Audio Engineering Society

Conference Paper

Presented at the International Conference on Spatial Reproduction – Aesthetics and Science, 2018 August 7–9, Tokyo, Japan

This paper was peer-reviewed as a complete manuscript for presentation at this conference. This paper is available in the AES E-Library (http://www.aes.org/e-lib) all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Subjective and objective evaluation of 9ch three-dimensional acoustic music recording techniques

Will Howie[1, 2], Denis Martin[1, 2], David H. Benson[1, 2], Jack Kelly[1, 2], and Richard King[1, 2]

[1] *Graduate Program in Sound Recording, McGill University, 555 Sherbrooke St. West, Montreal, QC, H3A 1E3, Canada*

[2] *Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT), 527 Sherbrooke St. West, Montreal, QC, H3A 1E3, Canada*

Correspondence should be addressed to Will Howie (wghowie@gmail.com)

ABSTRACT

A study was undertaken to compare four different 9-channel three-dimensional acoustic music recording techniques, all optimized for capturing a solo piano. The four techniques range in design philosophy: spaced, near-coincident, and coincident. Results of a subjective listening test showed the two spaced techniques as being equally highly rated for the subjective attributes "naturalness of sound scene", "naturalness of timbre", and "sound source image size". Listeners rated the coincident technique significantly lower than all other techniques under investigation for all perceptual attributes. Binaural recordings of the stimuli were analyzed using several different objective measures, some of which were found to be good predictors for the perceptual attributes "envelopment" and "sound source image size".

1 Introduction

Acoustic music recording for loudspeaker-based three-dimensional audio

As new loudspeaker-based 3D audio formats [1 – 4] have been introduced and become more commonplace in audio production, various researchers and recording engineers have developed acoustic music recording techniques optimized for said playback formats. Most techniques discussed in the literature tend to be optimized for 9-channel, 4+5+0 [3] reproduction formats [5 – 9]. Considerable work has also gone into developing sound capture techniques for Japan Broadcasting Corporation's 22.2 Multichannel Sound (22.2 or 9+10+3) [10 – 14], [4], most of which can easily be scaled to 3D formats with fewer channels. As with traditional stereo and 5.1 surround sound, acoustic music recording techniques for 3D audio can often be divided into one of three categories: spaced, near-coincident, or coincident.

Spaced recording techniques

Spaced 3D microphone techniques aim to capture and reproduce spatial sound information through time of arrival differences between microphone signals. A linear, one-to-one microphone signal to loudspeaker relationship is typically maintained. Another common feature in many proposed techniques is an emphasis on distant spacing between rear and height microphones to prioritize decorrelation between microphone signals. Several authors have commented on the importance of minimizing direct sound capture in height channel signals in order to ensure instrument or ensemble image stability at ear level [5], [8], [9], [13] while maintaining a traditional "concert" perspective.

King et al. [9] suggest the use of acoustic pressure equalizers when using omnidirectional microphones for rear and height channels, ensuring increased microphone directivity and channel separation at frequencies above 1kHz, but maintaining an efficient

capture of low frequency information [9], [15]. Bowles [5], on the other hand, suggests that to minimize direct sound in the height channels, hypercardioid microphones should be used, angled such that the nulls of the microphones are facing the soundstage. Hamasaki and Van Baelen [16] describe a similar approach, suggesting upward facing hypercardioid microphones for height channel capture, placed very high above the "main" microphones.

Near-coincident recording techniques

Near-coincident 3D recording techniques use smaller spacing between microphone capsules: typically less then 1 m. The sound scene is captured using a combination of timing and level differences between microphone signals. Michael Williams has written extensively on his "3D Multiformat Microphone Array" [17 – 19], which is designed to prioritize localization of direct sounds in the horizontal and vertical plane, while minimizing interaction effects between the two loudspeaker layers [17]. Theile and Wittek have expanded their "OCT Surround" (Optimized Cardioid Triangle Surround) [20] technique for 3D audio, adding four upward facing hypercardioid microphones, placed 1 m above the main layer microphones [8]. Main layer microphones use a mixture of cardioid (Centre, Rear Left, and Rear Right) and hypercaridoid (Left and Right) polar patterns. Wittek and Theile also recently introduced "3D ORTF" [30], an ambience capture system for 3D audio and VR applications that, as the name implies, is comprised of four closely spaced ORTF pairs. For height channel microphones spaced less than 2 m above main layer microphones, Lee [38] and Wallis [39] both suggest the use of directional polar patterns to minimize vertical inter-channel crosstalk, set at angles of at least 90° or 105° respectively.

Coincident recording techniques

Most publications addressing coincident microphone techniques for three-dimensional acoustic music recording have focused on ambisonics-based recording techniques [6], [7], [21]. A notable exception is Martin et al.'s single instrument capture arrays for 3D audio [22]. "Double-XY", for example, combines a traditional XY cardioid pair with a 2nd, vertically oriented coincident cardioid pair. Though not designed to capture a complete sound scene, as there is no information captured for the rear channels, Martin et al. have shown these techniques create sonic images with well-defined horizontal *and* vertical extent, which is highly valuable for achieving realistic or hyper-realistic recreations of acoustic instruments [23].

