The Perception of System Latency in Dynamic Binaural Synthesis

A. Lindau[1]

[1] *Audio Communication Group, TU Berlin, Email: alexander.lindau@tu-berlin.de*

Motivation

In an interactive virtual acoustic environment the total system latency (TSL) plays an important role for the authenticity of simulations, be it a purely acoustic [1] or an audiovisual simulation [2]. Knowledge about thresholds of just detectable latency will allow for adequate adjustment of the rendering effort. Moreover, headroom available for additional audio processing will be determined. Most former studies examined latency by means of auralization based on anechoic head related transfer functions (HRTFs). Thus, as no reliable thresholds exist for the binaural simulation of reverberant acoustic environments this study was conducted for different acoustic environments, and using different stimuli in a criterion free adaptive psychoacoustic procedure.

Latency in VAEs

Early studies on latency in virtual acoustic environments (VAEs), did not directly evaluate the detectability of latency but measured localisation accuracy or user response times as a function of altered TSL [3][4][5]. However, as localisation accuracy was shown to be barely degraded by latencies as high as 96 [3], 150 [4], or even 250 ms [5], it can hardly be regarded as a good predictor for the detectability of system latency. In localisation tasks, latency mainly increases the response times of the subjects [4][6]. So, some of the differences in values from cited studies were suspected [5][6] to be related to limited stimulus duration. Only recently the minimum detectable latency in VAEs was directly investigated by different authors [1][7][8]. An overview of the results is given in Table 1.

Ref.	test	stimulus	subjects	mTSL	threshold*
[1]	Yes/No	noise	9	12 ms	60 ms
[7]	2 AFC	castagnets	17	50 ms**	75 ms
[8]	paired comp.	multitone	9	9.9 ms**	70 ms

Table 1: Studies on just audible latency in VAEs, *minimum observed threshold, **method of determination unmentioned

Alternatively, in [9] it was suggested to deduce minimum detectable latency from psychoacoustic results on the minimum audible angle (MAMA, [10]). This quantity describes the angle a moving source has to cover if it is to be detected as different from a stationary source. MAMAs have been found to increase (slowly) with source velocity [10] and decrease with audio stimuli bandwidth [11]. Inferring that MAMAs hold true also for a moving receiver (i.e. a rotating head) and a stationary (virtual) source, minimum latency can be calculated from MAMA. Thus, a VAE with

$$TSL < \frac{MAMA}{v_{head}} \quad [\mathrm{s}] \qquad (1)$$

should be able to render stable virtual sources, i.e. yield an inaudibly low latency. MAMAs as low as 5.9°, 8.2°, and 9.6° as reported in [10] for source velocities of 90°/s, 180°/s, and 360°/s would thus demand system latencies of 65.6 ms, 45.6 ms, and 26.7 ms resp. To examine the predicted interrelation between head movements and latencies all head tracking data was recorded in this study.

Following [5], latency and update rate are distinct but, in practice, related parameters of VAEs. Total system latency is thus defined as the temporal delay between an event such as a distinct head movement and the corresponding reaction of the VAE, i.e. rendering audio with updated HRTFs/BRIRs. Update rates are introduced when temporal sampling of the auditory scene happens. This can happen either inside the renderer, which for instance, calculates scene updates for fixed instances of time, or at the input sensing devices, which is most often a head tracker, typically exhibiting an update rate between 120 - 180 Hz. Since several elements contribute to the total system latency, VAE response times are typically distributed around a mean and have to be determined empirically [12].

In contrast to applications in augmented reality where visual or auditory real-life cues without latency are concurrently presented within the simulation, a pure audio VAE represents a worst case latency task for subjects [1]. In this case a minimum TSL (mTSL) of at least 30 - 40 ms should be proven to be obtainable (see Table 1).

Method

Binaural impulse response datasets for auralization were measured in a large lecture hall (V = 8600 m³, RT = 2.1 s) and in an anechoic chamber using the automatic FABIAN HATS (head and torso simulator) [13], which is able to move its head freely above the torso. For frontal sound incidence (source: Meyersound UPL-1) datasets were measured and auralized for a grid of horizontal and vertical head movements (horizontal ±80°, vertical ±35°) with a virtually inaudibly fine angular resolution of 2° [14]. As acoustic travel times, included in field-measured BRIRs, would directly increase latency, these have been discarded from datasets. An onset detection algorithm was used to find the earliest direct sound in each dataset. This delay, reduced by 50 samples for safely preserving the onsets, was then the removed from the datasets.

