Online Access and Visualization of Enriched Multimodal Representations of Music Performance Recordings: the Quartet Dataset and the Repovizz System

Esteban Maestre[1,2,3]
Panos Papiotis[1]
Marco Marchini[1]
Quim Llimona[1,2]
Oscar Mayor[1]
Alfonso Perez[1,2]
Marcelo M. Wanderley[2]

[1]Music Technology Group, Universitat Pompeu Fabra, SPAIN
[2]Center for Interdisciplinary Research in Music Media and Technology, McGill University, CANADA
[3]Center for Computer Research in Music and Acoustics, Stanford University, USA

Abstract

In this paper, we provide a first-person outlook on the technical challenges and developments involved in the recording, analysis, archiving, and cloud-based interchange of multimodal string quartet performance data as part of a collaborative research project on ensemble music making. In order to facilitate the sharing of our own collection of multimodal recordings and extracted descriptors and annotations, we developed a hosting platform and data archival protocol through which multimodal data (audio, video, motion capture, descriptor signals) can be stored, visualized, annotated, and selectively retrieved via a web interface and a dedicated API. By way of this paper we make a twofold contribution: (a) we open our collection of enriched multimodal datasets to the community, the *Quartet Dataset*; and (b) we introduce and enable access to our multimodal data exchange platform, the *Repovizz* system, through which users can upload recorded data, and navigate, playback, or edit existing datasets via a standard Internet browser.

1. Introduction

The importance of music as a socio-cultural instrument in human expression and communication is undeniable. Ever since ancient times, music has played a central role in the development of human cognition, social interaction, and scientific knowledge. After the industrial revolution, with the advent of sound transmission and recording technologies, the democratization of access to music media has cemented one of the major entertainment industries in today's art-consuming society.

Most recently, the consolidation of digital technologies and the World Wide Web as the vehicle for information exchange and media delivery is changing the way in which music is consumed. Examples are on-demand music streaming, user-generated content, automatic music recommendation, etc. In that regard, the richness and complexity offered by music performance as an exploitable multimodal artifact for media consumption and digital humanities research offers a unique opportunity to further develop the paradigms of music performance representation, archiving, and retrieval.

Beyond its value in opening pathways for music entertainment, the inherent interdisciplinary character of music performance research attracts interest from diverse science and technology disciplines. From virtual reality applied to technology-enhanced music learning, or neuroscience and sensory-motor cognition research applied to music therapy, the increasing affordability of media recording technology besides the *traditional* audio and video modalities (e.g. motion capture and analysis, physiological signal processing, etc.) should be matched by improved means for documented archiving and remote visualization, exchange and retrieval of recorded data. For such wealth of data to be exploited in both entertainment and research contexts, the scientific community must first devote efforts to devising frameworks that support flexible exchange and enhanced browsing of music performance multimodal data collections in a way that allows showcasing research results and enables research reproducibility. This endeavour, which involves the exercise of expanded data acquisition and description, and remote data access and visualization, needs of daring, early-stage research pursuits that serve as an inspiration for the members of the scientific community to embark in the design of new tools for interactive, web-based consumption of multimodal representations of music performance recordings from which both collaborative research and entertainment-focused applications would undoubtedly benefit.

1.1 Multimodal description of music performance recordings

Content-based computational description of music audio signals has now become an established, interdisciplinary research area with presence in a wide variety of applications of signal processing, machine learning, psychology, and musicology. Spinning off from the more traditional field of speech signal analysis, *Music Information Retrieval* (MIR) [1] is today a field of research which, albeit successfully brought into play in many real-world contexts, mostly focuses on using audio signals and symbolic score information as data sources, leaving other modalities aside. A clear reason behind this limitation is the fact that much of today's MIR research is developed around existing music collections for which only audio recordings are available. Only in some contexts, often when targeting music performance as the object of study (e.g., expression modeling, analysis of playing technique, sensorimotor interaction, new interfaces for musical expression) [2], MIR starts to incorporate other modalities such as gesture analysis via motion capture and description [3], or physiological signals acquired during music playing [4].

In general, the process of computational description of music performance beyond audio-only analysis does not fundamentally change its nature: acquired data streams are processed (perhaps combining different modalities) to obtain new descriptor or annotations streams which then become part of a multimodal representation that includes media streams, descriptor signals, annotation streams, etc. However, as the number and nature of modalities and extracted descriptors and annotations increase, difficulties arise not only in adopting formatting schemes that allow the interchange of complex, multi-modal data [5], but also in finding effective platforms for remote data visualization and retrieval that make possible hosting, previewing, and downloading large sets of multimodal data made available either for consumption either in research or entertainment contexts.