Geluso [6] and Ryaboy [7] both discuss a native B-format capture approach for acoustic music recording: "Double MS+Z." As the name implies, a vertically oriented coincident bi-directional microphone is added to the Double MS [20] array. For ease of coincident spacing, both authors suggest the use of a Sennheiser MHK800 Twin microphone to capture both front and rear M components [6], [7]. As with coincident surround techniques, 1st order A-format capture systems such as a Soundfield microphone, or higher order spherical microphones such as the Eigenmike can also be used for coincident 3D sound capture [14], [21], [24].

Comparative evaluations of three-dimensional music recording techniques

Surprisingly few of the above-mentioned 3D music recording techniques have been subjected to formal comparative evaluations, either through subjective or objective means. Ryaboy [7] investigated perceptual differences between two recording techniques: Double MS+Z, and "Twins Square", a mixed spaced/coincident technique. Results of a double-blind listening test were reported as showing significant differences between the two techniques regarding "localization" (horizontal and vertical) and "perceived room size".

Howie et al. [14] compared three different orchestral music capture techniques optimized for reproduction over 22.2. Results showed listeners rated a spaced recording technique of the authors' own design [13] and a spaced technique created by Hamasaki and his co-authors [13], [16] equally and quite highly for the subjective attributes "clarity", "scene depth", "naturalness", "environmental envelopment", and "quality of orchestral image". A spherical higher order ambisonics (HOA) capture system was also included in the study, which listeners rated quite low for all subjective attributes under investigation.

Objective measures for multichannel audio evaluation

Several authors have investigated objective measures for multichannel audio that may act as predictors of subjective listener evaluations for spatial sound attributes [25 – 28]. Interaural cross correlation (IACC) has been used in concert hall acoustics as an objective measurement for aspects of spatial impression, and is meant to quantify the dissimilarity of signals at the two ears [29]. Investigating the impact of 3D audio on "envelopment", Power et al. found a strong negative correlation between mean listener "envelopment" scores for the various reproduction systems under investigation and measured IACC values for binaural dummy-head recordings made of the testing stimuli [25]. Choisel and Wickelmaier [27] reported a strong negative correlation between $IACC_f$ and perceived "spaciousness", comparing $IACC_f$ measurements of binaural recordings of the stimuli with listener evaluations. Masson and Rumsey [28] found perceptually grouped IACC measurements (PGIACC) on experimental stimuli correlated highly with listener subjective data. George et al. [26] showed that the measures "Area of sound distribution" and "extent of sound distribution" were successful in predicting listener scores for "envelopment" in surround sound recordings. Both measures were designed to "model the extent of sound distribution".

Research goals

This study aims to investigate possible perceptual differences between several currently proposed spaced, near-coincident, and coincident 3D acoustic music recording techniques, using both subjective and objective means. Correlation between subjective and objective measures will also be investigated.

2 Recording techniques under investigation

Four techniques were selected for investigation, drawn from the current literature on three-dimensional acoustic music recording. For this study, all techniques were optimized for reproduction using ITU 4+5+0 [3]. This 9-channel 3D audio standard calls for five loudspeakers at ear level at 0º, ±30º, and ±110º, and two pairs of elevated height channels at ±30º and ±110º. A "concert" perspective sound scene (i.e., direct sound in front, ambience above and surrounding) was maintained with all techniques.

Techniques 1 and 2 are both spaced techniques, based on designs described by Howie et al. [13] and King et al. [9]. Microphone type and placement strategy for the Left, Centre, and Right channels are identical for both techniques: spaced omnidirectional microphones. Both techniques also utilize large spacing between rear and height channel microphones. This ensures a high degree of decorrelation between signals that contain primarily ambient information. Technique 1, a reduction of a recording method originally optimized for 22.2, uses wide-cardioid microphones for the rear channels, and cardioid microphones for the height channels. This ensures a lack of direct sound information being captured by the ambience microphones, resulting in a more stable frontal sound image [13]. Technique 2 uses omni-directional microphones for rear and height channels, all fitted with acoustic pressure equalizers, as discussed in Section 1 and [9].

Technique 3, OCT 9, is described in detail by Theile and Wittek [8]. The technique is designed to prioritize clear directional imaging and a high degree of channel separation. Adequate decorrelation between microphone signals ensures that a "natural spatial impression" is reproduced [8]. See Section 1 and [8] for more detailed information.

Technique 4, Geluso's "Double MS+Z", is described in detail in [6], [7] and Section 1. Like other native B-format capture systems, the microphone signals require matrixing or post-processing to achieve the correct decoding for a given reproduction system. This is in contrast to Techniques 1–3, all of which maintain a linear relationship between microphone signals and respective loudspeaker channels.

3 Production of stimuli

Stimuli recording

All four recording techniques under investigation were setup for simultaneous recording of a solo piano. The recording venue was the Music Multimedia Room (MMR) at McGill University, a large scoring stage measuring 24.4 m x 18.3 m x 17 m. At the time

of recording, no acoustical treatment was installed in the room, resulting in an RT60 of approximately 4.5 s. For all techniques, microphone choice and placement were based on the recommendations of the technique's creators (Table 1, Figure 1).

