Anechoic audio and 3D-video content database of small ensemble performances for virtual concerts

David Thery, Brian F.G. Katz

PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Anechoic audio and 3D-video content database of small ensemble performances for virtual concerts

David THERY [(1)], Brian FG KATZ [(2)]

[(1)]Paris-Saclay University, LIMSI-CNRS, *CPU Team*, Orsay, France, david.thery@limsi.fr

[(2)]Sorbonne Université, CNRS, Institut Jean le Rond d'Alembert, *Lutheries - Acoustique - Musique*, Paris, France, brian.katz@sorbonne-universite.fr

Abstract

This paper presents the details related to the creation of a public database of anechoic audio and 3D-video recordings of several small music ensemble performances. Musical extracts range from baroque to jazz music. This work aims at extending the already available public databases of anechoic stimuli, providing the community with flexible audio-visual content for virtual acoustic simulations. For each piece of music, musicians were first close-mic recorded together to provide an audio performance reference. This recording was followed by individual instrument retake recordings, while listening to the reference recording, to achieve the best audio separation between instruments. In parallel, 3D-video content was recorded for each musician, employing a multiple Kinect 2 RGB-Depth sensors system, allowing for the generation and easy manipulation of 3D point-clouds. Details of the choice of musical pieces, recording procedure, and technical details on the system architecture including post-processing treatments to render the stimuli in immersive audio-visual environments are provided.

Keywords: Anechoic, Recordings, Point-Cloud, Auralization, Virtual Reality

1 Introduction

Auralization is based on the convolution of anechoic recordings with room impulse responses (RIR). This technique allows one to simulate how the recorded material would be heard in the measured room, for a given source/receiver pair. These RIRs can either be measured in the actual space or simulated, by employing various numerical methods [1]. Over the last decade, the perceptual quality, and hence the ecological validity, of such acoustical simulations has improved, as recent research has shown that simulated auralizations can be subjectively comparable to measured ones on a given set of attributes [2]. Nevertheless, the primary source (the anechoic material) remains one of the key factors to achieving realistic simulations. Several public sources exist for anechoic recordings which could be used for auralization purposes. Examples span almost 50 years. The BBC Orchestra conducted early anechoic recordings in 1969, however these recordings do not appear to be available anymore [3]. The Japan Audio Society released a CD in 1985 including various extracts of solo instrument music [4]. The Archimedes project from Bang and Olufsen followed in 1992 [5], providing high-quality solo instruments and speech recordings. Soon after, Denon published a well-known number of recordings of orchestral music [6], including classical, romantic and post-romantic pieces. However, these full orchestral recordings have a poor signal-to-noise ratio. In addition, the entire orchestra was recorded simultaneously using close-microphones on an acoustically damped stage and then down-mixed to a stereo recording, with the results not being well-suited to detailed auralizations due to the inability to separate instrument tracks.

Multi-channel recordings of symphonic music were recorded by [7] in 2005, but these recordings have not been made readily available, mainly due to copyright permission from the orchestra. A choir was recorded in an anechoic chamber in 2005, providing six choral arrangements with 80 singers [8]. One of the most thorough and high quality resource to date are the recordings described in [9] in 2008: instruments were individually recorded providing high-quality and perfectly separated instrument recordings of classical music from various periods, including Mozart, Beethoven, Mahler, and Bruckner. Lastly and most recently, in 2016, [10] conducted

recordings of opera orchestra and soloists, from composers including Donizetti, Verdi, and Puccini. These last recordings are not perfectly anechoic, being carried out in a dry room.

Overall, few high-quality anechoic recordings are available, especially for multiple instruments. It should be noted that a majority of recently built concert halls are multi-purpose halls, welcoming both classical and jazz ensembles. Yet, few if any anechoic jazz ensemble recordings have been made publicly available.

Aside from basic audio auralizations, there is growing interest and effort to couple audio renderings with associated visual models, potentially rendered in virtual reality (VR) [11, 12, 13]. It is generally acknowledged that auditory and visual modalities interact with each other, and both visual and auditory perceptions are significantly improved when the audio-visual stimuli are coherent (when the sound matches the visual). Research into the spatial coherence of audio-visual renderings in auralizations has shown that the visual distance between source and listener affect the perceived auditory distance and loudness, though this effect has shown inter-individual variations [14].