1.2 (Online) consumption of multimodal representations of music performance recordings

Even for low-complexity, audio-only descriptions of collections of music recordings, it is still difficult to find accessible platforms for flexible, remote data hosting that allow effective previewing and retrieval. Besides entertainment-focused web applications like *youtube* or *soundcloud* offering single media track previewing (or audio/video) and text annotations, one finds other research-oriented platforms like *freesound*, the *echonest*, or *acousticbrainz* which provide means for computation of audio descriptors, tag-based search and previewing of audio streams, but.offer limited flexibility for remotely previewing extracted descriptors or annotations. Beyond audio, it is difficult to find online data exchange frameworks that offer rich interaction mechanisms. A clear example is motion capture data, an important music performance modality of increasing popularity which has been extensively used in a number of biomechanics and behavioral studies: one finds a number of web-accessible motion capture databases for which interaction is limited to download links [6] and no preview is available, so

it is often needed to download the data to a local drive and open it with specialized software in order to start working with it.

A possible scenario for online open access to multimodal representations of music performance recordings is illustrated in Figure 1. In it, acquired media streams and extracted descriptors and annotations can be pushed to a content repository, which holds indexed data streams that are made available for download, reprocessing, and re-upload, and are also accessible through enriched browsing mechanisms that enable data preview and media streaming as a means for open, effective navigation and interaction with data.

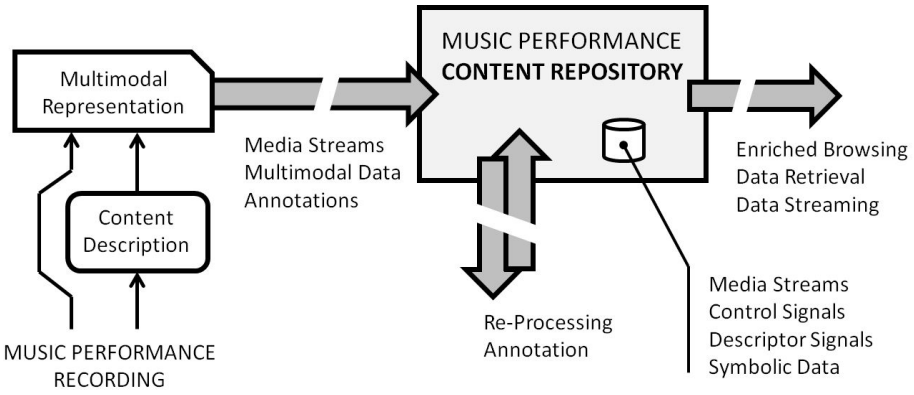

Figure 1. Ecosystem for online access to multimodal representations of music performance recordings.

1.3 Context and objectives of this work

Conforming an initiative started by Universitat Pompeu Fabra (UPF) in 2006 through research on how the joint analysis and modeling of violin gestures and audio attributes can be used to propose new sound synthesis paradigms, the work presented in this paper is the outcome of a continued, joint effort of a number of academic institutions traditionally involved in data-driven research on music performance. After a first successful attempt at carrying out multimodal analysis of solo violin performances in the context of expressive sound synthesis within an industrial project [7, 8] that lasted until 2009, the focus was expanded to the case of multimodal description and analysis of ensemble music performance via public funding and an extended collaboration between UPF, McGill University, and Stanford University, leading to pioneering contributions to the field [9, 10].

WIthin this collaborative research project involving the capture, exchange, and analysis multimodal data recorded from string quartet performances (the Quartet Project, see Section 2), several challenges had to be tackled around the topics of multimodal acquisition, description, archival, and online access to music performance recording data in the context of

analyzing joint music making. With the aim of easing collaborative research around our data, we developed a remote hosting platform and data archival protocol through which data of different modalities can be stored, visualized, annotated, and selectively retrieved via a web interface and a dedicated API so that users can upload recorded data, and navigate, playback, or edit existing datasets by means of a standard internet browser through which enriched visualization and reproduction means are provided.

Although early versions of the multimodal data exchange and remote visualization platform presented here have been partly introduced in a number of symposiums [11], this work provides a more comprehensive perspective and use-case description through which our methodology for music performance multimodal data recording, archival, and web delivery is also validated. In that regard, it is important to remark that through this publication we are opening our complete set of string quartet annotated multimodal data to the research community — not only we are sharing our data, but also providing a hosting system and web application for other researchers to either access our data or publish their own data.