The auralization system [13] is a Linux package for fast partitioned convolution. It uses two block sizes, a smaller one for the early part of the BRIRs, and a larger one for the diffuse reverberation tail. Updating the BRIR is done via parallel spectral domain convolution and time-domain cross-fading. In order to avoid switching artefacts a short linear cross fade, corresponding to the smaller block size, is used.

Thus, the first results of a filter exchange are available one audio block after recognizing a trigger event.

Before adequately operationalizing latency for the purpose of listening tests, the actual minimum TSL has to be determined. As shown in [9][12] multiple elements contribute to TSL, such as: head tracker update rate, serial port latency, tracker library latency, network latency, scheduling of cross-fading in the convolution engine, blocksize used in convolution engine, and delays introduced by time of flight in BRIR datasets and headphone compensation filters.

In order to realize a minimum system latency, the block size of the fast convolution algorithm was set to the minimum possible value (128 samples) while minimizing dropouts.

Frequency response compensation of the STAX SRS 2050II headphones was realized with a frequency domain least squares approach with high pass regularization, whose good perceptual properties were shown in [15]. To eliminate the filter's modeling delay introduced by this method, a minimum phase approach from [16] was used, again adding 20 samples of delay for onset protection.

For head tracking a Polhemus Fastrack device was used. The specified update rate of 120 Hz could be confirmed by measurement with the serial port set to maximum speed of 115 kBaud. A mean message delay of 8.34 ms (1000 measurements, σ = 1.7 ms, min: 3.3 ms, max: 13.5 ms) was observed.

The minimum TSL was measured using a setup similar to [9]. Therefore, the head tracking sensor was attached to a mechanical swing-arm apparatus. When moving the swing-arm from its initial position (comparable to starting a head movement), an electrical circuit was broken, which caused an immediate voltage drop at a simple voltage divider circuit. The convolution engine was set up with identical parameters as in the listening test, yet a specially designed IR-dataset was used, which lead to rendering silence when tracker indicated initial position (0°/0°). Deviations of more than 1° from this position immediately caused a square wave output to be rendered (see Figure 1). Due to using this artificial dataset, the 20 + 50 = 70 samples from onset protection included in the listening test's original BRIR datasets had to be added to measurement results to obtain the actual mTSL.

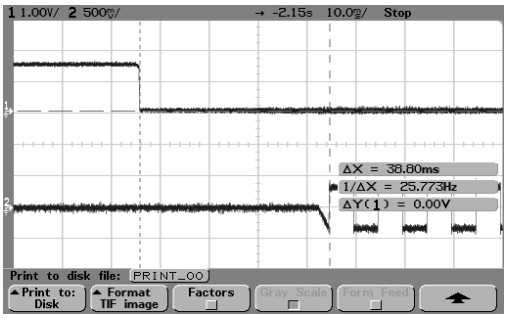

Figure 1: Screenshot from minimum TSL measurement, upper trace: falling edged indicates a starting tracker movement, lower trace: convolution output, after a linear fade-in (~block size) a square wave is rendered.

As mentioned, TSL appears as a distribution due to the interaction of different contributors' update rates or processing latencies. Figure 2 shows 60 measured values from our system; in the graph the missing 70 samples from onset protection within the BRIR data have already been added. The mean mTSL has thus determined as 43 ms (σ = 3.8 ms, range: 16 ms) which would be just sufficient, as was also confirmed by the listening test.

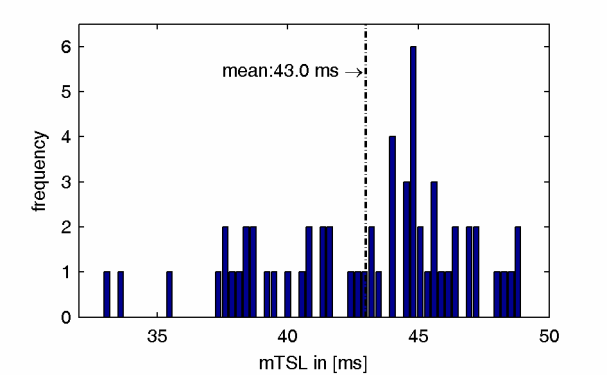

Figure 2: Distribution of measured minimum TSL values, broken line indicates mean; resulting end-to-end latency is shown (including additional 70 samples of delay from BRIR datasets and headphone compensation)

For operationalizing latency in a listening test FIFO buffering of the continuous control data stream should be employed. A prolongation of the system response time by dropping tracker update cycles, which is technically easy to implement and was used for instance in [4][7], mixes up delayed response time with reduced spatial resolution and should be avoided. In our case tracker position events, encapsulated in open sound control (OSC) messages, were FIFO cued using the 'pipe' object in the *puredata* software. Measurements proved this method to be able to reliably delay the OSC stream arbitrarily in increments of milliseconds.