In addition to the inclusion of visual rendering, recent studies have shown that the inclusion of dynamic voice directivity in simulated auralizations enhances the perception of envelopment and apparent source width, as well as the plausibility of the simulation [15].

In the context of these recent findings, this work presents the acquisition of synchronized audio and video recordings of small ensembles playing a variety of musical pieces, including different orchestrations, with two main goals:

- Extending the range of publicly available databases of anechoic audio recordings to cover a wider range of musical styles, hence enabling a more adapted music selection for a given space under study.
- Providing the community with audio and 3D-video recording datasets that can be easily integrated into VR environments (potentially including dynamic instrument directivity, see description in Sec. 3.4).

This paper begins with a description of the musical selections, followed by the procedure employed for the recordings. Technical details are provided in Sec. 3. Finally, potential applications and future work are discussed in Sec. 4.

2 MUSIC SELECTION AND MUSICIANS

2.1 Criteria of selection

Several requirements were defined for the selection of the musical pieces: they needed to provide valuable acoustical interest (presence of tutti, solo passages, wide occupation of the frequency spectrum), to represent different periods and musical styles (particularly those missing from currently available anechoic recordings), and to stage different orchestrations including various instruments. Particular attention was given to movable instruments (like saxophone, clarinet), with the final aim being the inclusion of coherent directivity in the audio that follow the musicians movement, also represented in the visual rendering. A summary of the different musical excerpts is provided in Tab. 1.

2.2 Classical Baroque

One of the missing musical period in the work of [9] was the Baroque era; which they decided not to record mainly because of the presence of the harpsichord, which was deemed difficult to record properly. Given the numerous composers, different geographical influences, number of orchestrations, and overall richness of this period, it was decided to acquire Baroque music material. The 3 pieces were interpreted by the same quartet, comprising 2 violins, 1 viola, and 1 cello.

The first selected piece was the second movement of the 3rd Orchestra Suite (BWV 1068 no.3) called Aria in D-minor. This 1’10" long piece is one of the most famous piece from J.S Bach, and providing various harmonies with a slow tempo (*Largo/Larghetto*).

Table 1. List of selected musical pieces, including title of the piece, composer, period/style, instruments used in the recorded interpretation, and the duration of the extract.

Musical piece 1	Composer	Period/Style	Instruments	Length
BWV 1068 no.3 Aria	J.S. Bach	Baroque (1739)	2 Violins, 1 Viola 1 Cello	1'10"
BWV 1080/15 Canon alla ottava	J.S. Bach	Baroque (1721)	2 Violins, 1 Viola 1 Cello	1'07"
RV 315 Opus 8	A. Vivaldi	Baroque (1723)	2 Violins, 1 Viola 1 Cello	2'35"
Minor Swing 2	D. Reinhardt	Manouche Jazz (1937)	2 Guitars Violin, Double-Bass	1'15"
Don't mean a thing (If Ain't Got That Swing)	D. Ellington	Swing Jazz (1931)	Saxophone, Guitar Double-Bass	2'00"
Si tu vois ma mère	S. Bechet	New Orleans (1952)	Clarinet, Guitar (Tenor) Double-Bass	1'30"

The second selected piece was part of the Art of the fugue, the last piece of work from J.S Bach. An extract of 1'07" from the Canon Alla Ottava in D-minor was selected. The first phrase of this Canon is written for the violin alone, followed by the cello; the different voices alternate often, at a moderate tempo (*Adagietto*), enabling one to focus on the different instruments and their frequency ranges successively.

The third and final piece from the Baroque era was the 3rd movement of the Summer of the 4 seasons from Vivaldi, in G-minor. This piece of 2'35", played *presto*, comprises a variation of dynamics, including violin solo passages, *crescendos*, and *tutti*.

These pieces, being interpreted with 2 violins, 1 viola, and 1 cello, are also interesting from the viewpoint of moving instruments.

2.3 Jazz

Common jazz instruments include saxophone, trumpet, trombone, clarinet, piano, guitar, double-bass, voice, and drums. Different periods can be characterized by different orchestrations.

