DOI: 10.22032/dbt.39949

DOI (proceedings): 10.22032/dbt.39936

Full Reviewed Paper at ICSA 2019

Presented * by VDT.

Auralization systems for simulation of augmented reality experiences in virtual environments

Peter Dodds, Sebastià V. Amengual Garí, W. Owen Brimijoin, Philip W. Robinson
Facebook Reality Labs, Redmond, WA, USA, Email: peterdodds@fb.com

Abstract

Augmented reality has the potential to connect people anywhere, anytime, and provide them with interactive virtual objects that enhance their lives. To deliver contextually appropriate audio for these experiences, a much greater understanding of how users will interact with augmented content and each other is needed. This contribution presents a system for evaluating human behavior and augmented reality device performance in calibrated synthesized environments. The system consists of a spherical loudspeaker array capable of spatial audio reproduction in a noise isolated and acoustically dampened room. The space is equipped with motion capture systems that track listener position, orientation, and eye gaze direction in temporal synchrony with audio playback and capture to allow for interactive control over the acoustic environment. In addition to spatial audio content from the loudspeaker array, supplementary virtual objects can be presented to listeners using motion-tracked unoccluding headphones. The system facilitates a wide array of studies relating to augmented reality research including communication ecology, spatial hearing, room acoustics, and device performance. System applications and configuration, calibration, processing, and validation routines are presented.

1. Introduction

We imagine a future world in which a multitude of people will wear augmented reality devices that overlay an entire metaverse of auditory and visual information on their daily experiences. Such devices will influence the way we live, work, and interact with other people. These devices will give us super-human listening abilities and facilitate telepresence with a much higher social signal bandwidth. To reach this future, we will need a much greater understanding of how users with augmented abilities will interact with the devices and each other.

Consider an example where two people are sitting in a noisy restaurant talking to one another. Devices with motion tracking, simultaneous localization and mapping, a microphone array, and binaural sound synthesis could provide many benefits in this situation. With the relative positions and orientations of the participants known, one listener's microphone array could capture the other talker with reduced background noise, and play it back in real time, binaurally positioned in the same spatial location as the actual talker, hence naturally and transparently improving the signal to noise ratio of transmitted speech. The shape and dynamic tuning of the beamformer, the allowable delay of the reinforcement signal, the accuracy of the spatialization, and the values of many other parameters needed to optimize the performance of the system in this scenario are all unknown.

The technology imagined in the scenario above largely exists today, just not in a form factor that is suitable for a head-worn, mobile device. Many technological developments are still needed to package such capabilities into a consumer product. Nonetheless, without having these future devices available for testing, we must utilize what is available to prototype the experiences and further research in the area. This paper describes a real-time interactive auralization system that utilizes

© Verband Deutscher Tonmeister e.V., 2019

currently available technology to prototype future experiences to better understand how users with augmented abilities will interact with their devices and each other.

2. System Overview

Our interactive auralization platform (IAP) is an evolving combination of hardware and software, designed to allow us to test new hardware and software, examine novel AR experiences, and rigorously measure human behavior in real and augmented reality environments. The facility allows us to test, evaluate, and demonstrate the various technologies that are being developed by the FRL audio research group, as well as assess human behavior in multiple subjects simultaneously in a variety of realistic acoustic environments, both real and with virtual components to them. In this way, it acts as a time machine because some of the functionality is not yet possible in current generations of AR hardware, even those that exist only in prototype form. The IAP has a high density loudspeaker array capable of reproducing realistic sound fields and is also capable of presenting real-time spatialized virtual audio wirelessly over headphones, including virtual sound sources, shared audio from nearby talkers, and remote talkers (telepresence), all rendered with very low latency using individualized head-related transfer functions (HRTFs), simulated room acoustics, and a virtual beamformer. The way the IAP is constructed gives us an ability to rapidly iterate changes in functionality to determine the utility and ideal parameters for a potential device feature without having to do extensive hardware development.