Recording Channel	Microphone
Tech 1 + 2 L	Schoeps MK 2H w/APE
Tech 1 + 2 C	Schoeps MK 2H w/APE
Tech 1 + 2 R	Schoeps MK 2H w/APE
Tech 1 Rear L	Schoeps MK 21
Tech 1 Rear R	Schoeps MK 21
Tech 1 Top Front L	Schoeps MK 4
Tech 1 Top Front R	Schoeps MK 4
Tech 1 Top Rear L	Schoeps MK 4
Tech 1 Top Rear R	Schoeps MK 4
Tech 2 Rear L	DPA 4006 w/APE
Tech 2 Rear R	DPA 4006 w/APE
Tech 2 Top Front L	DPA 4006 w/APE
Tech 2 Top Front R	DPA 4006 w/APE
Tech 2 Top Rear L	DPA 4006 w/APE
Tech 2 Top Rear R	DPA 4006 w/APE
Tech 3 L	Schoeps MK 41
Tech 3 C	Schoeps MK 4
Tech 3 R	Schoeps MK 41
Tech 3 Rear L	Schoeps MK 4
Tech 3 Rear R	Schoeps MK 4
Tech 3 Top Front L	Sennheizer MKH 8050
Tech 3 Top Front R	Sennheizer MKH 8050
Tech 3 Top Rear L	Sennheizer MKH 50
Tech 3 Top Rear R	Sennheizer MKH 50
Tech 4 M	Senn. MKH 800 Twin
Tech 4 S (Horizontal)	Senn. MKH 800 P48
Tech 4 S (Vertical)	Senn. MKH 800 P48

Table 1: Microphones for all techniques. "APE" = acoustic pressure equalizer

Spacing between Technique 3's microphones were exactly as prescribed in [8]. Microphone placement for all techniques (Figure 1) was optimized by a team of two professional recording engineers, both of whom had previous experience recording 3D audio. Techniques 1 and 2 shared the same microphones for the Left, Centre, and Right channels. The remaining channels for Techniques 1 and 2 used different microphones (Table 1) that shared the same placement and capsule angles (Figure 1).

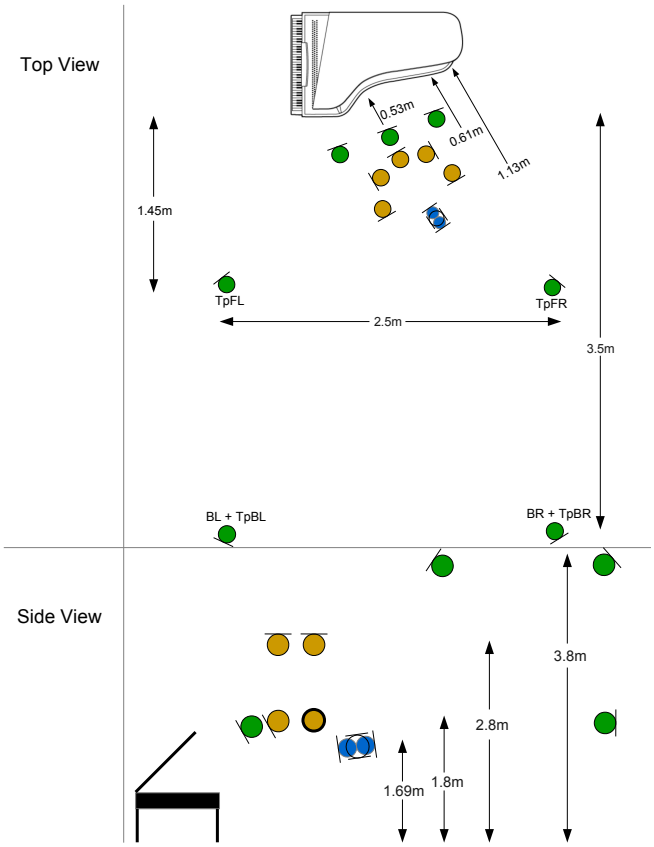

Figure 1: Overhead and side view of all microphone techniques setup in MMR. Green = Tech 1 and 2, Yellow = Tech 3, Blue = Tech 4. Spacing between Tech 1 and 2 Left and Right microphones is ≈ 1.4m.

All microphones were routed to a Sony SIU-100 System Interface Unit, using the internal microphone preamplifiers and analog-to-digital converters. Recordings were made to a Pyramix workstation at 96kHz/24bit resolution. Monitoring took place in McGill University's Studio 22. The studio 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 (9+10+3) and 4+5+0 [3]. The room's dimension ratios and reverb time fulfill ITU-R BS.1116 [31] requirements. Matrixed 9.1 monitoring of Technique 4 was made possible using Pyramix's internal mixer, following [7]'s channel matrixing scheme.