Stimulus selection was evaluated in pretests (pink noise, narrow band noise, pink noise pulses, acoustic guitar, speech, and castagnets). In contrast to [7], castagnets did not show to be particularly sensitive. Pink noise (~ 6 s) and male speech (~ 4.5 s) were chosen as they induced lowest latency thresholds.

For the listening test an adaptive three alternative forced choice (3AFC) test procedure was used. Three stimuli are presented without repetition in a triad, including the reference situation with minimum TSL twice and a stimulus with increased latency in random order for each trial. After an initial training phase including feedback, each run started with a test stimulus in the middle of the range of provided latency values. Latency was then adaptively changed according to the subject's responses using a maximum likelihood adaption rule ("Best-PEST", [17]).

Pretests were conducted in order to set up the test parameters. Thus, latency values were presented within a range of [mTSL; mTSL+225ms] and adapted with a step size of 3 ms. A maximum trial number of 20 was set as stop criterion in the adaptation process. Head movements were possible within the BRIR data ranges (spoken operator guidance). During the training phase, subjects were encouraged to find individual movement strategies that would maximise their detection rate.

22 subjects took part in this study (21 male, 1 female, avg. age: 29.8), 90% had former experience in listening tests and musical education.

To test for stimulus and reverberation effects the test was designed as a full factorial 2 x 2 (2 stimuli, 2 acoustic environments) repeated measures test. All subjects had to evaluate all 4 possible situations in randomized order, resulting in 4 x 22 = 88 threshold values. The individual test duration was about 45 minutes.

The listening test was implemented in Matlab®, providing user interfaces for training and listening test. The auralization engine, the insertion of latency, and the audio playback was remotely controlled via OSC messages. Head tracking data, received as OSC messages were also recorded by Matlab® with a sampling rate of ca. 50 Hz. The head tracker was reset before each run.

Results

The distribution of all obtained threshold values is shown in Figure 4. Results are depicted for a range starting from mTSL (43 ms). The lowest threshold of 52 ms, which is only 3 increments above mTSL, was found once for the reverberant/speech condition. 64 ms were the lowest threshold found for anechoic/speech and reverberant/noise, for anechoic/noise it was 73 ms. This is in good agreement with the figures from in Table 1. The largest threshold of 187 ms was also found for the anechoic/noise condition.

Threshold values can assumed to be normally distributed (Kolmogoroff-Smirnov Test, $p > 0.2$) for all conditions. Means and deviation within the four conditions were very similar (overall mean: 107.6ms, σ: 30.4 ms). Individual thresholds were very different (see also [1][8]) though. The low value of a reliability analysis supported this finding (Cronbachs Alpha: 0.7).

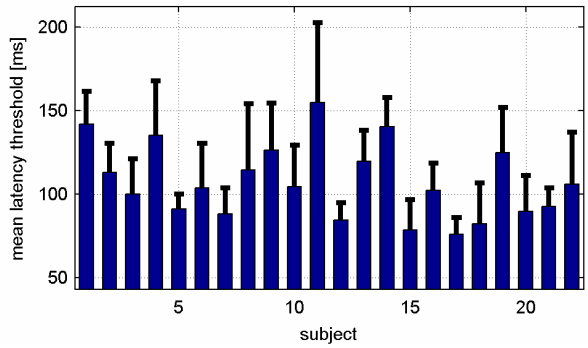

Figure 3: Minimum detectable latency thresholds w. standard deviations per subject, pooled over all 4 conditions

Finally, threshold data were analyzed by means of a 2 x 2 univariate ANOVA for repeated measures. The ANOVA proved that there were no effects of stimulus or acoustic environment. Hence, the four runs of each subject were regarded as repetitions and all data were pooled for further analysis. To further clarify the missing effects some subjects' comments on their own results shall be citied: Whereas some individuals reported the anechoic/noise situation as being best suited for a critical and undisturbed assessment of latency, others explained that speech in reverberant environment was the most natural stimulus, which made it easy to detect any unnatural behaviour within the simulation. From theory (section 2) it was expected to obtain lower thresholds with stimuli of higher spectral bandwidth (i.e. from noise).