It was decided to record *Manouche* jazz, a style from the 1930's characterized by a rythmic basis played by 2 guitars and a double-bass, accompanying a melodic violin, with the absence of percussion, woodwinds, and brass. The chosen title was the well-known *Minor Swing* in A-minor, interpreted by a trio comprising 1 double-bass, 1 guitar, and 1 violin, covering a large part of the audible frequency range. The recorded extract lasted 1'15".

Two pieces were selected to extend the range of instruments available, as well as other styles of jazz music: the first 2 min of *Don't mean a thing*, written in 1931 by Duke Ellington, played by a trio comprising a tenor saxophone, guitar, and double-bass. This piece was selected to represent the *Swing* jazz period. Finally, the first 1'30" of a Sydney Bechet song, recorded in 1952 (*New Orleans* period), played with a saxophone alto, guitar, and double bass was chosen for the particular timbre of the saxophone alto and the low density of notes in this piece, allowing for more perception of room effects.

2.4 Musicians

To cover the different styles chosen described above, 8 musicians from the *Sorbonne University's Choir and Orchestra (COSU)* were recruited and paid, including: 1 double-bass player, 1 cello player, 1 guitarist, 3 violonists, 1 altist, and 1 saxophonist. Their mean age was ≈ 20 years. All musicians had previous recording experience, though not in anechoic conditions.

3 RECORDINGS

3.1 Procedure

The objective of this work was to record both the audio and video of each musician, to allow maximum flexibility for future simulations. To achieve the high quality and perfect separation of instruments in the audio recordings, it was decided to individually record each instrument/musician.

The recording procedure was defined as follows:

- A preliminary reference take in which all musicians were aligned in a linear row and played together was recorded, in the anechoic room. Except the saxophone, violins, and viola which were recorded with a microphone mounted on the instrument, each instrument was recorded using figure-of-8 pattern close-microphones, with the nulls pointing perpendicular to the associated instrument, to reduce the sound captured from neighboring instruments. In order to provide a natural acoustic for the musicians, rather than the very dry anechoic chamber's acoustics, the live captured audio of each instrument was processed in real time by adding a reverberation of 1 sec to simulate a small studio (usual in studio recordings) and rendered over headphones identically to all musicians. The listening level was adjusted to maximize the musicians' playing comfort, optimizing the balance between their own instrument sound, the added reverberation, and the surrounding instruments. This live reverb processing was achieved using SPAT[1], with one instance per instrument. The processing was adjusted with the live feed virtually placed at 1 m in front of the listener, with a room size of 600 m^3 and an aperture of 90°.
- To provide a performance reference for the musicians when they were subsequently recorded in isolation, comparable to the reference piano track and conductor video in [9], a downmix of the reference performance was created for each musician for which their individual contribution was excluded. In this way, it was intended that musicians would be able to perform in a more "natural" way, mimicing their prior performance when playing in ensemble.
- A second recording session was then performed in which each musician was individually recorded while listening to their individual monitor mix. The same live reverberation processing was employed for the individual recording sessions. Subsequent to this recording, each musician was allowed to listen to their individual take and to the whole. They were able to retake the session if desired.
- For each of these individual takes, the video capture from the Kinect cameras system[2] was manually started and stopped.

3.2 Audio anechoic recordings

3.2.1 Anechoic chamber

All recordings were been carried out in the anechoic chamber of the Sorbonne Université, interior working dimensions of $7.95 \times 5.5 \times 4.07$ m, excluding the surrounding wedges of 0.85 m depth. The chamber is therefore assumed anechoic for frequencies above 80 Hz. Musicians were always located at least 1.5 m away from the tip of the anechoic wedges. The chamber has an equivalent background noise level of $LAeq = 16.6$ dBA. A surveillance camera installed in the chamber enabled visual monitoring of the recordings sessions from the adjacent control room.