1.4 Organization of this paper

The paper is organized as follows. Section 2 provides an overview of the research motivations and technical challenges of the Quartet Project, and how it led to the development of the dataset and data exchange framework presented in this paper. Section 3 outlines the technology and setup used to perform multi-modal data recording of string quartet performances, while Section 4 describes the techniques used for the description, segmentation, and annotation of recorded multi-modal data. Section 5 is devoted to describing the contents and structure of the multimodal dataset of string quartet performance recordings: the Quartet Dataset. Following, in Section 6 we describe the Repovizz system for remote hosting and web-based visualization and delivery of annotated multimodal representations of music performance recordings, and how the Quartet Dataset can be navigated. Finally, in Section 7 we reflect on the relevance of this work for the research community and beyond.

2. The Quartet Project

The Quartet Project is a long-term research endeavor to characterize and understand ensemble music performance and playing technique through the recording, description, and annotation of string quartet performances. Led by the Music Technology Group of UPF and involving a number of prestigious international partners, the project departed from early results obtained by UPF researchers when working on multimodal analysis of violin performance

(audio, video, motion capture, sensors) to develop new techniques for string sound manipulation and synthesis.

Since then, following successful results in understanding and modeling violin performance, researchers at UPF obtained a number of research grants and collaborated with other institutions (e.g., McGill University, Stanford University) to focus on developing a methodological framework for the recording, analysis, annotation, archival, and enriched web-based retrieval and reproduction of detailed multimodal recordings of string quartet performances. Today, the results of this project are assorted: (i) the methodological framework itself; (ii) pioneering research results on ensemble music performance analysis and modeling, including 2 PhD theses and numerous research articles; (iii) an annotated multimodal dataset of string quartet performance recordings made available to the research community in this paper(the Quartet dataset); and (iv) a system for remote storage and web-based enriched consumption of multimodal representations of music performance data (the Repovizz system).

Ensemble performance and the skills associated with it encompass an undeniably significant subset of music performance overall, as collective music making is omnipresent from music education to both professional and amateur performance. Nevertheless, the amount of empirical research carried out on musical ensembles is very limited compared to research on music performance in a broader sense. As its name suggests, the Quartet project directs its methodological focus on the string quartet ensemble, a group of four musicians (two violins, viola, cello) common in western classical music practice. String quartets offer a few advantages as subjects of empirical research; first, most performances are based on an underlying score which allows for an objective analysis of the performance. Second, their relatively small size and lack of a de facto leader makes the analysis of musical coordination feasible yet sufficiently complex. Finally, the fact that bowed string instruments produce sound and pitch in a continuous manner offers rich and interesting performance data for prospective analyses in multiple data modalities (sound, movement, video, etc).

Our empirical study of multimodal string quartet performance data has been motivated by three main objectives: studying the interdependence between the members of a string quartet ensemble, training models of joint expressivity in ensemble music performance, and compiling/sharing research datasets that serve as both ground truth as well as a starting point for further research on music ensemble performance [10, 9, 12, 13]. Either in accomplishing our own research goals, making the data available to other researchers, or in developing flexible platform for open multimodal data exchange, a number of technical and logistical challenges have been faced in this project. This paper attempts at giving an overview of how we chose to overcome those challenges, related to the synchronized recording of multimodal music performance data using a number of different devices, the processing and segmentation of acquired data and computed descriptors, the design of remote data storage and cloud-based accessibility mechanisms for collaborative research, or the development of a platform for flexible, web-based browsing engine with extended capabilities for visualization and streaming of multimodal data.

3. Data recording

String quartet performances were recorded via synchronous acquisition and pre-processing of raw data pertaining to diverse modalities: audio streams captured using ambient microphones as well as contact microphones placed in the bridge of the instruments, adding up to 8 input channels of audio; high-quality video footage including audio-based timecode synchronization; and motion capture data using two different high-end systems, one wired leading to 8 streams of 6DOF markers and one wireless leading to 140 streams of 3DOF markers. A schematic illustration of the technical setup can be seen in Figure 2; in it, four differentiated entities are shown: (a) a synchronization signal generator; (b) a desktop computer devoted to the synchronous acquisition of audio signals via a dedicated audio interface as well as the acquisition and processing of motion capture data using a wired Electromagnetic Field (EMF) -based system, (c) a desktop computer dedicated to the acquisition and processing of motion capture using an infrared (IR) camera-based motion capture system, and (d) a video camera for video footage. Next we provide further details on the recorded modalities and signal synchronization.