Stimuli mixing and level matching

A 25 s excerpt of J. S. Bach's "Variation 13" from the *Goldberg Variations* was chosen as stimulus. The passage covers a wide range of the keyboard, maintains a fairly even dynamic envelope, and was felt to contain enough "space" within the performance for listeners to perceive the ambient sound field with relative ease. For this passage, each technique was balanced by a team of two professional recording engineers, both of whom have over ten years' experience recording and mixing for multichannel audio. Careful attention was given to maintaining a similar balance of direct to reverberant sound for each technique. No filtering of any kind was applied to the microphone signals. This resulted in four stimuli, one for each technique. Playback of each 9-channel stimulus was then recorded using a Neumann KU-100 dummy head microphone placed at the listening position in Studio 22, at head height for a typical listener. Integrated loudness measurements were then performed on the resultant stereo binaural recordings using a professional software loudness meter calibrated to the EBU+9 scale. Global gain changes were then applied to each 9-channel stimuli as necessary, and the measurement procedure repeated until all binaural stimuli recordings were found to have integrated loudness measures within 0.5 LUFS of each other.

4 Subjective evaluation of stimuli

Design and implementation of listening test

A double-blind listening test was designed to identify perceptual differences between the four recording techniques. Four perceptual attributes were chosen for investigation: "envelopment", "naturalness of sound scene", "naturalness of timbre", and "sound source image size". These attributes were arrived at by a panel of four professional recording engineer/audio researchers who were asked to compare the four techniques in an informal listening session, and determine the sound attributes that best quantified perceived differences between the techniques. The four chosen attributes were agreed upon by the panel as being the most salient.

The listening test was implemented using Cycling 74's Max/MSP software. Subjects were seated in Studio 22's listening position, were explained the testing conditions, and given time to familiarize themselves with the testing interface and stimuli. Definitions of perceptual attributes were provided both verbally and in written form (see: Appendix A).

For each trial, subjects were asked to evaluate stimuli labelled "A", "B", "C", and "D" for a given attribute, using a set of continuous sliders (0-100). Anchor words were provided at the extremes of each slider. Since absolute anchors were not given at intervals along the scales, these measurements are relative and not absolute. Subjects were encouraged to use the full range of the scale for each trial. Subjects could switch between playback of A, B, C, and D stimuli or stop the audio at any point. Playback of stimuli was continuously looped and time synced to ensure seamless switching. The test was administered in blocks of three trials per attribute, for a total of 12 trials. This was done to allow subjects to focus on one perceptual attribute at a time. For each trial, stimulus assignments to A, B, C, and D, as well as the order of trial blocks, were randomized within the testing program. Subjects were instructed to set a comfortable listening level before completing the first trial, and then leave the level unchanged for the remainder of the test. At the test's midway point was an enforced rest period of 1 minute. Subjects took an average of 20 minutes to complete the test, after which they completed a short demographic survey.

5 Objective signal features

Three sets of objective features were calculated on binaural dummy-head recordings of the stimuli. The recordings were made as per Section 3.

IACC features

The first set of calculated features were related to the interaural cross-correlation coefficient (IACC). "Classical" IACC is defined as the maximum of the cross-correlation between two signals, measured at lag values between -1 and 1 ms. $IACC_f$ is a refinement of this feature that includes a pre-processing step designed to mimic envelope extraction in the human auditory system. The pre-processing step includes half-wave rectification and low-pass filtering, which has been shown to improve the measure's perceptual validity [27]. Both features were implemented as per Choisel and Wickelmaier [27], and were calculated

on signal windows 50 ms in duration and with a 25 ms hop size. The final reported IACC and $IACC_f$ values for each stimulus are the mean values over all signal windows.

Binaural model features

The next set of features were derived from a binaural model designed to predict room acoustic attributes. These features, named P_{REV}, P_{CLAR}, P_{ASW}, and P_{LEV}, have been shown to correlate with subjective assessments of "reverberance", "clarity", "apparent source width" and "listener envelopment", respectively. Details about the model are given in Schuitman et al. [33]. To calculate the features, a closed source implementation of the model written by the model's creator was used.

Monaural spectral features

The final set of features were designed to characterize the signals' monaural frequency content. These features were the Spectral Centroid, Spectral Crest factor, Spectral Flatness, Spectral Kurtosis, Spectral Skew, Spectral Spread, and Spectral Variation. All were calculated using the open source Timbre Toolbox [34].

To create frequency domain representations of the stimuli, the Timbre Toolbox's "ERBfft" setting was used. This resulted in spectral analysis bins with a perceptually informed frequency spacing. All other analysis parameters were left at their default values. The features were all calculated on sliding windows with a 5.8 ms hop size; the reported values are the medians over all windows. Prior to this analysis, the binaural stimuli were summed to mono.

6 Results: Subjective Evaluation

Listening test participants

13 subjects participated in the listening test, all of whom were either students or faculty within the Graduate Program in Sound Recording at McGill University. All subjects had previous experience performing triad or pairwise comparison-style listening tests or ear training activities. All reported having normal hearing. Other pertinent demographic data is summarized in Table 2.