3.2.2 Microphones

Several options were considered for micing the instruments. Use of purely electric instruments to avoid capturing other instruments sounds was discarded for two reasons: first, the altered timbre of such instruments as compared to acoustic ones, and second, the fact that even professional musicians are not used to playing this

[1] http://www.forumnet.ircam.fr/product/spat-en/
[2] https://developer.microsoft.com/fr-fr/windows/kinect

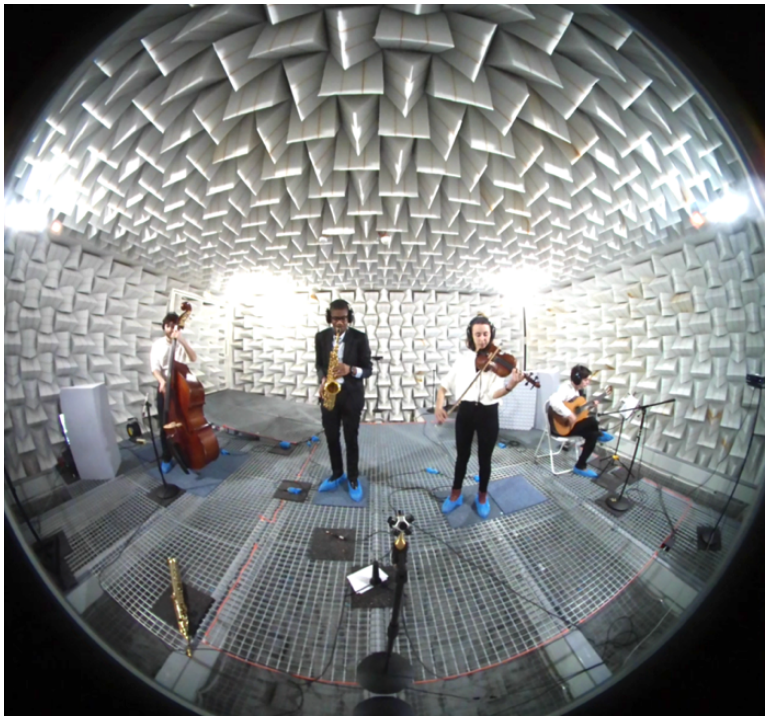

Figure 1. Musicians wearing headphones and playing together for the reference recording (1st session). Saxophone and violin were recorded using DPA4060 mounted on the instrument, while double-bass and guitar were recorded using Schoeps MK-8 figure-of-8 microphones.

type of instrument. Use of piezo-electric microphones was discarded due to the cost/lack of availability in the lab of high-quality microphones of this type and the additional processing required to transform the signal into natural instrument sound. After discussions with a few recording professionals[3], it was decided to record the instruments using closely placed microphones. The exact micing configuration was adapted for each instrument, following guidelines from recording engineers [19].

For the reference recording, as the goal was to record the ensemble with minimal cross-talk between pickups, the musicians were aligned in a single row. In addition to close-mounted microphones, figure-of-8 microphones pointing towards the instruments were used, to provide rejection of sounds coming from the adjacent musicians, as described in the procedure, and depicted on Fig. 1.

For the second recording phase, solo instrument recordings, protocols depended on the instrument category: *static* versus *moving* instruments with respect to needing the microphones being fixed to the instrument or not. For static instruments, including double-bass, cello, and guitar, one fixed omni-directional microphone (DPA 4006) was placed approximately 1 m away, pointing towards the instrument. Two supplementary microphones were placed off-center at 2 m distance from the instrument, to provide additional resources for any subsequent equalization (while also providing minimal data for comparing with theoretical radiation patterns of each instrument [18]). For moving instruments, such as violin, viola, and saxophone, closely placed miniature microphones (DPA 4060) were mounted on the instruments. These microphones were chosen for their high-quality: linear frequency response, good directional pattern[4,5], and low noise-floor (respectively $LAeq = 15$ dBA and 23 dBA).

[3] https://www.ens-louis-lumiere.fr/en/

[4] https://www.dpamicrophones.com/ddicate/4006-omnidirectional-microphone

[5] https://www.dpamicrophones.com/dscreet/4060-series-miniature-omnidirectional-microphone

3.2.3 Hardware/Software details

Audio and video were recorded using two distinct and independent systems. A dedicated computer (Macbook 2.5GHz, 16 GB RAM) managed all the audio, using a custom Max[6] patch. The signals from the microphones recording the instruments were acquired using a Fireface 802[7] audio interface, at 44.1 kHz/24bits. The audio computer also provided the monitor output mix to the headphones worn by the musicians, as well as a general monitor for the recording engineer.

3.3 RGB-Depth video recordings

In parallel to the individual audio recordings, each musician was also visually recorded in order to acquire 3D images for use as realistic performance avatars playing on stage in VR simulations. Visual recordings were carried using three Kinect v2 RGB-D sensors, allowing for the acquisition of 3D point-cloud representations of each the musician. Such visual data can be integrated, for example, into a Unity[8] VR scene without the need to create an animated avatar.

