Journal: J. Audio Eng. Soc.

Article doi: JAES-D-21-00029

Article title: Parametric Evaluation of Ensemble Vocal Performance Using an Immersive Network Music Performance Audio System

First Author: PATRICK CAIRNS

Corr. Author: PATRICK CAIRNS; email: pc1095@york.ac.uk

STANDARD AUTHOR QUERIES

1. Please check that the author names and affiliations appear correctly and are suitable for publication.
2. Please confirm the corresponding author email address is correct and has not changed since submission.
3. Please check that all figures and/or tables are correctly placed, have no missing sections, and correspond to their associated legends.
   Note that Figures may publish online in full color; however, they will be converted to grayscale for print. Please review the figures and associated captions and citations to ensure that figures will be understandable after being converted to black and white.

AUTHOR QUERIES - TO BE ANSWERED BY THE CORRESPONDING AUTHOR

The following queries have arisen during the typesetting of your manuscript. Please answer these queries by marking the required corrections at the appropriate point in the text.

AQ1	Please check the full article carefully and split any paragraphs over the 180-word limit into two paragraphs or more as appropriate.	
AQ2	Minor syntax adjustments have been made throughout. Please read carefully to ensure intended meaning has been retained.	
AQ3	Please confirm all product, software, and other proper names throughout the article.	
AQ4	Please expand "SIRs" in the sentence starting with "This is accomplished..."	
AQ5	Please define all abbreviations in all figures and tables throughout the article.	
AQ6	Please confirm changes in all tables throughout the article.	
AQ7	The running title has been edited for length. Please review the running title and confirm whether it is acceptable or provide an alternate short title.	
AQ8	Please expand "FEC" in the sentence starting with "Audio streaming parameters..."	
AQ9	Please confirm "Round Time Trip" (instead of "Round Trip Time") in Sec. 3.1.	
AQ10	Please expand "BPM" in the sentence starting with "Mean tempo slope describes..."	
AQ11	Please confirm that the links provided for Footnotes 4 and 5 are correct.	
AQ12	Many updates have been made to the references throughout the verification process; please confirm these are correct.	
AQ13	Please provide a page range/paper number for [18].	
AQ14	Please confirm the information in [22] is correct and provide a page range/paper number for that reference if possible.	

P. Cairns, H. Daffern, and G. Kearney, "Parametric Evaluation of Ensemble Vocal Performance Using an Immersive Network Music Performance Audio System"
J. Audio Eng. Soc., vol. X, no. X, pp. 1–10, (2021 X).
DOI: https://doi.org/10.17743/jaes.2021.0040

Parametric Evaluation of Ensemble Vocal Performance Using an Immersive Network Music Performance Audio System

PATRICK CAIRNS,* *AES Member*, **HELENA DAFFERN, AND GAVIN KEARNEY,** *AES Member*
(pc1095@york.ac.uk) (helena.daffern@york.ac.uk) (gavin.kearney@york.ac.uk)

AudioLab, University of York, York, United Kingdom

This paper describes an immersive audio Network Music Performance (NMP) system designed for group singing. A prototype of this design (audio only) was deployed to ten singers across Europe, who participated in a duet vocal performance study, operating the system from their home networks. Parametric evaluation of these vocal performances was conducted in order to provide characterization of musical interactivity between performers and explore the challenges and opportunities presented for immersive audio NMP systems in practical use-case settings. Results demonstrate that it is possible to achieve performance that conforms to expectations of live interactivity and estimate the conditions under which this may be achieved. Significant effect of latency, and in one case virtual room "type," is observed across performances. Informal questionnaire responses present discussion of the potential for virtual acoustics and latency to impact the perceptual experience of networked performers.

0 INTRODUCTION

Network Music Performance (NMP) technology [1], allows remote musicians to engage in group performance by sharing audio over the internet. It has become commonplace to include visual display in NMP technologies, typically following video conferencing methods. Though current state of the art is dominated by telepresence-motivated NMP designs [2], cutting-edge research has begun to explore the fusion of NMP technology with Virtual Reality (VR) experiences [3]. The term Immersive Network Music Performance (INMP) is presented to describe not just VR NMP applications but as a term that can be extended to Augmented Reality, holographic systems, and other immersive technology.

The factors of auditory immersion in VR experiences are well understood [4]. The introduction of VR visual display in INMP systems can therefore be acknowledged as requiring immersive audio functionality for optimized experiences.

Current work at the AudioLab, University of York, has contributed an audio system design for INMP vocal performance. This system provides simulation of a shared virtual acoustic space, low latency audio transport and rendering processes, and interactive spatial audio delivery. The design is based upon the Vocal Interaction in an Immersive Virtual Acoustic (VIIVA) system for VR singing [5] and has therefore been named "VIIVA-NMP" [6].

In live vocal performance it is well understood that the acoustic environment will have a significant effect on ensemble performance [7]. This is true with respect to both individual perception and objective physical properties, such as measures of performance synchrony [8]. In immersive audio performance experiences, such as INMP, shared virtual acoustic performance spaces may also have an effect on performance [9–11].

Empirical NMP research demonstrates that performance synchrony is significantly affected by latency associated with network transport and audio processing [11]. The Ensemble Performance Threshold (EPT) has been defined as "the level of delay at which effective real-time musical collaboration shifts from possible to impossible" [12]. In this context, "delay" refers to the complete One Way Trip (OWT) latency between one remote sound source (such as voice) and another remote sound receiver (such as ears) [1]. EPT is generally estimated in the range of 25–30 ms [1], although a range of dependencies exist, as highlighted

*To whom correspondence should be addressed E-mail: pc1095@york.ac.uk

in Rottondi et al.'s overview [1]. When OWT is above the EPT, synchronous performance requires the use of latency coping strategies [13]. This makes performance at delays above the EPT a substantially different experience to that of live musical interaction, because of high levels of interface awareness [14].

INMP audio systems present a new context, where implementation of immersive audio processes within latency and bandwidth constraints of "real-world" internet has historically been lacking [15]. In order to contribute a study in this new context, a prototype implementation of the VIIVA-NMP audio system design was deployed to ten remote performers across Europe. These musicians engaged in duet vocal performance from their home networks, and parametric analysis of the performance achieved was conducted in order to provide evaluation of the effect of latency and different virtual acoustic performance spaces in a practical use-case scenario.

