An initial investigation into the effects of digital audio sample rate on human perception of three-dimensional sound scenes

Will Howie

Preface (please read first)

This paper reports on a small-scale experiment I conducted during my last month as a PhD candidate at McGill University, in April 2018. The paper was submitted to a journal but subsequently withdrawn based on reviewer comments. There are some methodological choices that I now recognize introduced several unwanted variables into the experiment, making the results somewhat hazy. However, it was an interesting experiment, so I wanted to share it with those who are interested in this sort of thing. I intend to redo the experiment in the near future.

0 Introduction

In natural environments, our fine temporal resolution is necessary for rapid and accurate 360º localization of sounds such as rustling brush, crunching leaves, or snapping twigs [1], which would have been of vital importance to our hunter-gatherer ancestors. This fine temporal resolution is not only important for hearing sound in natural environments, but also plays a large role in the experience of hearing music in an acoustic space. Studies by Kuncher [2] and Krumbholz et al. [3] have demonstrated the human auditory system can discriminate timing differences as small as 5–20 µs between monaural sounds. Brughera et al. [4] showed listeners could detect interaural time differences of around 10–11µs for sine tones, a figure similar to what was found in previous studies by Zwislocki and Feldman [5], and Klumpp and Eady [6]. For band-limited random noise, Klumpp and Eady [6] found the threshold for detection of interaural time differences averaged 9 µs. Several musical instruments such as xylophone, trumpet, snare drum, and cymbals have been shown to have very steep transient onsets, reaching sound pressure levels greater than 120dB in less than

10 μs [7]. On the subject of reflected sound in a room, Kuncher states: "A transient sound produces a cascade of reflections whose frequency of incidence upon a listener grows with the square of time; the rate of arrival of these reflections dN/dt ≈ 4πc3t2/V (where V is the room volume) approaches once every 5 μs after one second for a 2500 m3 room [2]". Temporally dense transient aspects of natural, musical, and acoustic sounds, therefore, constitute an important part of our total listening experience.

Several authors have discussed the concept of "time-smearing", a broadening of transient impulses in captured sound caused by brick-wall filters present in analog to digital converters and downward sample rate conversion [2], [8], [9]. As the sample rate increases, "smearing" caused by pre and post-ringing around the impulse decreases, e.g. a sample rate of 96 kHz should introduce less time-smearing than 48 kHz. This "smearing" of transients may be responsible for important details in sound recordings being obscured, such as reflected sound [1], [9] or pitch information [10]. An increased perception of reverberant information, improved sound source localisation and timing information, and clarity of harmonic content are often given as anecdotal reasons why recording engineers and music producers chose to record audio at "higher" sample rates, such as 96 kHz or 192 kHz.

In his meta-analysis of previous research into human perception of high resolution audio, Reiss found "a small but statistically significant ability to discriminate between standard quality audio (44.1 or 48 kHz, 16bit) and high resolution audio (audio beyond standard quality). [11]" Reiss identified 80 relevant studies in his review of previous research of human perception of high resolution audio. For his meta-analysis, he focused on 18 studies that were related to discrimination between standard and higher sample rates [11]. A review of these 18 studies reveals that all except one used audio stimuli that were 2-channel stereo or mono. The sole study to use multichannel audio stimuli was Woszczyk et al. [12], which asked listeners to compare three versions of the same 6-channel sound scene: a straight analog feed from the microphones, and those same microphone signals sampled at 44.1 kHz and 352.8 kHz. The sound scene used in [12] was mechanical in nature, constantly shifting, and not necessarily representative of many real-life listening experiences. Also, the authors fail to give a detailed description of how the various mechanical sounds were presented spatially to the listener, especially those reproduced through two elevated ribbon-tweeters.

Stereophonic sound recordings deliver a decidedly limited reproduction of a given sound scene: 360º sonic information, as exists in natural hearing, is reduced to a single plane of sound with a horizontal extent of ±30º. This will naturally result in a great deal of perceptual masking of sound, including the complex, dense late reflections within a room that we normally hear from all around us. Previous studies [13 – 17] have shown that aspects of perceived spatial impression in sound reproduction related to late reflected sound energy, such as "envelopment" or "presence," improve as the number and spatial distribution of loudspeakers in a given audio reproduction system increases. This suggests that as captured and reproduced sound information approaches a level closer to natural hearing in a real acoustic environment, our ability to cognitively separate direct and reflected sound improves. A better separation of direct and diffuse sound should also result in a finer appreciation of the micro-timing differences in direct sounds that contain fast transients. It is possible, then, that with recordings made for three-dimensional playback, perceptual differences between standard and high-sample rate audio formats will become more obvious, as these presentations should lack much of the spatial or spectral masking present in stereo sound reproduction. And yet, the effects of digital audio bandwidth within the context of three-dimensional sound reproduction remain largely unexplored.

A number of recent studies have focused on areas related to the capture [18–21] and reproduction [17, 22–25] of three-dimensional audio. As a first step towards investigating the relationship between sample rate in digital audio systems and human perception of natural and musical auditory scenes within three-dimensional sound reproduction, this pilot study aims to address the following question: "Are listeners able to consistently correctly discriminate between three-dimensional audio stimuli captured at two different sample rates: 48 kHz, the current standard for most film, broadcast, and commercial music production, and 384 kHz, the highest PCM linear sample rate available with current commercial audio technology?"

1 Preparation of Testing Stimuli

1.1 Recording signal flow

For ease of facilitating this pilot study within the means of available equipment and facilities, it was decided to use a simple 9-channel (4+5+0) [26] 3D audio format for stimulus recording and reproduction, and to limit the range of stimuli to musical sounds. No attempt was made to use microphones or loudspeakers specially designed for high resolution audio capture or playback. Such equipment is rare within both commercial music production and home playback environments, and is therefore not representative of typical end-to-end audio transmission. A 9-channel microphone array was setup in a 560-seat concert hall with an average RT60 of 1.8 s. The microphones were optimized to capture single instruments, presenting the listener with a "concert" or "cinematic" perspective. The array is based on a larger-scale 3D music recording technique described in [27]. Three omni-directional microphones (Schoeps MK2H) were positioned to capture primarily direct instrumental sound, while widely spaced directional microphones (Schoeps MK21 in the main layer and MK4 in the height layer), were optimized to capture decorrelated, diffuse reflected sound energy (Figure 1). A one-to-one relationship was maintained between microphone signals and corresponding loudspeakers.