The IAP is also capable of making fine-grained measurements of the behavior of multiple people interacting in a common or physically separated space. These measurement capabilities currently include head, hand, and body tracking, eye gaze and pupillometry, and high-fidelity voice capture for up to six simultaneous subjects. Future additions will include finger and face tracking, galvanic skin responses measures, electroencephalography, and heart rate. All the measurement subsystems are tied to a common master clock (Evertz 5601MSC), allowing accurate time synchronization of current and future measurement data. This is critical for examining realtime functionality of devices with novel sensors and capabilities and also analyzing complex behaviors that are associated with real and virtual events in the space. The ability to capture data from multiple people at the same time ensures that we can capture and utilize details of interactive communication, exploration, and play in groups of people.

3. Subsystem Details

3.1. Loudspeaker Array

3.1.1. Hardware

In order to best understand how users will interact with future augmented reality devices, the wide variety of environments in which the devices will be used must be examined. To this end, a semi-spherical loudspeaker array for spatial audio reproduction has been constructed in the interactive auralization platform. The array consists of 49 MiniDSP SPK-4P loudspeakers. The SPK4-P is a compact loudspeaker with a 3.5” driver and an on-board 400 MHz Analog Devices SHARC processor and Class D amplifier. A single CAT5/6 network cable from a PoE or PoE+ enabled switch provides power, audio, and control using the Audio Video Bridging (AVB) communication protocol. For full-band reproduction, audio signals below 120Hz are sent to four miniDSP NDAC-2 AVB endpoints which convert the AVB stream to analog audio to drive four Genelec 7360A SAM Studio Subwoofers.

The loudspeaker array is interfaced to a single Windows PC computer in an isolated control room using an RME Digiface USB. Max/MSP and Matlab are used for real-time processing of the audio streams which allows the array to use different audio rendering pipelines simultaneously and interface with the other render and capture technologies in the system. Figure 1 illustrates the signal path of the IAP subsystems, including the loudspeaker array and audio pipeline.

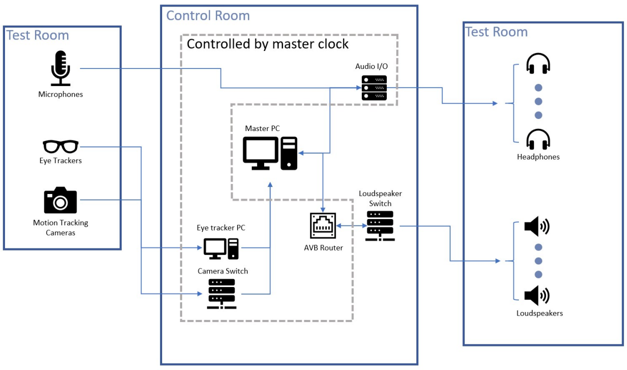


Fig. 1: Schematic diagram of IAP hardware connections.

3.1.2. Array Implementation

Figure 2 shows the 53 loudspeaker array for sound field reproduction that has been installed as part of the Interactive Auralization Platform. The positions of the loudspeaker were determined by circumscribing a sphere on to the room and then optimizing the positions of the loudspeaker for spatial audio reproduction given the architectural constraints of the room. The array consists of four rings of 12 loudspeakers each, roughly approximating a sphere, and a single loudspeaker directly above the center of the room. The subwoofers are placed at the cardinal compass directions at the edge of the room.

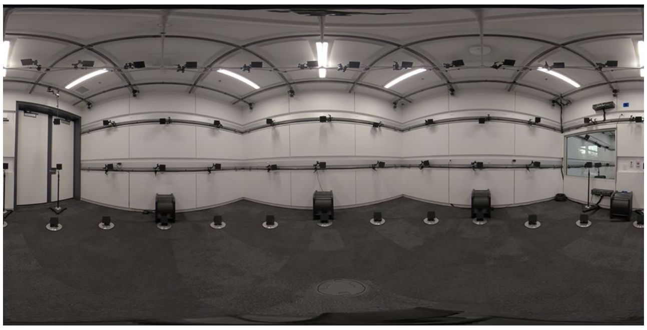


Fig. 2: Equirectangular photo of one of the loudspeaker arrays built for the IAP. (Photo credit: Scott Colburn)

The loudspeakers are individually calibrated with respect to level and delay and corrected to minimize coloration of the soundfield using an automated Matlab script. These filters can be loaded on to the SHARC processor of each loudspeaker or implemented further up the signal path, in the computer, depending on the needs of the IAP. Using the software back-end, a variety of spatial reproduction formats are achievable over the loudspeakers, including vector based amplitude panning (VBAP) [1], Spatial Decomposition Method (SDM) [2], and Higher Order Ambisonics (up to order $N = 6$).

