gave us information of possible errors in modeling and auralization. It seems that biggest (the most audible) differences are at low frequencies (< 400 Hz) and at very high frequencies (> 8 kHz). The low frequency modeling error might be explained with the lack of diffraction model in our system. The high frequency difference is not so straightforward to find out, but one possible error source is the applied headphone equalization which was not exactly the same for recorded and auralized soundtracks.

These tests were the first attempts to subjectively evaluate the physics-based auralization system with dynamic rendering. Without the knowledge of how the evaluation should be done we just pointed some, in our opinion relevant, questions to the subject. Along these tests we learned that there were perhaps too many questions to answer. In future tests we plan to ask only one or two questions in each task as well as use only two or three stimuli to reduce the amount of different comparison tasks. However, these tests showed that this kind of evaluation of auralization systems is possible.

In this paper auralizations of one specific room are evaluated and the results might have been totally different if the recorded and

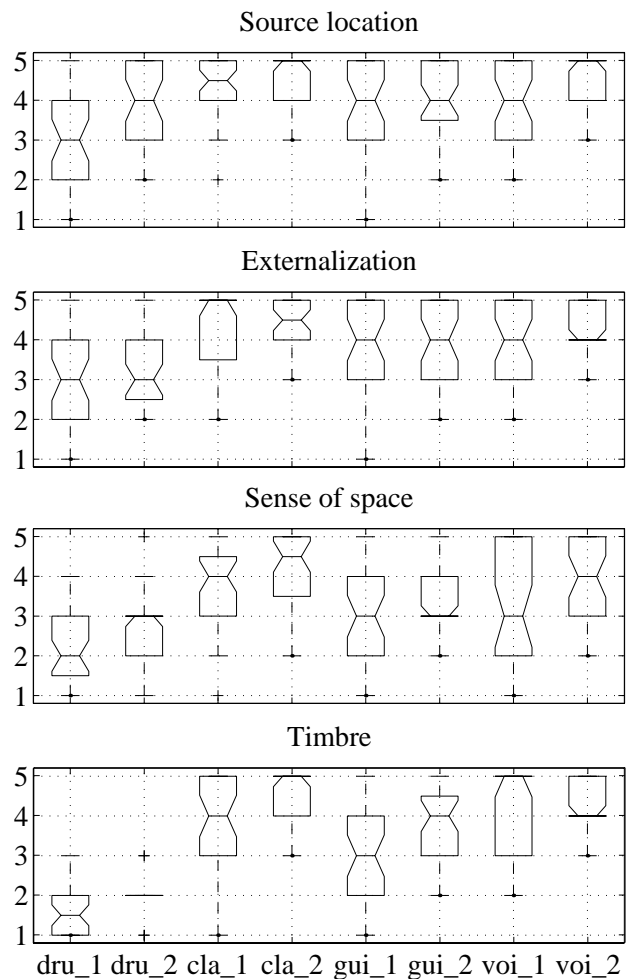


Figure 8: The results of the second listening test with each stimulus. For example, *dru_1* is the results from listening test 1 and *drum_2* is the results from listening test 2 (same six subjects). The vertical axis numbers have similar meaning than in Figs. 6 and 7.

modeled space has been more complex. Another question is if we can find overall measures of the quality of auralization systems if we don't know the target application. No general psychoacoustic explanation of the functioning of our binaural hearing in dynamic situations is known yet. We can only evaluate auralization systems in certain cases but we cannot be sure that they are producing a natural sounding virtual auditory environment in general, in all situations for all applications.

4. CONCLUSIONS AND FUTURE WORK

This paper presented the subjective evaluation of auralization. The listening test showed that it is possible to create natural sounding virtual auditory environment with physics-based room acoustics modeling and advanced digital signal processing. However, some differences between recorded and auralized soundtracks were found, especially with a transient-like stimulus signal. This result obliges us to refine our auralization model and make more listening tests before we can really claim that very natural sound-