The remainder of this section describes the architecture of the system used for the 3D-video recordings as well as the RGB-D image post-processing applied to obtain the final 3D-point-cloud.

Recording from multiple Kinects was achieved thanks to the `LiveScan3D` library[9] developed in C++/C# and the `OpenCV` library[10]. This system enabled for the production of a single fused 3D point-cloud of each musician from the data acquired from the separate cameras, viewable from different viewpoints (see Fig. 2).

The `LiveScan3D` library is based on a client/server architecture, the clients communicating with the server via TCP/IP. In principle, it allows for the use of any number of Kinect camera clients, with each Kinect connected to a separate computer (a limitation originating from the Kinect v2 SDK that does not allow multiple Kinect's on a single PC). For this database, 3 sensors were used, resulting in a few small shadow zones behind the musician in the 3D reconstruction of the point-cloud [17].

The server manages all the clients simultaneously for recording sequences. The composite point-cloud can directly be streamed in Unity[11] or to a viewer available in the `LiveScan3D` library. For the creation of the visual part of the database, the sequences were saved as `.ply` format files[12], a format which supports colored point-cloud data (either binary or ASCII format). These files can then, for example, be imported in Unity and played frame by frame to provide the visuals of each musician playing. A similar approach, though with only a single Kinect, has previously been used for the insertion of visual actors into an audio-visual rendering of a theatre auralization [14, 22, 23].

3.4 Post-processing

The added value of having video recordings is that dynamic source positions and orientations can be extracted and employed to provide dynamic acoustic directivity in auralizations, following the method described in [15], based on a 12 beams angularly equi-repartited decomposition of impulse responses.

Briefly, after convolving the mono recorded instrument track with the HOA-RIR of each source beam and filtering into separate frequency bands, dynamic source directivity can be adjusted in real-time by altering the gains of the different source beams to create the desired pattern and orientation. In contrast to the method proposed by [7], which used multi-channel anechoic recordings, this method offers a better representation of source directivity in the rendering architecture. RIR simulations need only to be calculated once, for a set of directional sources on a sphere. Current research efforts involve the development of a real-time implementation

[6] https://cycling74.com/products/max-features
[7] http://www.rme-audio.de/en/products/fireface_802.php
[8] https://unity3d.com
[9] https://github.com/MarekKowalski/LiveScan3D
[10] https://opencv.org/
[11] https://github.com/MarekKowalski/LiveScan3D-Hololens
[12] http://paulbourke.net/dataformats/ply

Figure 2. Example of cello player 3D point-cloud reconstituted image, viewed from two different angles.

of this method in order to study musician/room interactions [21].

4 Outlook

This paper presented the individual recordings (both 3D video and anechoic audio) of small musician ensembles (trios and quatuors) playing 6 extracts of musical pieces comprising Baroque and Jazz music, including the following composers: Bach, Vivaldi, Django Reinhardt, Duke Ellington, and Sydney Bechet. Close-mounted microphones were used to record each instrument individually in a methodical manner. Accompanying simultaneous RGB-depth-video recordings enabled for the creation of 3D point-clouds which can be easily incorporated into future virtual acoustic simulations combined with auralizations.

As introduced in Sec. 3.4, dynamic instrument directivity can be included in the simulations. This first perspective will allow for the conduction of subjective experiments to assess the impact of the incorporation of such dynamic instrument directivity on auralization perceptions, with a comparison to results previously obtained with voice stimuli [15]. Other potential uses of this multimodal dataset include investigating the impact of changing the spatial arrangement of musicians and instruments to simulate different orchestrations, or providing coherent and incoherent audio/visual cues to investigate multimodal effects, in particular for the auditory perception of dynamic source directivity.

5 Resources

Synchronized audio, 3D-video point-cloud, and source tracking information are made freely available for research purposes. Files can be downloaded via the institutional site `www.lam.jussieu.fr/Projets/AVAD-VR.html`.

ACKNOWLEDGEMENTS

The authors acknowledge the musicians for their participation in the creation of this database.