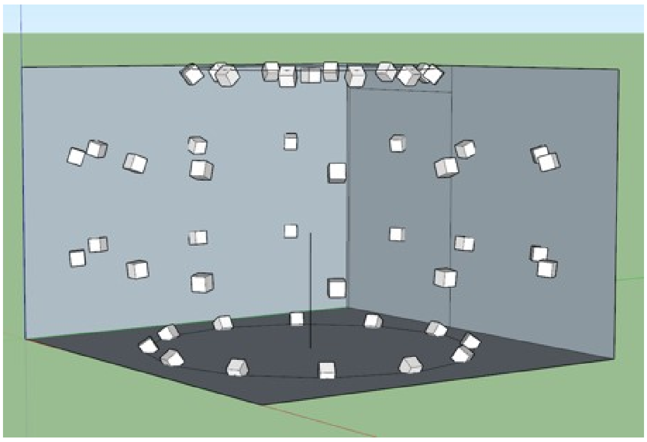


Fig. 3: 3D model showing the loudspeaker positions within the room.

3.1.3. Room Acoustics

In order to reduce the effects of room reverberance on the soundfield reconstruction, the room is acoustically treated. Two inch absorptive paneling is placed on all major surfaces of the room, save for an observation window and the floor. The frequency-dependent reverberation time of the room is shown in Figure 4.

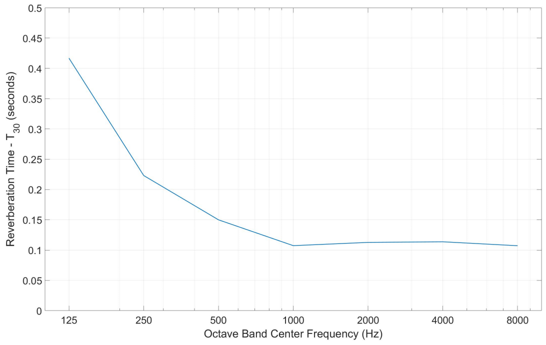


Fig. 4: Measured reverberation time T_{30} in the IAP.

3.2. Motion Tracking

The IAP uses a Vicon motion tracking system, which employs retroreflective infrared markers that may be attached to people or objects and can be tracked at sub-millimeter resolution throughout the capture space. The system currently in use consists of eight Vicon Bonita 10 cameras, mounted at ceiling height in the cardinal compass directions, capable of running at a sample rate of 250 Hz. An alternative version of the IAP uses Optitrack Prime 13W cameras running at 120 Hz. The motion tracking systems may be used to capture data on listener movements for behavioral capture but is also used to drive a real-time loudspeaker array and binaural rendering system. The position and rotation data of all objects tracked in the space is accessed using the Vicon DataStream SDK (or Optitrack Motive SDK), which streams the Cartesian coordinates and 4-element quaternions of each tracked object to other pieces of software on the same or other computers on the subnet. Currently this data is captured in Matlab, which uses a custom script to compute the angles (azimuth and elevation) and distance of all tracked objects relative to each other, in local coordinate frames. Untracked virtual objects may also be added at this stage. The relative Euler angles and distances are then rebroadcast at a frame rate of 100 Hz over UDP to Max/MSP, which handles the room acoustic rendering, loudspeaker DSP, and/or binaural spatialization.

3.3. Eye Tracking

Gaze tracking is accomplished using Ergoneers glasses (Dikablis II) containing infrared emitters and three cameras: two eye-facing cameras and one world-facing camera, all running at a frame rate of 60 Hz. These devices can be wireless or wired, depending on configuration, and interface with our data acquisition and streaming pipeline with D-Lab, with data being shared over UDP to Vicon. This allows us to assess and utilize gaze angle to alter sound presentation parameters in realtime in the same coordinate reference frame as the motion tracking data, and further allows us to ensure that data timestamps are uniform across the different software packages and measurement systems. Data captured for offline analysis includes gaze direction, pupil height, and pupil width, with HD videos captured from all three cameras for further analysis.