1 RELATED WORK

INMP proposes the extension of VR performance experiences to a networked, multi-user context, where performer avatars may allow for embodiment in the performance experience [16]. The replacement of Video over Internet Protocol (VoIP) streaming with avatar rendering metadata and virtual environments allows circumvention of the "catch-22" of telepresence-NMP design: Low-latency VoIP streaming requires large bandwidth (limiting accessibility by typical home users), and low-bandwidth VoIP streaming requires an additional latency associated with video compression (which will raise OWT latency well above the EPT). Haptic metadata associated with avatar rendering, however, may be easily sent synchronously alongside low-latency audio.

Recent research in INMP has provided perceptual evaluation of potential group singing telemedicine applications [17]. This study was conducted on an emulated network using NMP tools parallel to asynchronous VR visual rendering. It was found that the INMP choir with VR display showed strong potential to extend the health and well-being benefits of group singing to remote performers. The system used in this study provides only dry mono audio to performers. As such it can be recognized that potential improvements can be made to the VR experience through the inclusion of immersive audio and evaluation of solutions that provide this.

Previous work in [18] presents a system for the live-streaming of Higher Order Ambisonic signals to remote loudspeaker arrays. This system is used as an interactive installation rather than for real-time music making; however it still presents important ground work in INMP audio systems. Ambisonic loudspeaker arrays, used to place remote performer sounds in real acoustic spaces, however, are inaccessible to the typical home performer. The bandwidth requirements of Higher Order Ambisonic streaming present a further accessibility complication.

Research has also explored the need to place the sounds of remote performers in virtual performance spaces using acoustic simulation methods. Carôt et al. [10] acknowledges that the dry audio used in many NMP experiments may appear "unnatural" to singers. They conducted a study on "artificial reverb" using drums as the performance instrument but found no conclusive results. A study by Farner et al. [9] reports improvements in performance synchrony in virtual acoustic conditions using static measured Binaural Room Impulse Responses when compared to virtual anechoic conditions.

The extent of previous work in the field demonstrates the need to contribute INMP audio system designs that can be implemented for typical remote performers and are suitable for deployment alongside VR visual display. It can also be recognized that there is need to provide original evaluation of any such designs. Virtual acoustics are a required function of INMP audio systems, and previous work shows this is expected to have an influence on the performance experience. In the context of INMP audio system design and evaluation, it is therefore prudent to attempt to understand how different virtual acoustic spaces will impact the performance experience.

2 EXPERIMENT

2.1 VIIVA-NMP Audio System

A prototype implementation of the VIIVA-NMP audio system design using open source tools and resources was used to share audio between performers. JackTrip [19] was used to stream audio between performers using User Datagram Protocol streaming, jitter buffering, and Forward Error Correction packet loss concealment. Auralization of a virtual acoustic performance space is accomplished using convolution with first order Ambisonic impulse responses and virtual Ambisonic binaural decoding [20]. This is accomplished using Kronlachner VST [21, 22] measured SIRs from the Open Air Impulse Response Library [23] and SADIE II Binaural resources [24]. Jack Audio Connection Kit[1] was used to route audio between applications.

Though the VIIVA-NMP audio system design includes provision of 3 Degrees of Freedom using a custom head-tracking design, this feature was not enabled for testing, and static binaural audio was delivered to participants in this study. This decision was made in order to maintain control in this study, as it was not feasible to safely distribute hardware to participants across Europe during the necessary restrictions of the COVID-19 pandemic.

The VIIVA-NMP audio system is designed for implementation alongside VR visual display. Considering that visual cues are known to have an impact in live group singing experiences [25], it can be reasonably assumed that visual cues will also have an impact on INMP experiences. In the context of INMP audio system Wide Area Network deployment, however, absolutely no data as to what this might be is present. As visual stimuli present an unknown variable, it was decided to conduct this study with no visual display. This allows control over testing and provides a

[1] https://jackaudio.org/

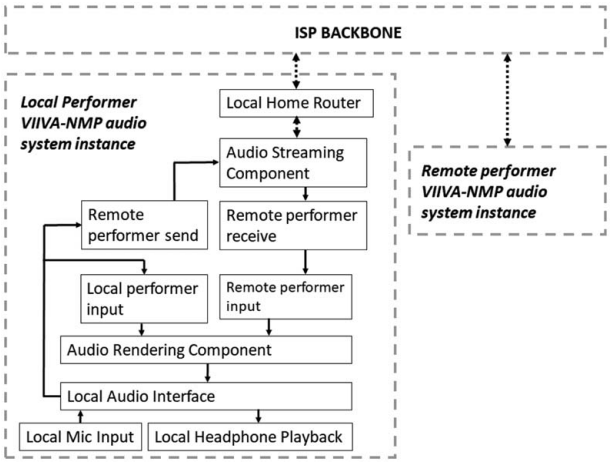

Fig. 1. VIIVA-NMP prototype configuration for testing protocol. AQ5

baseline for future research implementing this audio system design alongside VR visual displays.

2.2 Testing Protocol

Duet pairs of test participants formed connections using the VIIVA-NMP audio system prototype (Fig. 1). A microphone was used to sing into, and headphones were worn for binaural playback of the auralized performance. Each pair attempted a vocal performance task three times in each of the three different virtual acoustic performance spaces. Following completion of three attempts of the performance task in each virtual acoustic room, each participant completed a short questionnaire providing perceptual rating of the performance experience. Audio from each attempt of the performance task was recorded in Reaper.[2]

Ten individuals took part in this study, nine participants through recruitment and the lead researcher for this study who took the place of an absent participant. It was considered extremely low risk that prior knowledge of the experiment would impact the ability of the lead researcher to play music in time. No data from the lead researcher is included in the reporting of the perceptual ratings. These participants all met the preferred recruitment conditions (is under 45 years of age, has a minimum of amateur level singing experience, self-reports unimpaired hearing, and is familiar with the general type of duet performance required in the performance task). Gender is not reported as there is no reason to expect this to affect the ability of participants to perform in time with one another. No participant had significant network music experience prior to participation in this study. During system setup sessions with the lead researcher, each participant was given a brief practice period using the system, such that all participants had some experience prior to testing.

Participants were organized into five duet pairs, with one participant from each pair designated "server" and the other "client." Participant identification was assigned by group, and designation "client" or "server" within group (Table 1). The five duet pairs each contained one participant located in Glasgow, Scotland, in order to provide a common geographic location. The other participant from each pair was positioned at a different location in Europe (Table 2). Glasgow was chosen as a common location as this was considered to represent a typical small city with no exceptional internet infrastructure and therefore provided a good example of typical location for a use-case scenario. Locations across Europe were selected in order to control the latency experienced by participants. It was desired that OWT delays in the range of 20–60 ms were achieved in order to provide investigation at values where empirical NMP research gives reason to expect an effect on performance.