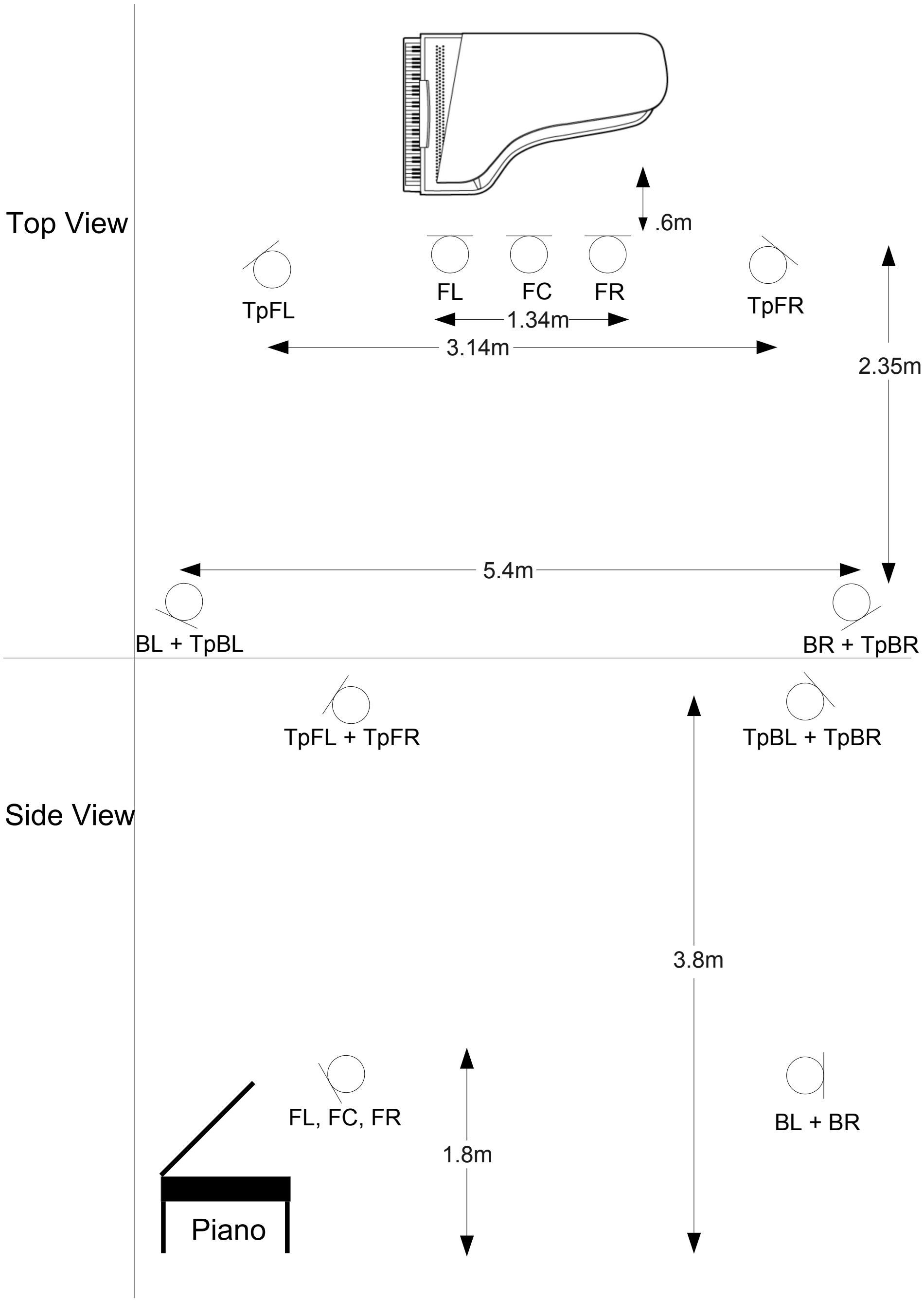


Figure 1. 3D Microphone Array. Microphone signal nomenclature as per [26]

Microphone signals were routed to a pair of 8-channel Millennia HV-3R microphone preamplifiers. From the XLR outputs of the preamps, each microphone signal was split passively, then routed to two sets of Merging Technologies *Sphynx 2* 8-channel analog-to-digital converters. One pair of converters was set to 384 kHz / 24 bit resolution, the other to 48 kHz /24 bit resolution. For each pair of converters, one unit acted as the sync master, clocked internally, while the other unit was a sync slave. It was not within the technical means of available equipment to clock both pairs of converters to the same master word clock. However, as all four of these units were purchased at the same time, and are of the same manufacturing generation, a negligible difference in internal base clock speed was assumed between pairs. 384 kHz signals were routed, via MADI, to an onstage computer for recording, monitored via headphones. 48 kHz signals were routed, via MADI, to McGill University's Studio 22 for recording and monitoring over loudspeakers. Studio 22 is equipped with 28 full-range, two-way loudspeakers (ME Geithain *M-25*) powered by Flying Mole class D amplifiers, and an Eclipse TD725SWMK2 stereo sub-woofer. The loudspeakers are arranged for reproduction of both 22.2 Multichannel Sound, i.e. 9+10+3, and 4+5+0, as per [26]. The room's dimension ratios and reverb time fulfill ITU-R BS.1116 [28] requirements. Both recording computers were running Merging Technologies *Pyramix* digital audio workstation.