3.4. Voice Capture

Audio is captured for analysis or re-spatialization from each participant using DPA microphones (4088 directional microphone) in a boom-mount configuration. The signals are transmitted over wireless transmitters (Sennheiser SK 100 G4) and captured in Max/MSP using an RME Fireface UFX II multi-channel sound card at a sample rate of 48 kHz. Time alignment between presented and captured signals is ensured by using the same sound card, which along with the motion and eye tracking systems is slaved to the IAP master clock.

3.5. Headphone Subsystem

Virtual sound sources may be presented binaurally over headphones in the IAP, but care must be taken to minimize the impact that the headphones have on the listener's perception of real world or loudspeaker array-generated signals. The ability to present sounds to a listener without interfering with the natural sound path is a fundamental requirement of auditory AR. As there are no currently available headphones that allow the presentation of fully broadband binaural signals to completely unoccluded ears without cross-talk, three types of commercially available

© Verband Deutscher Tonmeister e.V., 2019

devices are used in our work, varying inversely in the width of their frequency response and in the degree to which they impede the signal path of real world signals into the ear canal. The first are AKG K1000 open ear headphones, large diaphragm headphones that are suspended over the ears, rather than being pressed against the head. These are high fidelity headphones and have the strongest low frequency response of the headphones used, but they interfere the most with the natural sound field due to their large size. The second type that is used are Sony PFR-v1 headphones, small, spherical loudspeakers suspended in front of the pinnae, with a bass port that is routed through a hollow metal loop that sits in the ear behind the tragus. These are capable of a relatively flat response over a wide range of frequencies but have poorer low frequency response than the AKGs as they are smaller, however, they interfere less with the natural sound field. Finally, a pair of custom headphones are used that consist of a small in-ear headphones (Sennheiser IE4) that are suspended roughly 1 cm from the opening of the ear canal using a custom 3D printed mount. These have the poorest low frequency reproduction, but in principle occlude the ears the least. Two other categories of AR transducers are also currently used, but as they are proprietary technology they will not be discussed here. All headphone signals are presented via an RME Fireface UFXII sound card either wired or wirelessly with remote monitoring transceiver (Sennheiser EW IEM G4).

4. Audio Stimuli

4.1. Sound Generation

All audio is processed in Max/MSP, utilizing a variety of pre-made and custom objects including the SPAT toolbox from IRCAM. Max/MSP receives the azimuth, elevation, distance, and level of the desired virtual sound sources via UDP from Matlab and renders the appropriate virtual signals at these locations for up to six listeners. Arbitrary HRTFs can be loaded into the rendering software, ensuring flexibility in presenting individualized signals to multiple listeners. Soundfields are rendered using pre-recorded ambisonics content or by artificially generating sources using traditional sound design techniques and encoding them in to a spatial reproduction format. Room acoustics simulations are handled in two different ways, depending on the configuration of the system, but both are implemented in Max/MSP.