Attribute ratings

Since the attribute ratings were relative and not absolute, each participant's data was normalized for mean and standard deviation. The purpose of this is to normalize each individual participant's use of the rating scale. Z-scores were computed for each participant, within each attribute, similar to [32]. The attribute rating results averaged over all participants are visualized in Figure 2.

Age (years)	18-25 (28.6%)	25-32 (28.6%)	33-39 (28.6%)	51+ (14.3%)
Musical Training (years)	4-6 (14.3%)	6-10 (14.3%)	10+ (71.4%)	
Audio Experience (years)	0-4 (35.7%)	4-10 (35.7%)	10+ (28.6%)	
3D Audio Listening Experience	Yes (85.7%)	No (14.3%)		

Table 2. Data describing the previous experience and training of the listening test participants.

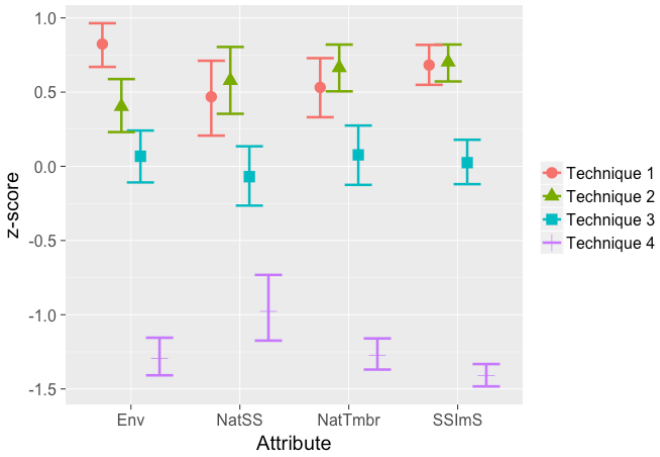

Figure 2: Attribute ratings averaged across all participants for each recording technique. Abbreviations: "Env" = envelopment, "NatSS" = naturalness of sound scene, "NatTmbr" = naturalness of timbre, "SSImS" = sound source image size.

A one-way ANOVA and subsequent Bonferroni adjusted pair-wise t-tests were performed on the ratings for each attribute. These results are summarized in Tables 3–7. The one-way ANOVA shows a main effect of "array" for all attribute ratings with very high significance. Pair-wise t-tests show a

similar pattern of ratings for each attribute. Technique 4 was rated significantly lower than all other techniques for each attribute. Technique 3 was rated significantly lower than Techniques 1 and 2 for most attributes. The exceptions to this are Technique 3 vs 2 in "envelopment", and Technique 3 vs Technique 1 in "naturalness of timbre". There is no significant difference between Techniques 1 and 2 for most attributes, the exception being "envelopment", where Technique 1 was rated significantly higher than Technique 2.

Attribute	**Tech. 1 Mean**	**Tech. 2 Mean**	**Tech. 3 Mean**	**Tech 4 Mean**	**F (df)**	***p***
Env	.82	.40	.067	-1.3	55 (3,39)	<.001
NatSS	.47	.58	-0.070	-.98	13 (3,39)	<.001
NatTmbr	.53	.66	.077	-1.3	51 (3, 39)	<.001
SSImS	.68	.70	.025	-1.4	120 (3, 39)	<.001

Table 3. One-way ANOVA results for each attribute.

	Tech. 1	Tech. 2	Tech. 3
Tech. 2	**.021**	---	---
Tech. 3	**<.001**	.335	---
Tech. 4	**<.001**	**<.001**	**<.001**

Table 4. Envelopment: pairwise t-test results (*p*). Statistically significant results displayed in **bold**.

	Tech. 1	Tech. 2	Tech. 3
Tech. 2	1	---	---
Tech. 3	**.0179**	**.0077**	---
Tech. 4	**<.001**	**<.001**	**<.001**

Table 5. Naturalness of Sound Scene: pairwise t-test results (*p*). Statistically significant results displayed in **bold**.

	Tech. 1	Tech. 2	Tech. 3
Tech. 2	1	---	---
Tech. 3	.144	**<.001**	---
Tech. 4	**<.001**	**<.001**	**<.001**

Table 6. Naturalness of Timbre: pairwise t-test results (*p*). Statistically significant results displayed in **bold**.

	Tech. 1	Tech. 2	Tech. 3
Tech. 2	1	---	---
Tech. 3	**<.001**	**<.001**	---
Tech. 4	**<.001**	**<.001**	**<.001**

Table 7. Sound Source Image Size: pairwise t-test results (*p*). Statistically significant results displayed in **bold**.

Additional analysis

There were no major effects of 3D audio listening experience, years of musical training, years of music production experience, or subject age on the attribute ratings; all techniques appeared in the same rank order as when all the data was pooled together. Subjects found the listening test moderately difficult, rating the difficulty 2.9/5 on average. There were no clues in post-test listener comments as to the source of this difficulty.

7 Results: Objective Analysis

A linear regression analysis was conducted to investigate relationships between the subjective attribute ratings discussed in the previous section and the objective signal features presented in Section 5. For each subjective attribute, the responses from all subjects were pooled together and regressed over each signal feature. The goodness of fit of each regression model was measured using the squared correlation (R^2). The results are summarized in Table 8.