1.2 Musical sound sources

Three musical instruments were chosen as sound sources for the listening test: piano, snare drum, and crotales. Piano was chosen for its combination of percussive attacks and complex timbre and tone colours, as well as large physical extent. Snare drum was chosen as this instrument is known to have very fast, very steep transient onsets [12]. The crotales were chosen for their long, clear ringing, which contains many overtones.

1.3 Recording and mixing of stimuli

The microphone array was initially positioned and optimized for recording the solo piano by a professional recording engineer with significant experience recording and mixing for various two and three-dimensional audio formats. For the snare drum and crotales, both instruments were positioned the same distance from the main front microphones as the piano, thereby capturing a consistent proximity perspective between all three instruments. The various musical excerpts performed on each instrument were recorded simultaneously to both recording systems at both sample rates under investigation.

Three musical excerpts were chosen to be used as testing stimuli: one per instrument. The piano excerpt is a 15 s passage from an improvised jazz solo, and makes use of a wide range of the keyboard. The strong attacks in the playing style of the pianist made for a prominent activation of the recording venue's acoustic signature. The snare drum excerpt is a continuous roll, 22 s in duration, which crescendos from *pianissimo* to *fortissimo*, decrescendos, and then repeats the same dynamic pattern once more, resulting in a sound scene of dense transient information. For the crotales, a dominant 7th chord is performed, one note at a time, at a very slow tempo. This is followed by a resolution to the tonic, which is allowed to resonate for several seconds; the excerpt is 19 s in duration. To confirm the capture of extended bandwidth audio content within the 384 kHz recordings, the audio files for these three excerpts were analyzed using a high-precision software audio analyzer. Images of the spectrogram of each sound source/excerpt are shown in Figures 2–4. Aside from musical content, these images also reveal the pattern of the noise shaping within the analog-to-digital converters.

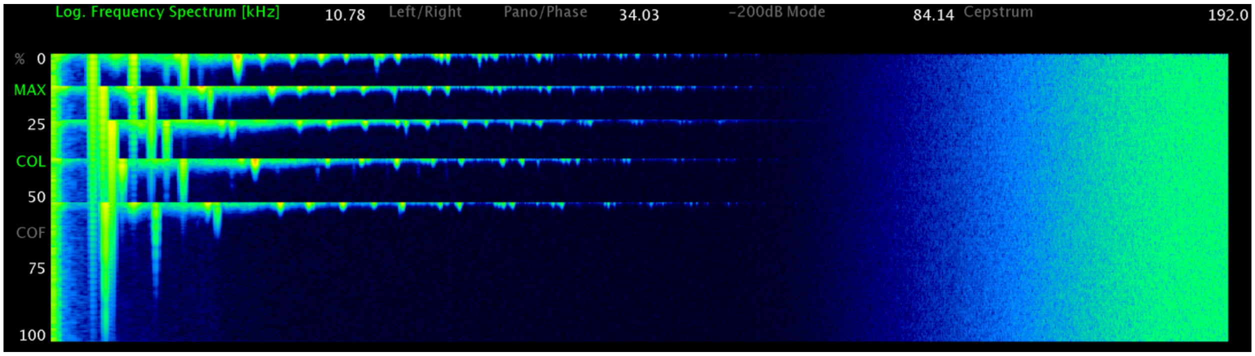


Figure 2: Spectrogram for Crotales. X-axis represents the logarithmic frequency scale, Y-axis represents time.

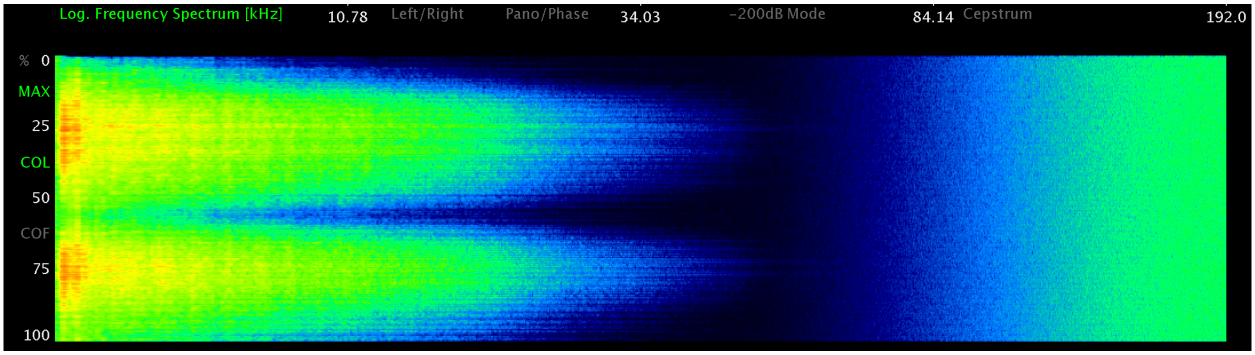
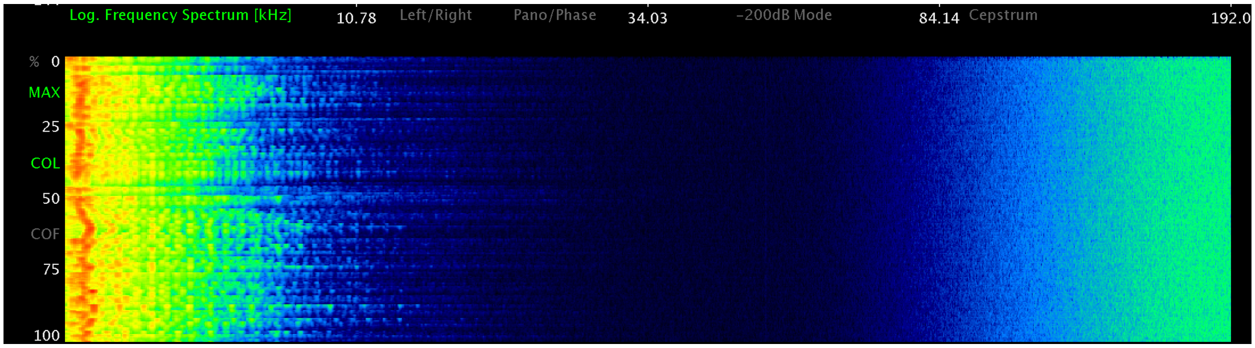


Figure 3: Spectrogram for Snare Drum. X-axis represents the logarithmic frequency scale, Y-axis represents time.