4.2. Room Acoustics Simulation

4.2.1. Room Re-Synthesis

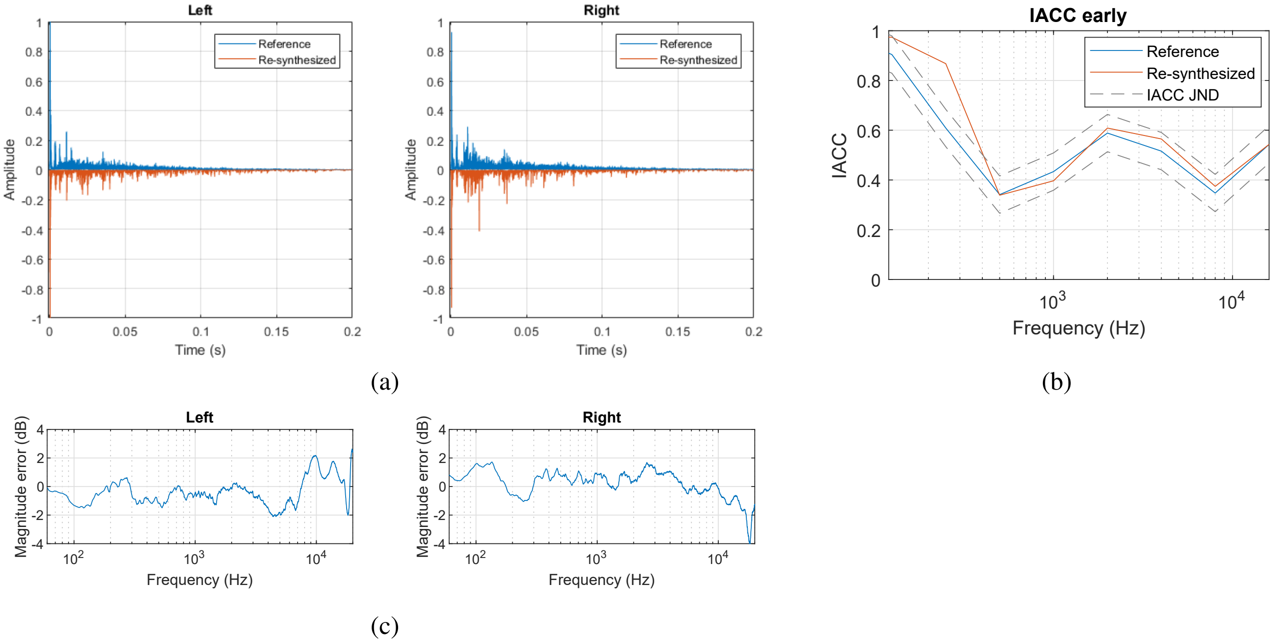
Generating perceptually plausible virtual acoustic objects in AR requires matching the acoustics of the simulated percepts to those of the real room. Once the acoustic divergence between real and virtual sounds becomes too large, virtual binaural percepts do not appear well externalized, and thus impair the quality of experience [3]. However, it is unknown what are acceptable deviations and to what extent it is necessary to match the acoustics of the virtual sounds to those of the real space. To enable research in perceptual thresholds of room acoustics for augmented reality it is useful to generate binaural renderings that are based on the measured acoustics of the room.

A straightforward implementation consists of measuring a binaural room impulse response (BRIR) using a head and torso simulator (HATS) with a variety of head orientations. However, this approach is time consuming and does not allow for personalization using individualized HRTFs. An alternative approach is based on microphone array measurements at the listening position and binaurally re-synthesizing the sound field. The method used here is largely based on the Spatial Decomposition Method (SDM) and the re-synthesis of BRIRs by combining the sound-field parametrization data (a pressure monaural RIR and direction-of-arrival information) with an arbitrary dataset of HRTFs. An extensive technical description and validation of the auralization approach is described in [4]. A variety of manipulations of the rendered BRIRs is presented as well in [4], including arbitrary modifications of the frequency-dependent reverberation time, the inclusion of fully synthetic late reverberation, or the manipulation of the spatial characteristics of the sound field.

The re-synthesized BRIRs are dynamically convolved in real-time in three separate pipelines: direct sound, early reflections, and late reverberation. This allows the implementation of real-time manipulations of the BRIR, enabling the study of HRTF manipulations, direct-to-reverberant ratio (DRR), or mixing time, among others. Reproducing room acoustics based on in situ measurements over non-occluding headphones allows us to compare real and virtual sources in real-time and objectively assess the perceptual differences between them.

Pilot listening tests have been completed to assess the degree of authenticity and plausibility of the auralizations. A discrimination test based on a two-alternative forced choice (2AFC) with a reference revealed that if listeners are provided with unlimited listening time the binaural renderings are not indistinguishable from the real loudspeaker. To our knowledge, there is only one study available in the literature testing perceptual authenticity of dynamic binaural synthesis [5]. In that case, although the BRIRs were measured in situ using binaural microphones, perceptual authenticity was not achieved. We have found that typically small deviations in spectral content and localization are the most common attributes used to judge deviations between a reference real sound and a binaural rendering. In order to test plausibility, we conducted a pilot study where a loudspeaker was covered behind an acoustically transparent baffle, and spatially degraded versions of a binaural render were compared to the real loudspeaker. In this case, we found that real and virtual sources appear to be equally plausible. Formal studies are being conducted to confirm these initial findings.

Figure 5 shows a comparison between a measured BRIR with a mannequin and a re-synthesized version of this BRIR using the above described method. As it can be observed, the time-energy properties of the left and right channels are largely preserved, although some spurious reflections



(a)

(b)

(c)

Fig. 5: (a) Absolute value of a BRIR measured with a mannequin and a re-synthesis. (b) Interaural Cross Correlation of the early part (0 to 80 ms) of the measured and re-synthesized BRIRs. (c) Spectral error after monaural equalization.

can be observed in the re-synthesis. The Interaural Cross Correlation (IACC) of the measured and re-synthesized versions fall within ± 1 JND (0.075 as defined in the standard ISO 3382). The spectral error falls within ±2 dB up to 16 kHz.