[2]https://www.reaper.fm/userguide.php

Table 1. Acoustic parameters for recording studios, University of York. AQ6

Oct. Band (*Hz*)	T30 (*s*)	EDT (*s*)	C50 (*Db*)
31.25	0.8	0.98	−2.33
62.5	0.57	0.73	5.82
125	0.28	0.34	16.5
250	0.27	0.22	20.03
500	0.21	0.22	23.87
1,000	0.19	0.22	26.0
2,000	0.21	0.22	23.59
4,000	0.22	0.22	22.54
8,000	0.21	0.22	22.92
16,000	0.18	0.22	27.99

A range of additional session configuration parameters were recorded for each test participant (Table 2). Hardware used in testing, namely audio interface, microphone, and headphones, were noted, as well as sample rate and buffer size used. Audio streaming parameters were also recorded, namely bandwidth, jitter buffer length, and FEC redundancy. The level of vocal performance proficiency was also noted for each participant. No duet pairs had sung together prior to testing. AQ8

2.3 Virtual Acoustic Performance Spaces

The three different virtual rooms used in testing were selected to represent discrete performance space types within the range of acoustic environments typically encountered by vocal performers. The three auralized spaces were a recording studio booth (recording studio, AudioLab, York, Table 1), medium hall (National Centre for Early Music, St. Margaret's Church, Table 3), and cathedral (Lady Chapel, St. Alban's Cathedral,Table 4).

3 METHOD

3.1 Latency Measurement

Measurement of OWT between performers (Table 2) was achieved by using a loopback pulse signal to measure the Round Time Trip. Oversized jitter buffers were used to ensure audio data was being read at a constant rate and latency was not expected to vary from this sampled value over the course of the performance task. The sampled Round Time Trip value is halved to provide an estimate of OWT latency between performers. AQ9

Table 2. System configuration data for testing participants.

ID	As	Ac	Bs	Bc	Cs	Cc	Ds	Dc	Es	Ec
Location	Oslo	Gla	Gla	York	Gla	Gla	Gla	Barc	Gla	York
Distance (*km*)	1000	-	300	-	1	-	1700	-	300	-
OWT (*ms*)	54.8	-	29.7	-	22.8	-	37.3	-	33.3	-
BW available (*Mb/s*)	100	10	10	15	10	20	10	90	10	300
BW used (*Mb/s*)	7	-	4.7	-	4.7	-	4.7	-	3.5	-
Fs (*kHz*)	96	-	96	-	96	-	96	-	48	-
Buffer Size (*samples*)	256	-	128	-	64	-	64	-	128	-
Jitter Buffer Size (*ms*)	32	-	10.667	-	6.667	-	14.667	-	10.667	-
FEC Redundancy	2	-	1	-	1	-	1	-	2	-
Singer Experience	P1	Am	P1	Am	P1	P1	P1	P2	Am	P2
Audio Interface	I-2	I-1	I-1	I-3	I-1	I-6	I-1	I-5	I-1	I-4
Headphones	HP3	HP1	HP1	HP1	HP1	HP2	HP1	HP1	HP1	HP1
Microphone	M7	M1	M2	M3	M2	M4	M2	M5	M1	M6

Coded locations are Gla, Glasgow, and Barc, Barcelona. Singer experience is denoted using the code Am, Amateur; P1, Proficient; and P2, Professional. BW available indicates the limiting bandwidth of the connection (typically upload speeds in UK), and BW used indicates the actual BW cost of audio streaming for each performance instance. ID indicates participant duet group (A−E) and individual designation within the group (server, s, or client, c). Audio interfaces are coded I-1, Focusrite Scarlett 18i20g1; I-2, Focusrite Scarlett 18i20g2; I-3, Focusrite Scarlett 2i2g2; I-4, Focusrite Scarlett 2i2g3; I-5, Focusrite Red 4 Pre; and I-6, Motu 828 mk2. Headphones are coded HP1, Beyerdynamic DT 990; HP2, Beyerdynamic DT770; and HP3, Senheiser HD25. Microphones are coded M1, Shure SM57; M2, Rode NT1A; M3, Beyerdynamic MM1; M4, Rode M2; M5, Neumann KM184; M6, DPA 6066; and M7, Audio Technica 2050.

Table 3. Acoustic parameters for National Centre for Early Music, St. Margaret's Church.

Oct. Band (*Hz*)	T30 (*s*)	EDT (*s*)	C50 (*Db*)
31.25	21	2.39	−3.16
62.5	2.69	2.13	−3.53
125	1.82	1.62	−0.03
250	1.6	1.87	−1.41
500	1.49	1.62	−0.58
1,000	1.4	1.49	0.53
2,000	1.3	1.36	1.52
4,000	1.15	1.11	3.2
8,000	0.81	0.59	7.91
16,000	0.52	0.47	12.48

Table 4. Acoustic parameters for Lady Chapel, St. Alban's Cathedral.

Oct. Band(*Hz*)	T30 (*s*)	EDT (*s*)	C50 (*Db*)
31.25	3.44	2.14	–8.39
62.5	3.28	2.52	–2.96
125	3.31	2.4	–6.22
250	2.78	1.76	0.99
500	2.49	1.5	1.3
1,000	2.34	1.5	2.35
2,000	2.1	1.24	2.98
4,000	1.71	0.6	7.66
8,000	1.05	0.35	10.08
16,000	0.62	0.22	14.4

3.2 Onset Detection

Metrics describing performance synchrony are derived from measurement of onset times within parts and comparison of onset times between parts. In order to provide onset detection functionality, an adaptation of the TIMEX [26] onset detection method was developed as a MATLAB[3] script for the purpose of this study, which was accordingly named TIMEX-Lite. Mono recordings of each performer are input to the TIMEX-Lite algorithm. Amplitude thresholding is used to partition musical events, and a pitch-based onset detection method labels onsets within these partitions by identifying pitch fluctuation associated with vocal onsets within a region at the start of the note partition (Fig. 2). The pitch detection and note partitioning components operate separately until peak-picking, providing discrete gating functions on audio input for each of the components in order to remove background noise.