Figures 4: Spectrogram for Piano. X-axis represents the logarithmic frequency scale, Y-axis represents time.

To facilitate mixing and playback of stimuli captured at both sample rates within the same Pyramix session, the 48 kHz audio files for each musical excerpt were sample-rate converted to 384 kHz. Sample-rate conversion was done in Pyramix, using the “apodizing” filter in the HeptaCon sample rate converter.

Stimuli mixing took place in the Critical Listening Lab (room A817) at the Centre for Interdisciplinary Research in Music Media and Technology (Figure 5). Five B&W 802D loudspeakers powered by a Classé CA5200 amplifier provided playback for main-layer microphone signals, while four Genelec 8030 powered loudspeakers were used as height channels. The loudspeakers were arranged for 4+5+0 reproduction, as per ITU-R BS.2051-0 recommendations [26]. The height channels in the Critical Listening Lab were positioned directly above the main layer loudspeakers. Merging Technologies *Sphynx 2* digital-to-analog converters were used for digital audio playback. The room measures 4.85 m by 4.5 m by 3.3 m, meeting all ITU-R BS.1116 [28] geometric properties requirements for a "reference listening room" for multichannel audio, except for "room size". Reverb time (RT60) and the operational room response curve are also within ITU-R BS.1116 requirements. Background noise does not exceed NR 20.

The microphone signals for each musical excerpt were balanced by the recording engineer. Balances were optimized to maintain a perceptually even, consistent direct-to-reverb ratio between stimuli. Microphone signal balances between 384 kHz and 48 kHz stimuli were kept identical per musical excerpt. Objective loudness measurements were taken to confirm that no significant difference in level exists between the final balanced 9-channel 384 kHz and 48 kHz stimuli, per musical excerpt. Measurements were performed using a B&K Type 2250 Hand-Held Analyzer set to dBC and slow time weighting. The unit displays digital values to one tenth of a dB. The B&K 2250 was positioned at a point equidistant from the main-layer loudspeakers used in the 4+5+0 configuration, on a stationary tripod, set at a height typical of the average seated listener's ears. For each stimulus, the entire musical excerpt was analyzed, with the peak dBC value of the excerpt being recorded. Peak values between 48 kHz and 384 kHz stimuli, per musical excerpt, were found to be within 0.1dB of each other. Subsequently, integrated loudness measurements were performed on each audio file used as experimental stimuli. Measurements were taken using Merging Technologies' "Final Check" software, which includes an EBU R-128 [29] compliant loudness meter. The meter was set to EBU Mode (R-128) and EBU +9 absolute (LUFS), with a peak hold of 5.0 s. Dithering error and energy from musical signals occupying upper frequency bands (see: Figures 2–4) may affect loudness measurements in a way that is not ecologically valid in terms of human frequency resolution or loudness perception. To avoid the influence of inaudible spectra on the loudness meter, a low pass filter at 20 kHz, 12dB per octave slope, was applied to all audio files, using the "EQ-X" digital equalizer. Per musical excerpt, all matching pairs of 384 kHz and 48 kHz audio files were found to be within 0.1LUFS of each other.