	Env	**NatSS**	**NatTimbr**	**SSImS**
IACC	.59	.33	.50	**.71**
$IACC_f$	**.63**	.33	.51	**.73**
P_{ASW}	.59	.35	.54	**.74**
P_{CLAR}	.51	.28	.40	.59
P_{LEV}	.58	.31	.47	**.68**
P_{REV}	.51	.26	.38	.57
SpecCent	.52	.31	.54	**.66**
SpecCrest	*.02*	.05	.09	.08
SpecFlat	.31	.19	.25	.39
SpecKurt	*.03*	*.02*	*.02*	*.04*
SpecSkew	.13	.10	.10	.18
SpecSpread	.24	.15	.30	.33
SpecVar	**.67**	.34	.54	**.75**

Table 8: Squared correlations (R^2) between subjective attributes and objective signal features. R^2 values greater than 0.6 are displayed in **bold**. *Italics* indicate combinations for which the regression coefficient was not significant ($p < .01$)

Many features were found to correlate strongly with "sound source image size". Two of these features, $IACC_f$, and Spectral Variation, were also predictive of "envelopment". The strong relationship between $IACC_f$ and both "envelopment" and "sound source image size" was unsurprising as this feature was explicitly designed to predict spatial attributes of

binaural signals. The strong relationship between these two spatial attributes and the monoaural Spectral Variation feature is somewhat more difficult to explain. Also worth noting is P_{ASW}'s strong correlation with "sound source image size".

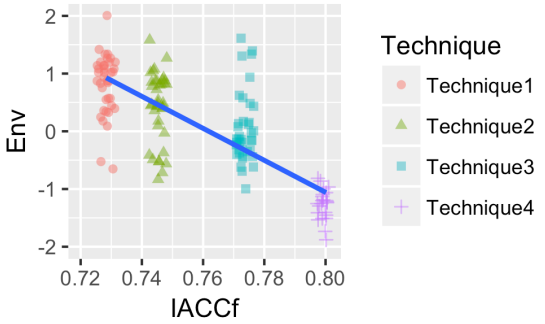

Figure 3: Plot of regression for "envelopment" vs. IACCf. Points have been jittered slightly on the horizontal axis for visibility.

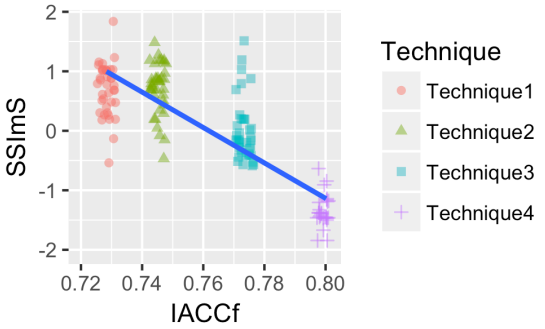

Figure 4: Plot of regression for "sound source image size" vs. IACCf

8 Discussion

Overall performance of recording techniques

Figure 2 shows the two spaced recording techniques were both rated very highly, and very similarly for all subjective attributes under investigation. The near-coincident technique received somewhat more modest ratings across the various attributes, while the coincident technique was rated very low across all attributes. This is remarkably similar to results from a previous study by Kamekawa et al. [35] comparing various two-dimensional orchestral music recording techniques. In that study, techniques combining spaced omnidirectional microphones for frontal sound capture with spaced ambience arrays of either omnidirectional or bi-directional microphones were rated consistently highly for most attributes under investigation. "INA5", a near-coincident technique, was rated lower than those spaced techniques for most attributes, while the coincident "Double MS" technique was rated quite low for all spatial attributes, as well as for preference [35]. Low attribute ratings for a coincident ambisonics-based 3D music recording technique were also seen in [14]. The same study also describes a similar trend found in a number of previous studies comparing two-dimensional music recording techniques.

Envelopment

Figure 2 shows a clear trend towards a linear relationship between spacing of microphones prioritizing ambience and listener "envelopment" ratings. Essentially, the more a technique is optimized towards decorrelation of ambience microphone signals, the greater sense of "envelopment" is perceived by listeners. This supports previous work by Griesinger [36] that suggests decorrelation of the ambient component of recordings across the audible frequency spectrum is necessary for achieving optimal levels of "spaciousness". It is particularly interesting to note that "envelopment" is the one attribute for which Techniques 1 and 2 show a significant difference in ratings. These techniques only differ in terms of the directional characteristics of their respective ambience microphones. While not examined for this study, this suggests that Technique 1's use of directional microphones for ambience capture likely resulted in a higher degree of decorrelation between signals.

Technique 1's rear and height channel signals would naturally contain less direct sound components then those of Technique 2, which uses all omnidirectional microphones. It is possible that this contributed to a greater perceptual separation of direct and reverberant sound components, and thus, a more enveloping sound scene, which is in line with Greisinger's theories on foreground and background auditory streaming [36].