The algorithm outputs a list of onset times associated with each performance. These onset time lists were then used to derive values for a range of synchrony metrics that were used to provide parametric evaluation of the duet performances achieved by each participant pair.

3.3 Synchrony Analysis

Two-way repeated measures ANOVA was conducted on the performance synchrony measurements, with post-hoc pairwise comparison, Shapiro-Wilk normality testing, and Least Significant Difference confidence interval adjustment. For each synchrony metric, the synchrony metric is the continuous dependent variable, and Group (OWT) and Room are the two independent variables. Room is an ordinal categorical variable. Group (OWT) is considered ordinal categorical, as each category reflects a different performer pair; however it could also be considered a continuous numerical variable, with respect to OWT. Synchrony measures within parts are defined by Inter Onset Intervals, detailing the temporal separation of onsets within a part. Synchrony measures between parts are defined by Onset Time Differences (OTDs), detailing the temporal disagreement between performers on the placement of musical events that are intended to be concurrent. Where a between-parts synchrony metric, such as ratio or difference, is used in this study, it is expressed in terms of "server" with respect to "client."

[3]https://uk.mathworks.com/products/matlab.html

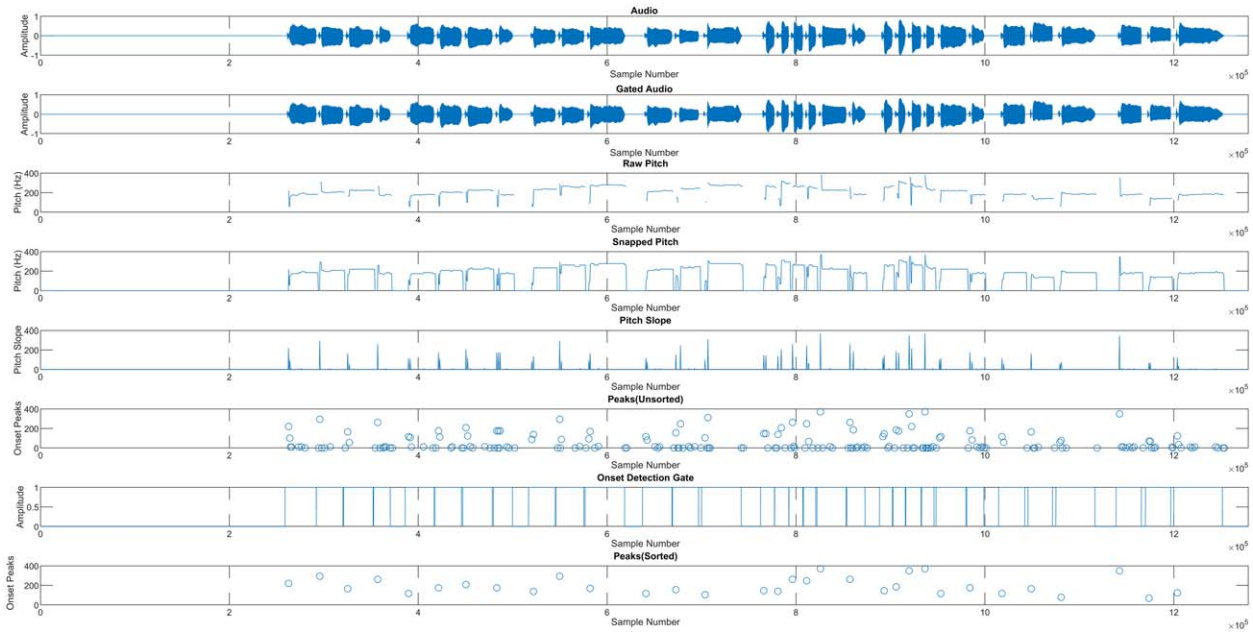

Fig. 2. TIMEX-Lite onset detection process.

Mean tempo slope describes acceleration or deceleration in tempo, measured in BPM/s. Empirical EPT research [11] demonstrates that stable tempo is associated with live musical interactions and tempo deceleration is indicative of impaired performance, typically through failure to adopt effective latency coping strategies when OWT latency exceeds the EPT.

Between parts, *mean tempi ratio* [27] describes the ability of two musicians to perform at the same tempo (distinct from acceleration or deceleration). Live musical interactions can be considered as characterized by a value of 1, where significant deviation from this value indicates that the performance is impaired or synchrony has collapsed entirely. The metric *asynchrony* [8] is defined as the mean of the standard deviation of absolute OTDs within an ensemble performance (simply the standard deviation in duet context). Values in the range of 30–50 ms are considered to characterize *asynchrony* in live performance [26, 28]. *Precision* [26] details the mean of the absolute OTDs between parts in an ensemble performance. For this metric, live musical interactions can be considered characterized by values of 50–70 ms [26]. *Tendency to lead*, defined as the median signed OTD, is used to detect latency coping strategies. Live vocal interactions are characterized by absolute values in the range of 20–40 ms [26], and values that consistently exceed this range are likely to indicate the implementation of leader-follower methods of managing impaired performance conditions because of OWT latency in excess of the EPT.

3.4 Perceptual Evaluation

Parametric analysis of performance synchrony was accompanied by an informal questionnaire where participants provide ratings of perceptual quality of the INMP experience. This includes rating immersive qualities such as *naturalness* [29] and *communication* [14], as well as entertainment qualities such as *enjoyment* and *engagement*. Participants could provide additional comments on each questionnaire item. The questionnaire is available to view online.[4]

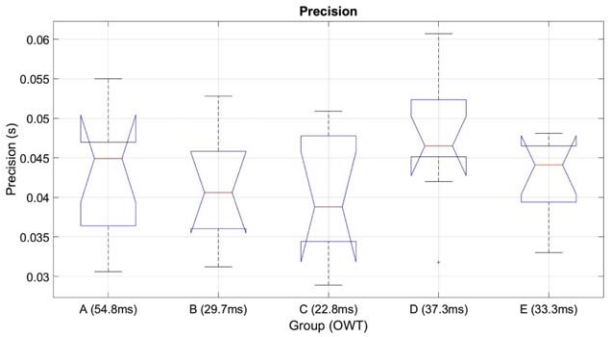

Fig. 3. *Precision* by Group (latency) including data from all rooms. N = 9 per group (45 total). No significant difference between groups.