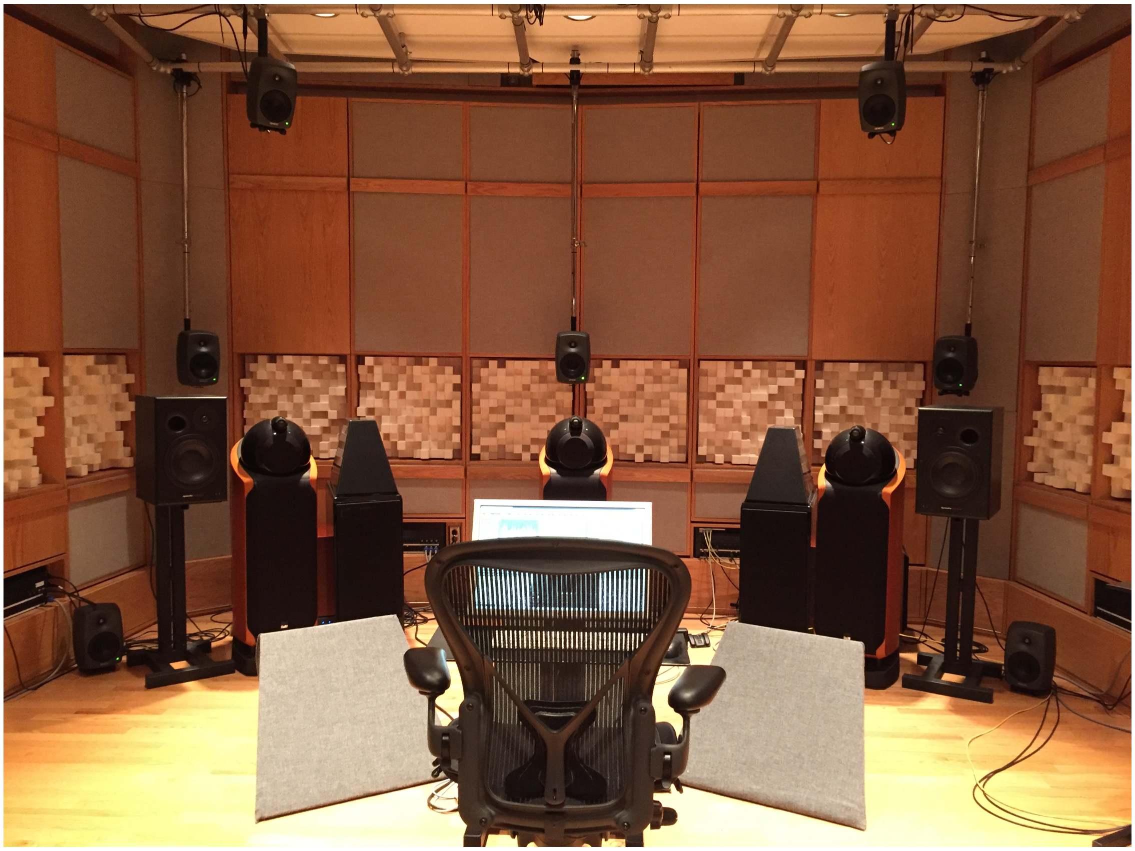


Figure 5. Critical Listening Lab

2 Listening Test

A listening test was designed to determine whether subjects could consistently discriminate between 9-channel three-dimensional sound recordings of musical instruments for two different sample rates: 48 kHz and 384 kHz.

2.1 Subjects

10 subjects performed the listening test. All were current students within the Graduate Program in Sound Recording at McGill University. All subjects had completed at least one year of technical ear training, and so were familiar with audio stimuli comparison/matching-type exercises. All subjects had at least 2 years audio recording and production experience, and had at least one hour of previous experience listening to three-dimensional music recordings. Nine of the ten subjects had more than 10 years of musical training. All reported having normal hearing.

2.2 Listening Test

Listening tests took place in the above described Critical Listening Lab. Merging Technologies' Pyramix digital audio workstation was used as the testing interface. Subjects were seated at a point equidistant from the main-layer loudspeakers used in the 4+5+0 configuration.

Before undertaking the listening test, subjects were presented with a Pyramix session in which they could listen to all of the stimuli that would be heard within the listening test. For each musical excerpt, subjects could switch between VCA faders labelled "A" and "B", each of which represented a different sample rate. If asked, the testing administrator would specify what the sample rates were (some subjects preferred not to know). Subjects were told to take several minutes to carefully compare the two sample rates for each musical excerpt, and to get an impression of what perceptual differences may exist between the two. The researcher was not present in the room while this orientation activity took place.

Once subjects had completed the orientation, they were presented with a new session window with 12 multichannel audio clips labelled “1” through “12”. For a given trial, subjects were instructed to listen to a clip using a looped-playback function. Alternately, subjects could select and loop a shorter segment of the clip if they wished to focus on one specific moment within the musical excerpt. The mixer window within the Pyramix session contained three VCA faders labelled “A” “B” and “C”. Subjects were instructed to compare these three versions of each excerpt at their leisure and determine which two were the same: a standard ABC triad test. A triad test was chosen over the traditional ABX test since it results in a lower threshold for random chance guessing (33.3% instead of 50%) and is recommended within ITU-R BS.1116 [28]. Answers were recorded on an online form. Each subject performed 4 trials per musical excerpt, for a total of 12 trials. Subjects took an average of 20–30 minutes to complete the test. Upon completing the test, subjects were asked to comment on any aspects of the sound scene they felt changed consistently between sample rates, per musical excerpt.

For each trial, 48 kHz and 384 kHz stimuli assignments to VCAs A, B, and, C were determined by a random list generator whose number lists are based on atmospheric noise [30]. Per trial, stimuli A, B, and C were time aligned to within less than 4 samples at 384 kHz, i.e. 1 sample at 48 kHz, allowing for seamless switching. The presentation order of musical excerpts within the test was also randomized using the same random list generator. This was not a double-blind listening test, as the researcher who prepared the Pyramix session knew the arrangement of stimuli. This compromise in test design was primarily due to the inability of typically used audio testing platforms, such as Max/MSP, to playback 384 kHz audio files. The researcher was not present in the room while participants took the listening test.

3 Results

3.1 Pooled discrimination rates

10 subjects performed 12 trials each, for a total of 120 trials. For the first analysis, all subject data was pooled together. Table 1 shows the success rates and results of 4 binomial tests for pooled subject discrimination between sample rates, both overall and per musical excerpt. As can be seen, an overall discrimination rate of 66% was achieved, which the binomial test shows to be highly significant. Significant discrimination rates were also achieved when considering each musical excerpt individually; piano: 62%, snare: 68%, crotales: 68%. Results of a chi-squared test show that the difference between the discrimination rates per musical excerpt is not significant: X^2 (2) = 0.296, p = 0.86.

Table 1. Binomial test on sample rate discrimination (chance probability = 0.33). With α = 0.05 and Bonferroni correction, the significance threshold for each musical instrument is p = 0.017.

Data Group	**Discrimination**	**95% Conf. Interval**	**p**
Total	0.66	0.57–0.74	<0.001
Piano	0.62	0.46–0.77	<0.001
Snare	0.68	0.51–0.81	<0.001
Crotales	0.68	0.51–0.81	<0.001

3.2 Individual discrimination rates

Overall sample rate discrimination rates per subject were also considered. Table 2 shows the results of 10 binomial tests, one for each subject’s responses. Nine out of ten subjects performed better than chance (0.33), however these results were only significant for four subjects after applying Bonferroni correction. Subjects 1 and 3 performed the task with a very high degree of accuracy, reaching discrimination levels of 100% and 92% respectively. Interestingly, both had significant previous experience comparing musical performances recorded at high sample rates versus standard sample rates. Subjects 4 and 6, who each achieved a success rate of 75%, both have a background in technical ear training instruction.

Table 2. Binomial test on sample rate discrimination per subject (chance probability = 0.33). The significance threshold after Bonferroni correction is p = 0.005.