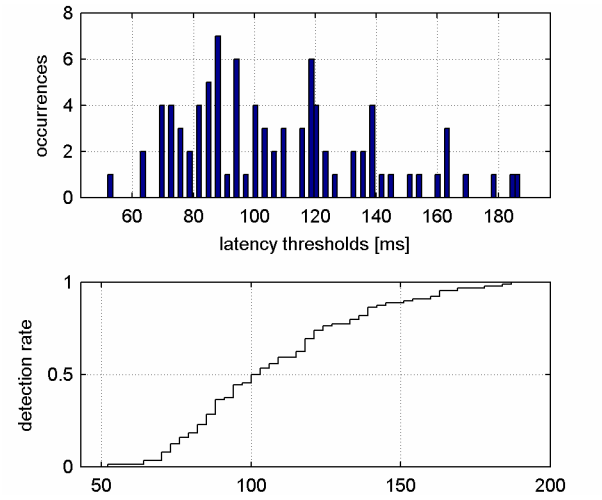

Figure 4: Above: Histogram of all latency thresholds. Below: Cumulated detection rate of latency thresholds serving as an estimate of the underlying psychometric function, (pooled for all conditions as no effect could be observed)

Head tracking data was recorded for all subjects throughout the listening test whenever a stimulus was present. Horizontal and vertical angular head position was recorded with approximately 50 Hz sampling frequency. After preprocessing of head position data, acceleration and velocity traces could be retrieved. Histograms revealed different individual strategies in use of range. Mean horizontal movement ranges were ±26°, mean velocity was 190°/s, and mean acceleration 92°/s² (excluding times were head stood still). Mean vertical movement ranges were ±11°, mean vertical velocity was 20 °/s, and mean acceleration 10 °/s².

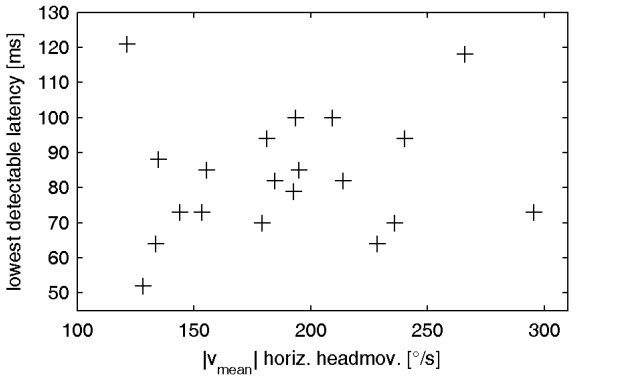

Figure 5: Scatter plot of lowest latency threshold reached by individuals vs. mean magnitude of horizontal head movement velocity in the related run, (1 subject omitted due to missing data)

From visual inspection of recorded movement traces and as values for vertical head movements were by far smaller than those for horizontal movements it is assumed, that vertical head position was somehow used as a constant offset and not altered consciously during the test. The maximum horizontal head movement velocity was 942°/s, maximum acceleration was 612.5°/s², 428°/s respectively 206°/s² were maximally observed for vertical head movements.

Somewhat surprisingly, the lowest individual latency thresholds show no linear relation to the mean horizontal head movement velocities (see Fig. 5, the best subject was a 'slow mover'). Indeed, a linear regression showed a best fit to a constant. When analyzing the *maximum* horizontal head movement velocities the same inconclusive behaviour was observed.

Discussion

Thresholds for the detection of latency in a VAE were determined in a criterion free listening test design, providing a distinct minimum TSL and operationalizing latency with a fine temporal resolution. A minimum threshold of 53 ms has been retrieved once. Mean and standard deviations of pooled threshold values are 107.63 ms resp. 30.39 ms. As normal distribution could be assumed, from these values interval estimates can be calculated. The 95% confidence interval of the mean is [101.3 ms; 114 ms]. Only three times – which is 3.4% of all measured thresholds – latencies $\leq$ 64 ms were detectable. No effect could be observed either for anechoic vs. reverberant environments or for different stimuli. From MAMA behaviour it was inferred that higher bandwidths and faster head movements would lead to lower latency threshold values. The missing stimulus effect is thus somehow unexpected, although it is admitted that a group of two stimuli constitutes no systematic variation of bandwidth. In [1] it was argued, that three subjects with lowest thresholds also showed maximal rotation speeds. Contradictory, from the published data can also be seen, that this behaviour was reversed for all 6 remaining subjects. Likewise from our data a relation between mean respectively maximum head movement velocity and latency thresholds could not be found; maybe a different predictor will be better suited. Until then, a VAE is assumed a complex system, and the underlying processes that lead to a perceptibility of latency do not seem to be reducible to stimulus bandwidth and mean or maximum velocity of head movements.