REFERENCES

[1] Vorländer, M. Computer simulations in room acoustics: concepts and uncertainties. J. Acoust. Soc. Am., Vol 133 (3), 2013, 1203-1213.

[2] Postma, B.; Katz, B. Perceptive and objective evaluation of calibrated room acoustic simulation auralizations. J. Acoust. Soc. Am., Vol 140, 2016, 4326-4337.

[3] A. N. Burd. Nachhallfreie Musik fur Akustische Modelluntersuchungen. Rundfunktech, Mitteilungen, Vol 13, 1969, 200-201.

[4] Japan Audio Society, GES 9080. Various - Impact, CD compilation, 1985.

[5] Hansen, V.; Munch, G.. Making recordings for simulation tests in the Archimedes project. J. Audio Eng. Soc., Vol 39 (10), 1991, 768-774.

[6] Anazawa, T.; Inokuchi, K.; Clegg, A. H. An Anechoic Digital Recording of Orchestral Music. Denon, PH 6006, 1991.

[7] Vigeant, M.; Wang, L.; Rindel, J.; Christensen, C.; Gade, A. Multi-channel orchestral anechoic recordings for auralizations. ISRA, 2010.

[8] Freiheit, R.; Alexander, J.; Ferguson, J. Making an anechoic choral recording, J. Acoust. Soc. Am., Vol 118, 2020-2021.

[9] Pätynen, J.; Pulkki, V.; Lokki, T. Anechoic recording system for symphony orchestra, Acta Acustica United with Acustica, Vol 94 (6).

[10] D'Orazio, D.; De Cesaris, S.; Massimo, G. Recordings of Italian opera orchestra and soloists in a silent room. Proc. of Meetings on Acoustics, Vol 28 (1), 2016.

[11] Poirier-Quinot, D.; Postma, B.; Katz, B. Augmented auralization: Complimenting auralizations with immersive virtual reality technologies. ISMRA, 2016, 1-10.

[12] Katz, B.; Postma, B.; Poirier-Quinot, D.; Meyer, J. Experience with a virtual reality auralization of Notre-Dame Cathedral; J. Acoust. Soc. Am., Vol 5, 2017, 3454-3454.

[13] Pelzer, S.; Vorländer, M. Auralization of virtual rooms in real rooms using multi-channel loudspeaker reproduction. J. Acoust. Soc. Am, Vol 134 (5), 3985-3985.

[14] Postma, B.; Katz, B. The influence of visual distance on the room-acoustic experience of auralizations. J. Acoust. Soc. Am, Vol 142 (5), 2017.

[15] Postma, B.; Katz, B. Subjective Evaluation of Dynamic Voice Directivity for Auralizations. Acta Acustica united with Acustica, Vol 103 (2), 2017, 181-184.

[16] Palacino, J.; Paquier, M.; Koehl, V.; Changenet, F.; Corteel, E. Assessment of the impact of spatial audio-visual coherence on source unmasking - Preliminary discrimination test. AES Conv 140, 2016.

[17] Kowalski, M.; Naruniec, J.; Daniluk, M. Livescan3D: A Fast and Inexpensive 3D Data Acquisition System for Multiple Kinect v2 Sensors. Int. Conf. on 3D Vision, Vol 1, 2015, 318-325.

[18] F. Otondo; J. H Rindel, G. A New Method for the Radiation Representation of Musical Instruments in Auralizations. Acta Acustica with Acustica, Vol 91 (5), 2005, 902-906.

[19] Owinski, B. The recording engineer's handbook, 4th edition, BOMG Publishing, 2017.

[20] Meyer, J. The Sound of the Orchestra. J. Audio Eng. Soc, Vol 41 (4), 1993, 203-213.

[21] Katz, B.; Le Conte, S; Stitt, P.; EVAA: A platform for Experimental Virtual Archeological-Acoustics to study the influence of performance space. Intl Sym on Room Acoustics (ISRA), 2019.

[22] Thery, D.; Poirier-Quinot, D.; Postma, B.; Katz, B. Impact of the Visual Rendering System on Subjective Auralization Assessment in VR. EuroVR2017, 105-118.

[23] Katz, B.F.G.; Poirier-Quinot, D.; Postma, B.N.J. Virtual reconstructions of the Théâtre de l'Athénée for archeoacoustic study. Intl Cong on Acoustics (ICA), 2019, 1-8.