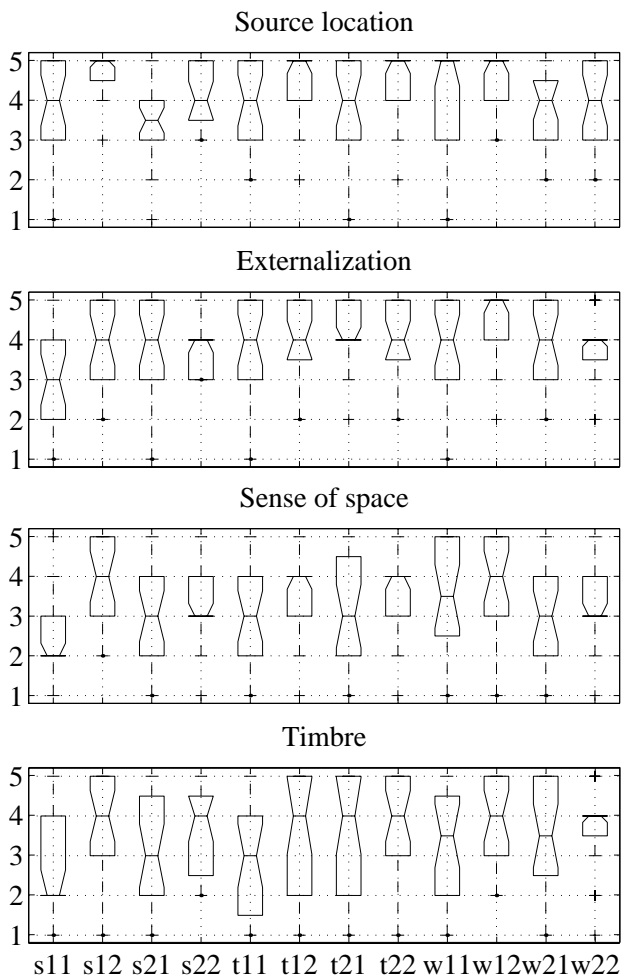


Figure 9: The results of the second listening test with each case (see Fig. 2). For example, s11 is the results from listening test 1 and s12 is the results from listening test 2. The vertical axis numbers have similar meaning than in Figs. 6 and 7.

ing virtual auditory environment with physics-based room acoustics modeling is possible to create.

The auralization method, based on image source method and artificial late reverberation, is applied in multimedia applications where real-time calculation is needed. The auralization has to be done in a way that the system fulfills real-time requirements. Another future challenge is to optimize the auralization process to enable real-time operation without reducing the perceived quality of the system. The computational load can be reduced by using shorter filters in auralization or by calculating less image sources. The effect of these reductions will be evaluated with listening tests in near future.

5. ACKNOWLEDGMENTS

This work has been financed by the Helsinki Graduate School in Computer Science and the Pythagoras Graduate School. The authors also wish to thank Nokia Research Center, Nokia Foundation, and Finnish Foundation of Technology Development (Tekni-

ikan edistämissäätiö) for financial support.

6. REFERENCES

- [1] J.-M. Jot, "Real-time spatial processing of sounds for music, multimedia and interactive human-computer interfaces," *Multimedia Systems, Special Issue on Audio and Multimedia*, vol. 7, no. 1, pp. 55–69, 1999.
- [2] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.
- [3] J. Borish, "Extension of the image model to arbitrary polyhedra," *J. Acoust. Soc. Am.*, vol. 75, no. 6, pp. 1827–1836, 1984.
- [4] P. Maijala, "Better binaural recordings using the real human head," in *Proc. Int. Congr. Noise Control Engineering (Inter-Noise 1997)*, Budapest, Hungary, Aug. 25-27 1997, vol. 2, pp. 1135–1138.
- [5] H. Møller, C.B. Jensen, D. Hammershøi, and M.F. Sørensen, "Using a typical human subject for binaural recording," in the *100th Audio Engineering Society (AES) Convention*, Copenhagen, Denmark, May 11-14 1996, preprint no. 4157.
- [6] H. Møller, "Fundamentals of binaural technology," *Applied Acoustics*, vol. 36, no. 3-4, pp. 171–218, 1992.
- [7] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen, "Creating interactive virtual acoustic environments," *J. Audio Eng. Soc.*, vol. 47, no. 9, pp. 675–705, Sept. 1999.
- [8] T. Lokki, J. Hiipakka, and L. Savioja, "A framework for evaluating virtual acoustic environments," in the *110th Audio Engineering Society (AES) Convention*, Amsterdam, the Netherlands, May 12-15 2001, preprint no. 5317.
- [9] R. Väänänen, V. Välimäki, and J. Huopaniemi, "Efficient and parametric reverberator for room acoustics modeling," in *Proc. Int. Computer Music Conf. (ICMC'97)*, Thessaloniki, Greece, Sept. 1997, pp. 200–203.
- [10] J. Hynninen and N. Zacharov, "Guineapig - a generic subjective test system for multichannel audio," in the *106th Audio Engineering Society (AES) Convention*, Munich, Germany, May 8-11 1999, preprint no. 4871.
- [11] E. Wenzel, "What perception implies about implementation of interactive virtual acoustic environments," in the *101st Audio Engineering Society (AES) Convention*, Los Angeles, Nov. 8-11 1996, preprint no. 4353.
- [12] R.S. Pellegrini, "Comparison of data- and model-based simulation algorithms for auditory virtual environments," in the *106th Audio Engineering Society (AES) Convention*, Munich, Germany, May 8-11 1999, preprint no. 4953.
- [13] T. Lokki, "Objective comparison of measured and modeled binaural room responses," in *Proc. 8th International Congress on Sound and Vibration*, Hong Kong, China, July 2-6 2001, Accepted for publication.