4 RESULTS

4.1 Precision and Asynchrony

Mean *precision* across duet pairs ranged from 36–52 ms (Fig. 3), and mean *asynchrony* ranged from 27–41 ms (Fig. 4) across all duet pairs. In both cases no deviation from values expected from live performance is present. *Asynchrony* was not found to be significantly effected by Group (OWT) ($F = 2.14$, $p = 0.1009$) or Room ($F = 2.29$, $p = 0.1188$). Nor was there a significant effect of either

[4]https://github.com/SpaceCadetAlba/INMP-supplement.git

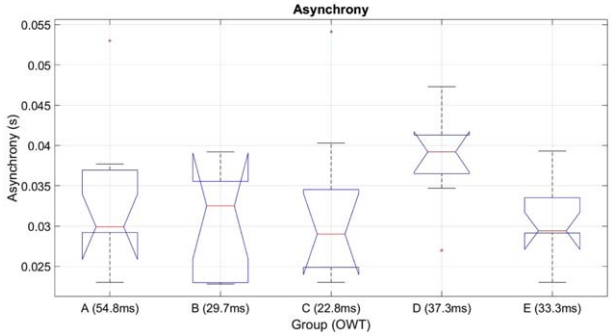

Fig. 4. *Asynchrony* by Group (OWT) including data from all rooms. N = 9 per group (45 total). No significant difference groups.

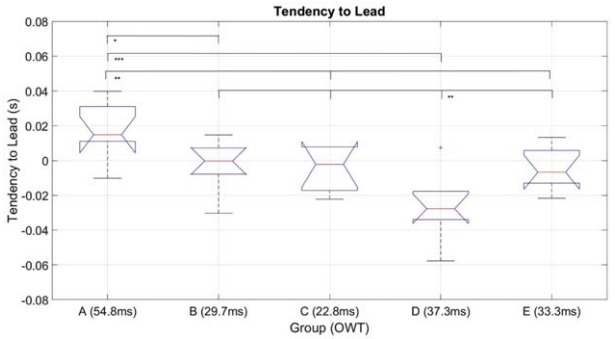

Fig. 5. *Tendency to lead* by Group (latency) including data from all rooms. N = 9 per group (45 total). *, **, and *** indicate significance evaluated at $p = 0.05$, 0.01, and 0.001 respectively.

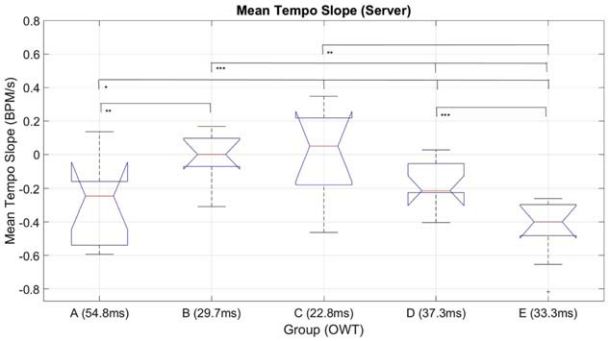

Fig. 6. *Mean tempo slope* (server performer) sorted by Group (OWT) and including data for all rooms. N = 9 per group (45 total). *, **, and *** indicate significance evaluated at $p = 0.05$, 0.01, and 0.001 respectively.

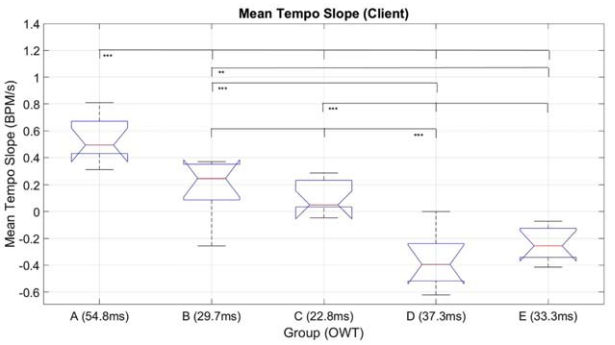

Fig. 7. *Mean tempo slope* (client performer) sorted by Group (OWT) and including data for all rooms. N = 9 per group (45 total). ** and *** indicate significance evaluated at $p = 0.01$ and 0.001 respectively.

Room (F = 1.75, $p = 0.1917$) or Group (OWT) (F = 1.5, $p = 0.2286$) with respect to *precision*.

The higher *asynchrony* of Group D (37.3 ms) compared to other duet pairs, though of no significance, is likely an indication of impaired synchrony due to high latency, as identified with significance in other synchrony metrics.

4.2 Tendency to Lead

Analysis revealed no effect of Room (F = 0.63, $p = 0.5392$) on *tendency to lead*; however it did identify two duet pairs where measurements were significantly different from other groups. Indeed an effect of Group (OWT) was apparent across duet pairs (F = 9.8, $p = 0.0000$, large effect size with partial eta squared = 0.3404). These were notably the two groups with highest OWT, Group A (54.8 ms) and Group D (37.3 ms) (Fig. 5).

Mean values across all performances ranged in magnitude from 1–29 ms, Though these values are within the limits characterizing live musical performance, the significant difference from other groups with *tendency to lead* closer to zero indicates that it is likely that Group A and Group D are implementing leader-follower latency coping strategies across performance.

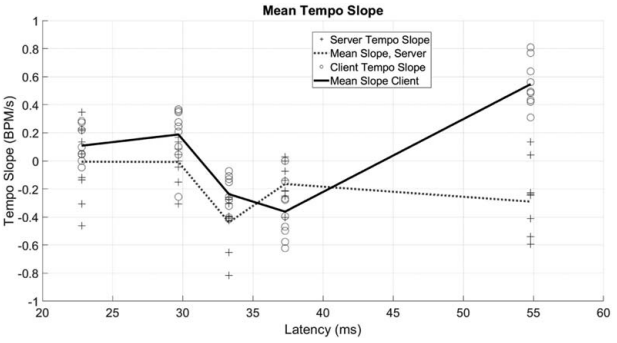

Fig. 8. *Mean tempo slope* across OWT latency.

4.3 Mean Tempo Slope

Mean tempo slope within parts was again not affected by the Room (F = 1.63, $p = 0.2041$) but did demonstrate significant dependency on Group (OWT) (F = 22.56, $p = 0.0000$, large effect size with partial eta squared = 0.4172) (Figs. 6 and 7).