Figure 3 and Table 8 show a relatively strong correlation between $IACC_f$ values and listener

"envelopment" ratings. This agrees with the results of previous studies discussed in Section 1 showing measurements of inter-channel coherence of binaural dummy-head microphone signals as being good predictors of "envelopment".

What is less intuitive is the relatively strong correlation observed in this study between "envelopment" ratings and the feature Spectral Variation. Also known as "Spectral Flux", this feature measures how quickly the power spectrum of a signal changes over time, and has traditionally been used to examine the timbre of audio signals containing speech or musical instrument sounds [37]. Like the other spectral measures used in this study, Spectral Variation examines a mono summation of the binaural signals. It is known that summation of similar audio signals often results in spectral notches or "comb filtering". The relationship between the Spectral Variation and $IACC_f$ features may indicate that when signals with a higher degree of decorrelation are summed, the frequencies at which spectral notches occur will have more variation over time. This would decrease the similarity in frequency spectra between successive windows, and thus increase the Spectral Variation.

Sound source image size

Results shown in Figure 2 suggest that for distances from the sound source typical of classical music recording, frontal sound microphone arrays relying primarily on timing differences between signals yield wider sound source images. This is not necessarily the "best" or "most desirable" result: recording techniques should be chosen that yield an instrument or ensemble size that best corresponds to the aesthetic goals of the recording engineer, producer, and artist.

Schuitman et al.'s P_{ASW} measure uses a model of the auditory system to determine which components of the input belong to the source stream, allowing it to evaluate the fluctuations in interaural time differences for direct sound only [33]. For this study, this feature appears to be the best fit for predicting listener perception of "sound source image size". Table 8 also shows strong correlations between listener scores for "sound source image size" and the objective measures IACC and $IACC_f$. This is not surprising, given that IACC is often associated with "apparent source width" in concert hall acoustics [29]. In addition, Mason and Rumsey found IACC-based measures strongly correlated to listener judgements of "sound source width [28]".

Naturalness Attributes

For this study, the concept of "naturalness" was divided into two areas: sound scene and timbre. Figure 2 and Table 5 show that listeners in this study found the sound scenes reproduced by spaced recording techniques to be the most "natural". This is consistent with previous results from [14]. For "naturalness of timbre", a definitive result is less clear. Among Techniques 1–3, a small trend can be observed in terms of perceived naturalness of timbre and the ability of each technique to efficiently capture low frequency content. As microphone directivity increases, low frequency roll-off of direct sound also increases, as a function of distance from the sound source. For this study, listeners likely equated "naturalness of timbre" with "flatness" of spectrum, though not to such a strong degree as to result in significant differences between Techniques 1 and 2, or Techniques 1 and 3. Surprisingly, none of the objective measures applied to the stimuli show a particularly strong correlation with either "naturalness" attribute.

Subjective attributes under investigation

As described in Section 4, the subjective attributes used for this study were chosen to emphasize areas of clear difference between the four techniques. It is possible that for other common subjective attributes, such as "localization", "brightness", etc., the general ranking trend in the results may have been different.

9 Acknowledgments

This work was supported by the Social Sciences and Humanities Research Council (SSHRC) and The Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT).

References

[1] K. Hamasaki and K. Hiyama, "Development of a 22.2 Multichannel Sound System," *Broadcast Technology*, vol. 25, pp. 9-13, Winter 2006.

[2] ATSC Standard: A/342 Part 1, Audio Common Elements," Advanced Television Systems Committee, Washington, DC, 24 January 2017.

[3] "Advanced sound system for programme production," ITU-R BS.2051-0, Geneva, 2014.

[4] "Multichannel sound technology in home and broadcasting applications," ITU-R BS.2159-7, Geneva, 2015.

[5] D. Bowles, "A Microphone Array for Recording in Surround-Sound with Height Channels," in *AES Convention 139*, New York, 2015.

[6] P. Geluso, "Capturing Height: The Addition of Z Microphones to Stereo and Surround Microphone Arrays," in *AES Convention 132*, Budapest, 2012.

[7] A. Ryaboy, "Exploring 3D: A subjective evaluation of surround microphone arrays catered for Auro-3D reproduction," in *AES Convention 139*, New York, 2015.

[8] G. Theile and H. Wittek, "Principals in Surround Recording with Height (v2.01)," in *AES Convention 130*, London, 2011.

[9] R. King et al., "A Survey of Suggested Techniques for Height Channel Capture in Multi-channel Recording," in *AES Convention 140*, Paris, 2016.

[10] K. Hamasaki et al., "Advanced multichannel audio systems with Superior Impressions of Presence and Reality," in *AES Convention 116*, Berlin, 2004.

[11] T. Hinata et al., "Live Production of 22.2 Multichannel Sound for Sports Programs," in *AES 40th International Conference*, Tokyo, 2010.

[12] K. Irie and T. Miura, "The production of 'The Last Launch of the Space Shuttle" by Super Hi-Vision TV," in *Broadcast Engineering Conference, NAB show 2012*, Los Vegas, 2012.