Subject	**Discrimination**	**95% Conf. Interval**	**p**
1	1.00	0.73–1.00	**<0.001**
2	0.67	0.35–0.90	0.014
3	0.92	0.61–1.00	**<0.001**
4	0.75	0.43–0.94	**0.003**
5	0.50	0.21–0.79	0.108
6	0.75	0.43–0.94	**0.003**
7	0.33	0.10–0.65	0.238
8	0.67	0.35–0.90	0.014
9	0.42	0.15–0.72	0.188
10	0.58	0.28–0.85	0.045

3.3 Pertinent perceptual differences between stimuli

During brief post-test interviews, each subject was asked to comment on what differences within the sound scene were useful cues for discriminating between stimuli. No subjects suggested that any perceptional difference in level existed between stimuli. Responses from the subjects whose individual rates of discrimination were significantly above chance (1, 3, 4, 6) were analyzed in an attempt to extract salient perceptual differences between the sound scenes captured at 384 kHz versus 48 kHz.

For the piano, the main differentiating factor was an overall change in timbre between stimuli. It was felt that one version was somewhat brighter than the other, and this brighter version was generally assumed to be 384 kHz. Amount of perceived “air” in the recording was also reported by several subjects.

For the crotales, all four of these subjects commented on a subtle difference in the decay of the musical excerpt, after the dominant 7th chord resolves to the tonic. The 384 kHz audio was felt to give a more well-defined pitch centre to the tonic note.

With the snare drum, there was somewhat less consensus. Subjects 1 and 4 commented on how the density of the sound located behind the listener would change between stimuli, with one version giving a better perception of individual attacks within the reflected sound. Subject 6 focused on the very peak of the roll's crescendo. Based on the orientation session, the subject felt that at 384 kHz the peak of the crescendo was more noticeable. Conversely, subject 3 focused on the quietest moments within the snare roll, listening for differences in the sound of the snare hits that were too quiet to substantially activate the ringing of the drum.

4 Discussion

4.1 Overall Rate of Discrimination

The results summarized in Table 1 show that listeners in this study could discriminate between three-dimensional reproductions of musical sound scenes captured at 384 kHz and 48 kHz with a statistically significant, relatively high degree of accuracy: 66%. This rate of successful discrimination did not change significantly between musical excerpts. This is a much higher figure than the overall result reported in Reiss's [11] meta-analysis of high resolution audio perceptual evaluation: 52.3%. It is also important to note that the chance success rate in this study was 33.3%, in contrast to 50% in studies examined by Reiss. For example, within the previous literature catalogued by Reiss [11], only Theiss and Hawksford [31] reported a higher rate of discrimination than the current study: 74%, though with a much larger confidence interval [11]. Subjects in that study achieved a mean discrimination rate 24% higher than chance (74% – 50%). In the current study, subjects achieved a mean discrimination rate 33% higher than chance (66% – 33%). Several other previous studies also reported relatively high discrimination rates: Yoshikawa et al. [32] with 64%, Mizumachi et al. [33] with 63%, and Jackson et al. [8] with 61%. Although these four previous studies had relatively different aims and methodologies, an important feature common to all was that subjects acquired significant training prior to performing the listening test(s). Reiss [11] showed that studies using subjects who had received detailed training, such as explanations or examples of what to listen for, reported a stronger ability to discriminate high resolution audio than those studies where subjects received little or no training. In the current study, listeners were given only a short orientation session, and were not told of any specific sonic attributes or artifacts to listen for. This all suggests that certain perceptual effects of capturing sound at higher sample rates become more audible within the context of three-dimensional audio, though additional studies with a greater number of subjects would be necessary to confirm this hypothesis.

4.2 Subject training and previous experience

When considering the individual results of each subject's performance (Table 2), it becomes clear that the two subjects (1 and 3) who reported having previous experience comparing audio content recorded at high sample rates vs 44.1 or 48 kHz had the strongest ability to discriminate between the two sample rates under investigation. The next best performing subjects (4 and 6) both teach courses in technical ear training, and thus are regularly engaged in identifying and explaining perceptual differences between audio stimuli. These results are in keeping with Reiss's [11] findings on the importance of listener training for performing audio resolution discrimination tasks. Mizumachi et al. [33] compared 192 kHz/24bit PCM audio with 48 kHz/16bit PCM audio, as well as two lossy-compressed MPEG audio formats, within the context of in-vehicle listening. They found that the ability of subjects to significantly discriminate between the 192 kHz and 48 kHz PCM formats depended on whether or not they had significant familiarity with listening to high resolution audio. It would be valuable to confirm and quantify the effect of training on human perception of high resolution audio. One possibility would be to perform an investigation similar to the current study, but using two different listener groups: trained and untrained, whose success rates could then be compared. Training could consist of one or more guided listening sessions, wherein the subjects are shown specific aspects of the 3D sound scene that change when the resolution of the stimuli increases or decreases. These perceptual aspects could be determined in advance by a panel of expert listeners who are well experienced with listening to high resolution audio.

4.3 Pertinent perceptual differences between stimuli

In two different studies investigating frequency discrimination in human hearing, Moore and his co-authors suggest that fine temporal information is necessary for good discrimination of the fundamental frequency of complex tones [34, 35]. This appears to be reflected in the current study, where there was universal agreement among the top performing subjects that for the crotales example, the sample rate that they knew or assumed to be 384 kHz gave a better resolution of the fundamental pitch of the tonic note. One subject described the notes of the chord as being "more accurately centred – I could better understand the intonation of each note, especially the last two in the sequence." A similar effect was observed by subjects in Theiss and Hawksford's study [31], who commented on a greater reproduction of the melodic lines within the high sample rate stimuli. In studies by Kanetada et al. [36] and Pras and Guastavino [37], listeners commented on aspects of clarity, spatial impression, and timbre as being key subjective differences between standard audio quality (44.1 or 48 kHz) and higher resolution audio stimuli. These observations are consistent with the impressions of listeners in the current study; a number of subjects commented on the change in timbre between sample rates for the piano excerpt, with one example being distinctly brighter and possessing more clarity. For the snare drum excerpt, several subjects commented on a change of spatial impression within the side and rear room sound between stimuli. Here we may be seeing the benefit of the smaller sampling window and reduced transient smear found in the 384 kHz audio, as compared to 48 kHz, which should allow for a more accurate capture and reproduction of dense reflected sound energy.