Acknowledgement

Alexander Lindau was supported by a grant from the Deutsche Telekom Laboratories.

References

[1] Brungart, D.S.; Simpson, B. D.; Kordik, A.J. (2005): "The detectability of headtracker latency in virtual audio displays." In: *Proc. of ICAD 2005 - 11th Meeting of the International Conference on Auditory Display.* Limerick

[2] Meehan, M. et al. (2003): "Effect of Latency on Presence in Stressful Virtual Environments." In: *Proc. of the IEEE Virtual Reality Conference 2003*. Los Angeles, pp. 141

[3] Bronkhorst, A.W. (1995): "Localization of real and virtual sound sources." In: *J. Acoust. Soc. Am.,* vol. 98, No. 5, pp. 2542-2553

[4] Sandvad, J. (1996): "Dynamic Aspects of Auditory Virtual Environments." In: *Proc. of the 100th AES Convention*. Kopenhagen, preprint no. 4226

[5] Wenzel, E. M. (2001): "Effects of increasing system latency on localization of virtual sounds." In: *Proc. of ICAD 2001 - Seventh Meeting of the International Conference on Auditory Display*. Boston

[6] Brungart, D.S. et al. (2004): "The interaction between head-tracker latency, source duration, and response time in the localization of virtual sound sources." In: *Proc. of ICAD 2004 - 10th Meeting of the International Conference on Auditory Display*. Sydney

[7] Mackensen, P.: *Auditive Localization. Head Movements, an additional cue in Localization.* PhD thesis, Technische Universität Berlin, 2004

[8] Yairi, S.; Iwaya, Y.; Suzuki, Y. (2006): "Investigations of system latency detection threshold of virtual auditory display." In: *Proc. of ICAD 2006 - 12th Meeting of the International Conference on Auditory Display*. London, pp. 217-222

[9] Wenzel, E. M. (1997): "Analysis of the Role of Update Rate and System Latency in Interactive Virtual Acoustic Environments." In: *Proc. of the 103rd AES Convention.* New York, preprint no. 4633

[10] Perrott, D. R.; Musicant, A. D. (1977): "Minimum audible movement angle: Binaural localization of moving sound sources." In: *J. Acoust. Soc. Am*., vol. 62, No. 6, pp. 1463-1466

[11] Chandler, D. W.; Grantham, D.W. (1992): "Minimum audible movement angle in the horizontal plane as a function of stimulus frequency and bandwidth, source azimuth, and velocity." In: *J. Acoust. Soc. Am*., vol. 91, No. 3, pp. 1624-1636

[12] Miller, J.D. et al. (2003): "Latency measurement of a real-time virtual acoustic environment rendering system." In: *Proc. of ICAD 2003 - 9th Meeting of the International Conference on Auditory Display*. Boston

[13] Lindau, A., Hohn, T., Weinzierl, S.(2007): "Binaural resynthesis for comparative studies of acoustical environments" *Proc. of the 122nd AES Convention*, Vienna, preprint 7032

[14] Lindau, A.; Maempel, H.-J.; Weinzierl, S. (2008): "Minimum BRIR grid resolution for dynamic binaural synthesis." In: *Proc. of the Acoustics '08.* Paris, pp. 3851-3856

[15] Schärer, Z.; Lindau, A.(2009): "Evaluation of Equalization Methods for Binaural Signals", to be presented at 126th AES Convention, Munich, Germany

[16] Norcross, S.G.; Bouchard, M.; Soulodre, G.A. (2006): "Inverse Filtering Design Using a Minimal-Phase Target Function from Regularization." In: *Proc. of the 121st AES Convention*. San Francisco, preprint no. 6929

[17] Pentland, A. (1980): "Maximum likelihood estimation: The best PEST." In: *Perception & Psychophysics*, vol. 28, No. 4, pp. 377-379