The linear plot of client and server Mean Tempo Slope across latency (Fig. 8) illustrates expected deceleration where OWT is above the EPT. The extreme difference in tempo slope between client and server for Group A (54.8 ms) illustrates extreme latency coping, namely the tendency

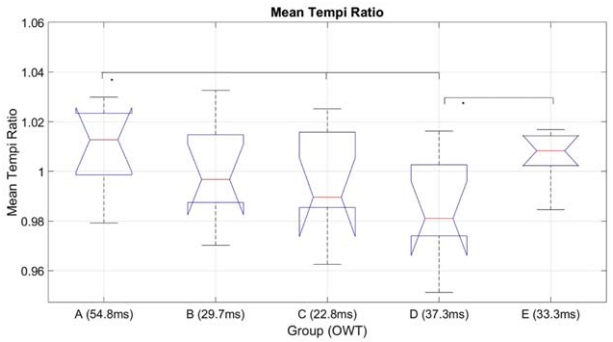

Fig. 9. *Mean tempi ratio* sorted by Group (OWT) including data for all Rooms, N = 9 per group (45 total). * indicates significance evaluated at $p = 0.05$.

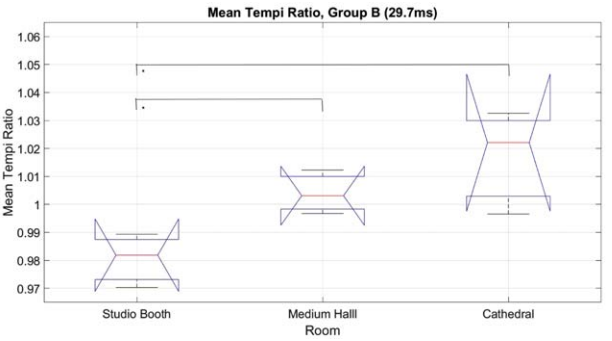

Fig. 10. *Mean tempi ratio* sorted by Room for Group B (29.7 ms). N = 3 measurements per room (9 total). * indicates significance evaluated at $p = 0.05$.

for the client performer to accelerate in an attempt to counterbalance tempo deceleration. Significant dependency on the designation "server" or "client" within duet pairs was also identified. This is exclusive to pairwise difference between participants in Group A (54.8 ms), evaluated at $p = 0.001$. Results from between-parts synchrony metrics allow us to identify that this is because of synchrony collapse in performance from this group. Insignificant variation can be seen between performers in other groups. Though insignificant, this likely simply indicates differences in vocal performance ability within duet pairs.

4.4 Mean Tempi Ratio

Mean tempi ratio measurements indicate a significant effect of Group (OWT) across testing pairs ($F = 3.29$, $p = 0.0237$, large effect size with partial eta squared = 0.1767), where the high-latency groups, Group A (54.8 ms) and Group D (37.3 ms), show a significantly greater deviation from an even *mean tempi ratio* of 1 than other performer pairs, evaluated at $p = 0.05$ (Fig. 9).

Room was also noted as significant across performances ($F = 2.728$, $p = 0.0270$, large effect size with partial eta squared = 0.476). Within performer pairs it was observed that for Group B (29.7 ms) measurements for the studio booth exhibited significantly different ratios than for Room B and C (medium hall and cathedral respectively), evaluated at $p = 0.05$ (Fig. 10).

4.5 Questionnaire Response

Questionnaire responses were overwhelmingly positive, rating the performance experience almost unanimously as enjoyable, with one participant commenting "It was remarkable how familiar the musical setting felt, once I had got over the initial strangeness of the situation."

Notably, multiple cases where participants indicated perceptual rating variance between the three virtual acoustic performance spaces were identified, and in multiple cases participants directly attributed this variance in rating directly to room acoustic parameters. Because of this attribution, it was acknowledged that it would not be appropriate to group data across rooms. This leaves only one data point per combination of conditions. The questionnaire responses therefore provide only an indication of the participants' perception, and robust evaluation in this area will be the topic of future research. The free comment sections of the questionnaire items from each participant also provide useful additional descriptive information. Full results are available to view online.[5]

The studio booth was generally considered too dry to provide an enjoyable and natural vocal performance experience; however the high clarity of this room was, in one case, noted as allowing for enhanced precision in perception of musical events. The medium hall and cathedral were identified as providing better immersive quality with respect to naturalness, spaciousness, and separation. In one case a participant commented that the medium hall seemed to allow for easier establishing of synchronous performance. One participant noted the reverb time in the cathedral environment deteriorated the precision of perception of musical events, while another participant noted an enjoyment preference for the cathedral over the medium hall due to positively identified better clarity and lesser low frequency energy.

5 DISCUSSION

Parametric evaluation of performance synchrony allows robust identification of cases where performance conformed to expectations of live musical interactivity and cases where performance is impaired. Group E (33.3 ms) suffers from tempo deceleration, and Group D (37.3 ms), also exhibiting tempo deceleration, appears to attempt to adopt a leader-follower coping strategy. Group A (54.8 ms) experiences collapsing of synchronous performance. Group B (29.7 ms) and Group C (22.8 ms) both achieve performance synchrony demonstrating characteristics of live musical interactivity across all synchrony measures. This allows for identification of the instances where INMP audio systems may provide a natural vocal performance experience. Notably, these findings follow expectations of the effect of latency described in empirical EPT research.

[5]https://github.com/SpaceCadetAlba/INMP-supplement.git

Table 5. Minimum and recommended technical requirements for optimal performance using the VIIVA-NMP audio system prototype.

OWT	< 30 *ms*
Estimated Range	300 km+
Sample Rate	96 kHz
Buffer Size (min)	128 samples
Buffer Size (recommended)	64 samples
Jitter Buffer Size	<11 *ms*
FEC Redundancy	1–2
BW per Channel	4.7–7.0 Mb/s

An effect of room was identified for Group B (29.7 ms), where the client participant provides a relatively faster and less synchronous performance exclusively in the dry studio booth. This client participant also attributed an increased ability to perceive "pitch and time misalignment" in this room. This was expressed as due to positively identified highest clarity among the virtual acoustic environments used in testing. Perceptual evaluation (primarily through performer comment) identifies that reverberation may be aiding the ability to establish synchronous performance. Given this, it is considered that it may be possible that the smoothing effect of reverberation on the perception of musical event onsets [11] may allow for a "masking" of pitch and time misalignment. This may allow for a greater error tolerance in INMP performance. Alternatively, the clarity of dry acoustics may allow for more precise adjustment of response to errors in performance. The perceptual quality and enjoyment of the performance experience also appears that it may be aided by reverberant environments, though particularly long reverbs or poor clarity may impede performance synchrony.