4.4 Type and length of stimuli for future studies

This study was meant as a preliminary investigation into the effects of digital audio sample rate on perception of three-dimensional sound scenes. For the sake of simplicity, a limited number of musical sound sources were used, while the testing methodology was aimed more towards identifying sonic differences between audio sample rates related to temporal resolution. Ideally, this study would the first step in a larger body of research that could examine this topic with a more expansive perspective. Within the field of neuroscience, Oohashi and his collaborators have conducted several studies into what measurable effects high frequency sound has on brain activity, concluding that inaudible high frequency sounds with a nonstationary structure (e.g. music) cause "nonnegligible effects on listeners when coexisting with audible low-frequency sounds. [38]" This phenomenon, which Oohashi terms the "hypersonic effect" was further investigated by Kuribayashi et al. [39], who concluded that this effect only becomes significant after a period of listening greater than ~150 s in duration, for a given high resolution audio stimulus. The current study used short musical excerpts, between 15 to 22 s in duration, as per ITU-R BS.1116 [28] guidelines, and allowed subjects to freely switch between stimuli. In order to investigate possible perceptual effects caused by the "hypersonic effect", much longer stimuli would be required, recorded and reproduced with microphones and loudspeakers designed specifically with high resolution audio content in mind. There is also the question of type of stimulus content to consider. As discussed in the introduction, human hearing evolved to identify and localize natural sounds, not necessarily musical in nature. Ideally, additional studies investigating temporal resolution in 3D audio reproduction would include recordings of natural sound scenes, which could be captured either outdoors with portable equipment (e.g. a forest soundscape) or artificially created in a sound stage using techniques drawn from Foley and sound effects design. Of particular benefit would be sounds that are transient in nature, such as snapping branches or dried leaves crushed by footsteps. Finally, there is the question of the spatial resolution of the stimuli. Of the currently standardized channel-based 3D audio formats, 22.2 Multichannel Sound (22.2), or 9+10+3, has the greatest number of and most even spatial distribution of points of sound reproduction [26], and has been shown to be perceptually unique among common 3D audio formats for the reproduction of acoustic music [17]. Ideally, the next phase of this research would be conducted using stimuli and a playback environment optimized for 22.2 or a format with a similar channel count and configuration, so as to deliver sound scenes that better match the spatial density of real-world listening.

4.5 Limitations of this study

The relatively small sample size in this study limits the statistical power of the results, and ability to generalize any analysis to a larger population. A larger number of subjects spread out between several testing venues would have been preferable, but was not within the means of this pilot study. Another consideration for a similar future study would be the inclusion of in-depth objective analysis of the audio signals used within the experiment, which may help to clarify what differences between stimuli are being observed by subjects. Several factors related to the technical setup of the stimulus recording could have introduced unknown variables into the listening experiment. As the two pairs of analog-to-digital converters used for stimulus recording were not clocked to a single master, it is possible that a small amount of drift was introduced between the 384 kHz and 48 kHz recordings. And although all analog-to-digital and digital-to-analog converters used for this study were of the same model and manufacturing generation, minute but detectable sonic differences may still exist between them. Any of these technical factors could have contributed to the ability of listeners to discriminate between stimuli in a way that was unknown to the investigator.

Acknowledgements

This research was supported by the Social Sciences and Humanities Research Council, and the Centre for Interdisciplinary Research in Music Media and Technology. Thanks to Malcolm Sailor, Denis Martin, and Jack Kelly for their assistance with stimulus recording. Thanks to Richard King for audio production and testing facilitation.

References

[1] M. Lewicki, "Efficient coding of natural sounds," *Nature Neuroscience,* vol. 5, pp. 356-363 (2002), https://doi.org/10.1038/nn831.

[2] M. Kunchur, "Audibility of temporal smearing and time misalignment of acoustic signals," *Technical Acoustics,* vol. 17 (2007).

[3] K. Krumbholz, et al., "Microsecond Temporal Resolution in Monaural Hearing without Spectral Cues?" *J. Acoust. Soc. Am.*, vol. 113, pp. 2790–2800 (2003), https://doi.org/10.1121/1.1547438.

[4] A. Brughera et al., "Human interaural time difference thresholds for sine tones: The high-frequency limit," *J. Acoust. Soc. Am.*, vol. 133, pp. 2839-2855 (2013), https://doi.org/10.1121/1.4795778.

[5] J. Zwislocki and R. S. Feldman, "Just noticeable differences in dichotic phase," *J. Acoust. Soc. Am.*, vol. 28, pp. 860–864 (1956), https://doi.org/10.1121/1.1908495.

[6] R. B. Klumpp and H. R. Eady, “Some measurements of interaural time difference thresholds,” *J. Acoust. Soc. Am.*, vol. 28, pp. 859–860 (1956), https://doi.org/10.1121/1.1908493.

[7] W. Woszczyk, “Physical and perceptual considerations for high-resolution audio,” in *AES Convention 115*, New York (2003).

[8] H. Jackson et al., "The audibility of typical digital audio filters in a high-fidelity playback system," in *AES Convention 137*, Los Angeles (2014).

[9] P. C. Craven, "Antialias Filters and System Transient Response at High Sample Rates," *J. Audio Eng. Soc.,* vol. 52, pp. 216-242 (2004, Mar.).

[10] B. Moore, "The Role of Temporal Fine Structure Processing in Pitch Perception, Masking, and Speech Perception for Normal-Hearing and Hearing-Impaired People," *Journal of the Association for Research in Otolaryngology,* vol. 9, pp. 399-406 (2008), https://doi.org/10.1007/s10162-008-0143-x.