4.2.2. Artificial Reverberation

In this configuration, flexibility and realtime performance is prioritized, and is used when participants are expected to walk around the room, interacting with multiple virtual and real sound sources. Here we use the feedback delay network in SPAT, with the parameters manually matched to the natural acoustics of the IAP room. Alternative room models may also be used that do not match the acoustics of natural signals in the room. In the future, the real-time propagation engine found in the Oculus Audio SDK will be implemented as an additional room simulation configuration.

5. Example Use Cases of the IAP

5.1. Virtual spatial mixing of musical compositions

In this scenario, two participants are fitted with wireless open-ear headphones and motion tracking markers. The room contains six physical models of six different musical instruments, each with a motion tracking marker array, each corresponding to an individual track of a song. Virtual sound sources for each track are individually spatialized to the location of the corresponding instrument. This allows the participants to pick up a given instrument – which triggers the playback of the associated track—and place it wherever they like in the capture space, with the audio for that instrument appearing to emanate from the correct location. Once each instrument is placed, the participants may then freely move through the complete sound field that they have created, allowing them to, e.g., stand back and hear the recording as the musicians may have been when on stage, or to sit down next to the drummer and hear the percussion more clearly. This installation allows people to interact with music in a way that is novel and engaging, and also gives us a good testing ground for examining the plausibility and authenticity of new spatial rendering techniques and room acoustics simulations.

5.2. Real/virtual sound source comparison

In this application, the room renderings described in 4.2.1 are used to present an augmented soundscape composed of real sources (loudspeakers) and binaural virtual sources that mimic the acoustic properties of other visible loudspeakers. For instance, in a musical excerpt the voice of a singer is presented over a loudspeaker, while the guitar accompanying their vocals is presented binaurally over headphones. Listeners are then asked to identify which source is being played from the headphones. Additionally, extra controls are provided to modify the level of the direct sound, early reflections and late reverberation or to fully remove the room acoustic component of the binaural renders. This allows listeners to interactively explore the perceptual importance of sound propagation on virtual audio for augmented reality and its importance on the perceived realism.

6. Conclusion

Augmented Reality devices will provide a novel framework with which users can interact with the world and each other. Understanding these interactions is a high dimensional problem involving many sensory modalities and requires novel solutions to gain insight. The Interactive Auralization Plat-

© Verband Deutscher Tonmeister e.V., 2019

form provides a vehicle to gather data about how users with augmented abilities may interact with their devices, environments, and each other by integrating currently available technology in an experientially meaningful way. The use of audio, visual, and other sensing technologies as well as the ability to augment the acoustic environment of the user, allow for robust and scalable data collection that allows for the experimental evaluation of new technologies and features that would otherwise be inhibited by form factor, compute, and sensor integration challenges. The IAP has proved valuable to understand the challenges and opportunities of augmented reality technology and further evolution of the platform is planned in the future to integrate our findings and address new areas of research.

7. References

[1] V. Pulkki, “Virtual sound source positioning using vector based amplitude panning,” *J. Audio Eng. Soc*, vol. 45, no. 6, pp. 456–466, 1997.

[2] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki, “Spatial decomposition method for room impulse responses,” *J. Audio Eng. Soc*, vol. 61, no. 1/2, pp. 17–28, 2013.

[3] S. Werner, F. Klein, T. Mayenfels, and K. Brandenburg, “A summary on acoustic room divergence and its effect on externalization of auditory events,” in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, June 2016.

[4] S. V. Amengual Garí, O. Brimijoin, H. Hassager, and P. Robinson, “Flexible binaural resynthesis of room impulse responses for augmented reality research,” in *EAA Spatial Audio Signal Processing Symposium*, 2019.

[5] F. Brinkmann, A. Lindau, and S. Weinzierl, “On the authenticity of individual dynamic binaural synthesis,” *The Journal of the Acoustical Society of America*, vol. 142, no. 4, pp. 1784–1795, 2017.