As noted, no visual display was included in this study. It was, however, also acknowledged that visual contact can improve synchrony in live performance [25, 26]. It can therefore be reasonably expected that inclusion of VR display in future research may also improve upon synchrony results reported in this study.

It was considered that the ability to maintain stable *mean tempo slope* may vary between performers. It is reasonable to expect singer experience to contribute in some way toward performance synchrony, and this shall be controlled for robust evaluation in future study.

6 CONCLUSION

The parametric analysis of performance conducted demonstrates that it is possible to achieve naturally interactive vocal performance using INMP systems. This study allows for estimation of the technical requirements and geographic range associated with achieving this using the VIIVA-NMP audio system design (Table 5).

In the networked virtual acoustic context presented by INMP, it appears that the effect of latency detailed in empirical EPT research holds true. Perceptual rating of the effect of room indicates that the experience of virtual acoustic performance is likely to conform to expectations from real acoustic performance [7]. Observation of the effect of virtual acoustic performance space on *mean tempi ratio*, a physical performance parameter, indicates the potential for acoustics to influence objective aspects of musical performance using INMP audio systems.

7 FURTHER WORK

Testing utilizing home networks reduced the control on some experimental parameters, namely hardware and network delays (with only one latency value associated with each performer pair). Future work will include control on hardware and inclusion of delay incrementing functionality (to facilitate multiple latency conditions for each duet pair). As singer experience is expected to have an impact on performance synchrony this will also see more robust control in future research. Investigation into modeling the effect of virtual acoustics on performance synchrony will require further parametric synchrony analysis. An extended qualitative evaluation will also be required. A conceptual design for achieving this is currently an interactive networked adaptation of the SALTE listening test framework [30]. A parametric evaluation of the effect of OPUS codec compression (which is suitable for low-bitrate compression of Ambisonic signals [31]) on INMP performance will also be a prudent future contribution. Integration of VR environments and avatars will also be required in future work, as will evaluation of the effect of this on INMP performance. A component of this work will focus on the role of VR display visual contact.

8 ACKNOWLEDGMENT

The authors would like to express their gratitude to the participants who took part in this study and Kronlachner VST and JackTrip developers for providing high-quality and open-source resources that were used in this study. This study was supported by XR Stories, University of York.

9 REFERENCES

AQ12

[1] C. Rottondi, C. Chafe, C. Allocchio, and A. Sarti, "An Overview on Networked Music Performance Technologies," *IEEE Access*, vol. 4, pp. 8823–8843 (2016 Dec.). https://doi.org/10.1109/ACCESS.2016.2628440.

[2] S. Delle Monache, L. Comanducci, M. Buccoli, et al., "A Presence- and Performance-Driven Framework to Investigate Interactive Networked Music Learning Scenarios," *Wireless Commun. Mobile Comput.*, vol. 2019, paper 4593853 (2019 Aug.). https://doi.org/10.1155/2019/4593853.

[3] B. Loveridge, "Network Music Performance in Virtual Reality: Current Perspectives," *J. Network Music Arts*, vol. 2, no. 1, paper 2 (2020 Aug.).

[4] C. Eaton and H. Lee, "Quantifying Factors of Auditory Immersion in Virtual Reality," in *Proceedings of the AES International Conference on Immersive and Interactive Audio* (2019 Mar.), paper 103. https://doi.org/10.17743/aesconf.2019.978-1-942220-27-5.

[5] G. Kearney, H. Daffern, L. Thresh, et al., "Design of an Interactive Virtual Reality System for Ensemble Singing," in *Proceedings of the Interactive Audio Systems Symposium*, paper 1 (York, UK) (2016 Sep.).

[6] P. Cairns, H. Daffern, and G. Kearney, "Immersive Network Music Performance: Design and Practical Deployment of a System for Immersive Vocal Performance," presented at the *149th Convention of the Audio Engineering Society* (2020 Oct.), eBrief 632. https://doi.org/10.17743/aesconv.2020.978-1-942220-33-6.

[7] M. Kob, S. V. Amengual Garí, and Z. Schärer Kalkandjiev, "Room Effect on Musicians' Performance," in J. Blauert and J. Braasch (Eds.), *The Technology of Binaural Understanding*, Modern Acoustics and Signal Processing, pp. 223–249 (Springer Nature Switzerland AG, Cham, Switzerland, 2020).

[8] R. A. Rasch, "Synchronization in Performed Ensemble Music," *Acta Acust. united Ac.*, vol. 43, no. 2, pp. 121–131 (1979 Sep.).

[9] S. Farner, A. Solvang, A. Sæbo, and U. P. Svensson, "Ensemble Hand-Clapping Experiments Under the Influence of Delay and Various Acoustic Environments," *J. Audio Eng. Soc.*, vol. 57, no. 12, pp. 1028–1041 (2009 Dec.). http://www.aes.org/e-lib/browse.cfm?elib=15235.

[10] A. Carôt, C. Werner, and T. Fischinger, "Towards a Comprehensive Cognitive Analysis of Delay-Influenced Rhythmical Interaction," in *Proceedings of the International Computer Music Conference*, vol. 2009, pp. 473–476 (Montreal, Canada) (2009 Aug.). http://hdl.handle.net/2027/spo.bbp2372.2009.107.

[11] C. Chafe and M. Gurevich, "Network Time Delay and Ensemble Accuracy: Effects of Latency Asymmetry," presented at the *117th Convention of the Audio Engineering Society* (2004 Oct.), paper 6208. http://www.aes.org/e-lib/browse.cfm?elib=12865.

[12] N. Schuett, *The Effect of Latency on Ensemble Performance*, Honors Thesis, Stanford University, Stanford, CA (2002 May).

[13] C. Chafe, J.-P. Cáceres, and M. Gurevich, "Effect of Temporal Separation on Synchronization in Rhythmic Performance," *Perception*, vol. 39, no. 7, pp. 982–992 (2010 Jan.). https://doi.org/10.1068/p6465.

[14] M. A. Iorwerth and D. Knox, "The Application of Networked Music Performance to Access Ensemble Activity for Socially Isolated Musicians," in *Proceedings of the 5th International Web Audio Conference: Diversity in Web Audio*, pp. 8–13 (Trondheim, Norway) (2019 Dec.).