[11] J. Reiss, "A Meta-Analysis of High Resolution Audio Perceptual Evaluation," *J. Audio Eng. Soc.,* vol. 64, pp. 364-379 (2016, Jun.), https://doi.org/10.17743/jaes.2016.0015.

[12] W. Woszczyk et al., “Which of the Two Digital Audio Systems Best Matches the Quality of the Analog System,” in *AES 31st International Conference*, London (2007).

[13] H. Shim et al., “Perceptual evaluation of spatial audio quality,” in *AES Convention 129*, San Francisco (2010).

[14] K. Hamasaki et al., “Effectiveness for height information for reproducing presence and reality in multichannel audio system,” in *AES Convention 120*, Paris (2006).

[15] K. Hamasaki et al., “Advanced multichannel audio systems with superior impressions of presence and reality,” in *AES Convention 116*, Berlin, Germany (2004).

[16] S. Oode et al., “Dimensional Loudspeaker Arrangement for Creating Sound Envelopment,” IEICE Technical Report, EA2012-46 (2012).

[17] W. Howie, R. King, D. Martin, “Listener Discrimination Between Common Speaker-Based 3D Audio Reproduction Formats,” *J. Audio Eng. Soc.*, vol. 65, pp. 796-805 (2017, Oct.), https://doi.org/10.17743/jaes.2017.0030.

[18] W. Howie et al., “Subjective and objective evaluation of 9ch three-dimensional acoustic music recording techniques,” in *AES International Conference on Spatial Reproduction – Aesthetics and Science*, Tokyo (2018). (accepted)

[19] K. Hamasaki and W. Van Baelen, “Natural Sound Recording of an Orchestra with Three-Dimensional Sound,” in *AES Convention 138*, Warsaw (2015).

[20] B. Martin et al., “Subjective Graphical Representation of Microphone Arrays for Vertical Imaging and Three-Dimensional Capture of Acoustic Instruments, Part I,” in *AES Convention 141*, Los Angeles (2016).

[21] H. Lee and C. Gribben, “Effect of vertical microphone array spacing for a 3D microphone array,” *J. Audio Eng. Soc*., vol. 62, pp. 870-884 (2014, Jan.), https://doi.org/10.17743/jaes.2014.0045.

[22] H. Lee, “2D-3D ambience upmixing based on perceptual band allocation,” *J. Audio Eng. Soc*., vol. 63, pp. 811-821 (2015, Nov.), https://doi.org/10.17743/jaes.2015.0075.

[23] H. Wierstorf et al. “Listener Preference for Wave Field Synthesis, Stereophony, and Different Mixes in Popular Music,” *J. Audio Eng. Soc*., vol. 66, pp. 385–396 (2018, May), https://doi.org/10.17743/jaes.2018.0019.

[24] H. Lee, “Sound Source and Loudspeaker Base Angle Dependency of Phantom Image Elevation Effect,” *J. Audio Eng. Soc.*, vol. 65, pp. 733–748 (2017, Sept.), https://doi.org/10.17743/jaes.2017.0028.

[25] J. Francombe et al., “Evaluation of spatial audio reproduction methods (Part 2): Analysis of listener preferences,” *J. Audio Eng. Soc*., vol. 65, pp. 212-225, (2017, Mar.), https://doi.org/10.17743/jaes.2016.0071.

[26] “Advanced sound system for programme production,” ITU-R BS.2051-0, Geneva (2014).

[27] W. Howie et al., “A Three-Dimensional Orchestral Music Recording Technique, Optimized for 22.2 Multichannel Sound,” in *AES Convention 141*, Los Angeles (2016).

[28] “Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems,” ITU-R Recommendation BS.1116-1, International Telecom Union: Geneva, Switzerland (1997).

[29] “Loudness Normalization and Permitted Maximum Levels of Audio Signals,” EBU R-128, European Broadcasting Union: Geneva (2014, June).

[30] “List Randomizer,” [Online]. Available: https://www.random.org/lists/. [Accessed 04-01-2018].

[31] B. Theiss and M. O. J. Hawksford, “Phantom Source Perception in 24bit @ 96kHz Digital Audio,” in *AES Convention 103*, New York (1997).

[32] S. Yoshikawa et al., “Does High Frequency Sampling Improve Perceptual Time-Axis of Digital Audio Signal,” in *AES Convention 103*, New York (1997).

[33] M. Mizumachi et al., “Subjective Evaluation of High Resolution Audio Under In-car Listening Environments,” in *AES Convention 138*, Warsaw (2015).

[34] B. Moore et al., “Frequency discrimination of complex tones; assessing the role of component resolvability and temporal fine structure,” *J. Acoust. Soc. Am.*, vol. 119, pp. 480-490 (2006), https://doi.org/10.1121/1.2139070.

[35] B. Moore and G. Moore, “Discrimination of the fundamental frequency of complex tones with fixed and shifting spectral envelopes by normally hearing and hearing-impaired subjects,” *Hearing Research*, vol. 182, pp. 153-163 (2003), https://doi.org/10.1016/S0378-5955(03)00191-6.

[36] N. Kanetada et al., “Evaluation of Sound Quality of High Resolution Audio,” in *Proceedings of the 1st IEEE/IIAE International Conference on Intelligent Systems and Image Processing* (2013), https://doi.org/10.12792/icisip2013.014.

[37] A. Pras and C. Guastavino, “Sampling rate discrimination: 44.1 kHz vs. 88.2 kHz,” in *AES Convention 128*, London (2010).

[38] T. Oohashi et al., “Multidisciplinary study on the hypersonic effect,” *International Congress Series 1226*, pp. 27-42 (2002).

[39] R. Kuribayashi et al., “High-resolution music with inaudible high-frequency components produces a lagged effect on human electroencephalographic activities,” *NeuroReport*, vol. 29, pp. 651-655 (2014), https://doi.org/10.1097/WNR.0000000000000151.