[13] W. Howie et al., "A Three-Dimensional Orchestral Music Recording Technique, Optimized for 22.2 Multichannel Sound," in *AES Convention 141*, Los Angeles, USA, 2016.

[14] W. Howie et al., "Subjective Evaluation of Orchestral Music Recording Techniques for Three-Dimensional Audio," in *AES Convention 142*, Berlin, 2017.

[15] W. Woszczyk, "Acoustic Pressure Equalizers," *Pro Audio Forum,* pp. 1-24, 1990.

[16] K. Hamasaki and W. Van Baelen, "Natural Sound Recording of an Orchestra with Three-Dimensional Sound," in *AES Convention 138*, Warsaw, 2015.

[17] M. Williams, "Microphone Array Design for localization with elevation cues," in *AES Convention 132*, Budapest, 2012.

[18] M. Williams, "The Psychoacoustic Testing of a 3D Multiformat Microphone Array Design, and the Basic Isosceles Triangle Structure of the Array and the Loudspeaker Reproduction Configuration," in *AES Convention 134*, Rome, 2013.

[19] M. Williams, "Microphone Array Design applied to Complete Hemispherical Sound Reproduction – from Integral 3D to Comfort 3D," in *AES Convention 140*, Paris, 2016

[20] T. Kamawaka, "An Explanation of Various Surround Microphone Techniques," Sanken, [Online]. Available: http://www.sanken-mic.com/en/qanda/index.cfm/18.56. [Accessed 26 March 2017]

[21] Bates et al., "Comparing Ambisonic Microphones – Part 2," in *AES Convention 142*, Berlin, 2017.

[22] B. Martin et al., "Microphone Arrays for Vertical Imaging and Three-Dimensional Capture of Acoustic Instruments," in *AES Conference on Sound Field Control*, Guilford, 2016.

[23] 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.

[24] M. Ikeda et al., "New Recording Application for Software Defined Media," in *AES Convention 141*, Los Angeles, 2016.

[25] P. Power et al., "Investigation into the Impact of 3D Surround Systems on Envelopment," in *AES 137*, Los Angeles, 2014.

[26] S. George et al., "Development and Validation of an Unintrusive Model for Predicting the Sensation of Envelopment Arising from Surround Sound Recordings," *J. Audio Eng. Soc.*, vol. 58, no. 12, 2010, pp. 1013-1031.

[27] S. Choisel and F. Wickelmaier, "Relating auditory attributes of multichannel sound to preference and to physical parameters," in *AES Convention 120*, Paris, 2006.

[28] R. Masson and F. Rumsey, "A comparison of objective measurements for predicting selected subjective spatial attributes," in *AES Convention 112*, 2002.

[29] T. D. Rossing, "Acoustics in Halls for Speech and Music," *Springer Handbook of Acoustics*. Springer: New York, 2007.

[30] H. Wittek and G. Theile, "Development and application of a stereophonic multichannel recording technique for 3D Audio and VR," in *AES Convention 143*, New York, 2017.

[31] "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).

[32] J. Berg and F. Rumsey, "Verification and correlation of attributes used for describing the spatial quality of reproduced sound," in *AES 19th International Conference*, Schloss Elmau, 2001.

[33] J. van Dorp Schuitman and D. de Vries, "Deriving content-specific measures of room acoustic perception using a binaural, nonlinear auditory model," *J. Acoust. Soc. Am.* vol. 133, no. 3, 2013, pp. 1572-1585.

[34] G. Peeters et al., "The Timbre Toolbox: Extracting audio descriptors from musical signals," *J. Acoust. Soc. Am.*, vol. 130, no. 5, 2011, pp. 2902-2916.

[35] Kamekawa et al., "Correspondence Relationship between Physical Factors and Psychological Impressions of Microphone Arrays for Orchestra Recording," in *AES Convention 123*, New York, 2007.

[36] D. Griesinger, "Spatial Impression and Envelopment in Small Rooms," in *AES Convention 103*, New York, 1997.

[37] F. Alías et al., "A Review of Physical and Perceptual Feature Extraction Techniques for Speech, Music and Environmental Sounds," Appl. Sci., vol. 6, no. 143, 2016.

[38] H. Lee, "The Relationship between Interchannel Time and Level Differences in Vertical Sound Localisation and Masking," in *AES Convention 131*, New York, 2011.

[39] R. Wallis and H. Lee, "The Reduction of Vertical Interchannel Crosstalk: The Analysis of Localisation Thresholds for Natural Sound Sources," *Applied Sciences,* vol. 7, no. 278, 2017.

Appendix A: Definitions of Subjective Attributes

Sound Source Size: How large is the horizontal and vertical extent of the sound source's sonic image. (Sound source is a piano.)

Envelopment: The sense of immersion and involvement in the sound field. Amount that the listener feels inside/enveloped by the sound image.

Naturalness of Sound Scene: Consider the total sound scene: direct and diffuse sound, from all directions, and their relation to one another. How natural is the listening experience, as opposed to feeling artificial or removed from your frame of reference for such a sound scene?

Naturalness of Timbre: Consider the total sound scene: direct and diffuse sound. How natural is the overall timbre of the reproduced sound?