[15] A. Carôt, C. Hoene, H. Busse, and C. Kuhr, "Results of the Fast-Music Project—Five Contributions to the Domain of Distributed Music," *IEEE Access*, vol. 8, pp. 47925–47951 (2020 Mar.). https://doi.org/10.1109/ACCESS.2020.2979362.

[16] H. Daffern, D. A. Camlin, H. Egermann, et al., "Exploring the Potential of Virtual Reality Technology to Investigate the Health and Well Being Benefits of Group Singing," *Int. J. Perform. Arts Digit. Media*, vol. 15, no. 1, pp. 1–22 (2019 Dec.). https://doi.org/10.1080/14794713.2018.1558807.

[17] J. Tamplin, B. Loveridge, K. Clarke, Y. Li, and D. J. Berlowitz, "Development and Feasibility Testing of an Online Virtual Reality Platform for Delivering Therapeutic Group Singing Interventions for People Living With Spinal Cord Injury," *J. Telemed. Telecare*, vol. 26, no. 6, pp. 365–375 (2020 Jul.). https://doi.org/10.1177/1357633X19828463.

[18] M. Gurevich, D. Donohoe, and S. Bertet, "Ambisonic Spatialization for Networked Music Performance," in *Proceedings of the 17th International Conference on Auditory Display* (Budapest, Hungary) (2011 Jun.). http://hdl.handle.net/1853/51573. AQ13

[19] J.-P. Cáceres and C. Chafe, "JackTrip: Under the Hood of an Engine for Network Audio," *J. New Music Res.*, vol. 39, no. 3, pp. 183–187 (2010 Nov.). https://doi.org/10.1080/09298215.2010.481361.

[20] M. Noisternig, A. Sontacchi, T. Musil, and R. Holdrich, "A 3D Ambisonic Based Binaural Sound Reproduction System," in *Proceedings of the 24th AES International Conference: Multichannel Audio, The New Reality* (2003 Jun.), paper 1.

[21] M. Kronlachner, "Ambisonics Plug-In Suite for Production and Performance Usage," in *Proceedings of the Linux Audio Conference*, pp. 49–53 (Graz, Austria) (2013 May).

[22] M. Kronlachner, "Plug-in Suite for Mastering the Production and Playback in Surround Sound and Ambisonics," presented at the *136th Convention of the Audio Engineering Society (Student Design Competition)* (2014 Apr.). AQ14

[23] D. T. Murphy and S. Shelley, "OpenAIR: An Interactive Auralization Web Resource and Database," presented at the *129th Convention of the Audio Engineering Society* (2010 Nov.), paper 8226. http://www.aes.org/e-lib/browse.cfm?elib=15648.

[24] C. Armstrong, L. Thresh, D. Murphy, and G. Kearney, "A Perceptual Evaluation of Individual and Non-Individual HRTFs: A Case Study of the SADIE II Database," *Appl. Sci.*, vol. 8, no. 11, paper 2029 (2018 Oct.). https://doi.org/10.3390/app8112029.

[25] S. D'Amario, H. Daffern, and F. Bailes, "Synchronization in Singing Duo Performances: The Roles of Visual Contact and Leadership Instruction," *Front. Psychol.*, vol. 9, paper 1208 (2018 Jul.). https://doi.org/10.3389/fpsyg.2018.01208.

[26] S. D'Amario, H. Daffern, and F. Bailes, "A New Method of Onset and Offset Detection in Ensemble Singing," *Logoped. Phoniatr. Vocol.*, vol. 44, no. 4, pp. 143–158 (2019 Mar.). https://doi.org/10.1080/14015439.2018.1452977.

[27] A. Barbosa and J. Cordeiro, "The Influence of Perceptual Attack Times in Networked Music Performance," in *Proceedings of the 44th AES International Conference on Audio Networking* (2011 Nov.), paper 10. http://www.aes.org/e-lib/browse.cfm?elib=16133.

[28] R. A. Rasch, *Aspects of the Perception and Performance of Polyphonic Music*, Ph.D. Thesis, University of Groningen, Groningen. the Netherlands (1981 Apr.).

[29] B. G. Witmer and M. J. Singer, "Measuring Presence in Virtual Environments: A Presence Question-

naire," *Presence: Teleoperators Virtual Environ.*, vol. 7, no. 3, pp. 225–240 (1998 Jun.). https://doi.org/10.1162/105474698565686.

[30] D. Johnston, B. Tsui, and G. Kearney, "SALTE Pt. 1: A Virtual Reality Tool for Streamlined and Standardized Spatial Audio Listening Tests," presented at the *147th Convention of the Audio Engineering Society* (2019 Oct.), e-Brief 536. https://doi.org/10.17743/aesconv.2019.978-1-942220-31-2.

[31] T. Rudzki, I. Gomez-Lanzaco, J. Stubbs, et al., "Auditory Localization in Low-Bitrate Compressed Ambisonic Scenes," *Appl. Sci.*, vol. 9, no. 13, paper 2618 (2019 Jun.). https://doi.org/10.3390/app9132618.

THE AUTHORS


Patrick Cairns


Helena Daffern


Gavin Kearney

Patrick Cairns graduated from Glasgow Caledonian University in 2019 with a B.Sc. Honors degree in Audio Systems Engineering. He is currently an XR Stories-funded postgraduate researcher at the AudioLab, University of York, where his research investigates the design, development, and evaluation of immersive audio NMP systems for group vocal performance. Patrick has worked across a range of audio disciplines, including sound production, film and television, and broadcast.

•

Helena Daffern is currently an Associate Professor of Audio and Music Technology at the University of York. She graduated with a B.A. (Hons.) degree in Music, M.A. degree in Music, and Ph.D. in Music Technology, all from the University of York, UK, in 2004, 2005, and 2009, respectively, before completing postgraduate training as a classical singer at Trinity College of Music, London. Her research utilizes interdisciplinary approaches and virtual reality technology to investigate voice science and acoustics, particularly singing performance, vocal pedagogy, choral singing, and singing for health and wellbeing.

•

Gavin Kearney graduated from Dublin Institute of Technology in 2002 with an Honors degree in Electronic Engineering and has since obtained M.Sc. and Ph.D. degrees in Audio Signal Processing from Trinity College Dublin. He joined the University of York as Lecturer in Sound Design in January 2011 and was appointed Associate Professor of Audio and Music Technology in 2016. He has written over 70 research articles and patents on different aspects of immersive and interactive audio. He is currently Vice-Chair of the AES Audio for Games Technical Committee as well as an active sound engineer and producer of immersive audio experiences.