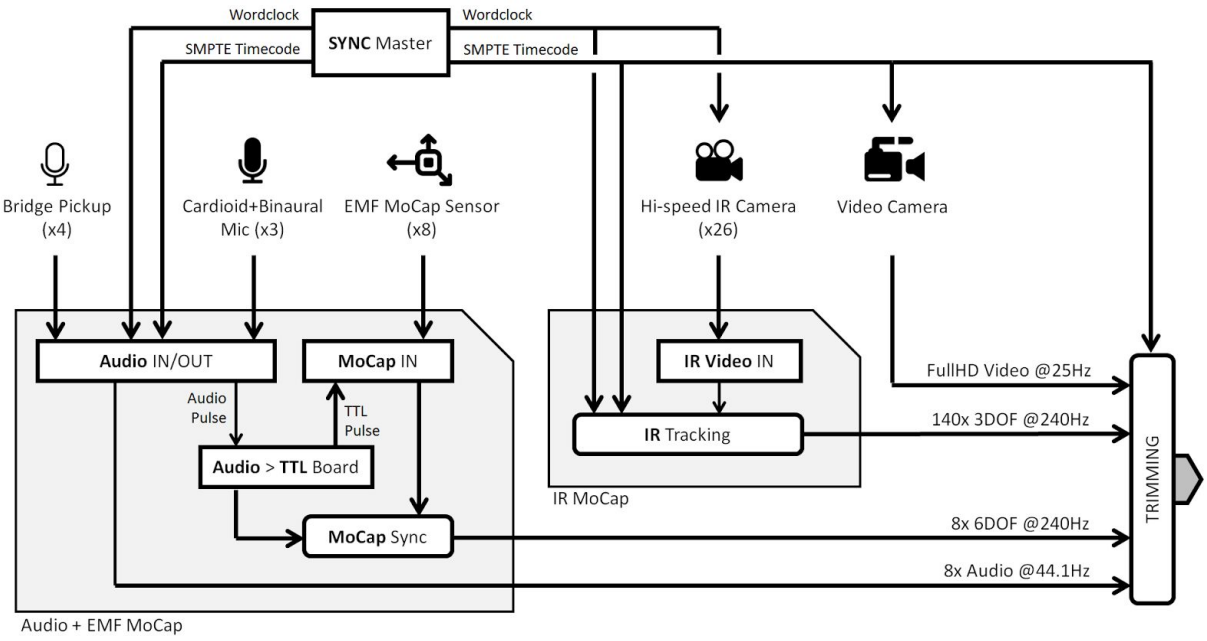

Figure 2. Schematic illustration of the data recording setup

3.1 Audio and video capture

Both individual audio from each performer as well as ambient audio was captured in order to study the performance in terms of the produced sound. An individual audio signal from each performer was acquired using a piezoelectric pickup fitted on the bridge of each instrument. We used Fishman V-100 pickups for the violins and viola, and a Fishman C-100 pickup for the cello. All recordings were carried out using a sampling rate of 44.1 Khz. The use of pickup microphones holds the advantage of capturing solely the sound of the instrument it is attached on and removing any sonic interference from the rest of the instruments or the room ambiance. Pickup gains were manually set with the aid of level meters in the recording equipment to avoid clipping of individual audios. To balance the sound level of the quartet as a whole, we employed an iterative procedure, based on listening to the mix of the pickups on the headphones while adjusting the four gains.

Ambient sound - i.e. the overall sound produced by the musicians - was recorded using two different devices, one monophonic and one stereophonic. For monophonic audio, we used a Neumann KM184 cardioid condenser microphone sampled at 44.1 KHz. The placement of the microphone was approximately 2 meters away from the centroid of the semicircle defined by a quartet's typical seating arrangement, in the direction faced by the quartet. For stereophonic sound, we used a binaural dummy head (Neumann KU-100) placed directly below the condenser microphone, facing the quartet.

High quality video was acquired with a Canon VIXIA HF R200 camera; the audio track of the video was fed with a SMPTE linear timecode audio signal, used as a synchronization index to temporally align the different data streams.

3.2 Motion capture

We used two different motion capture systems simultaneously to record each performance, a wired system based on electromagnetic field (EMF) sensing, as well as a wireless system based on optical tracking of infrared light-reflective markers. The systems capture two types of movement: bowing motion via wired sensors and markers attached to the body and bow of each instrument, and upper body motion via markers attached on the performers.

The first of the two motion capture systems used in our recordings is the Polhemus Liberty system, a 6DOF tracking system based on EMF sensing. It consists of a transmitting source, and a set of receiving wired spheric sensors with a radius as low as 0.5 cm and a weight down to 6 grams. Each sensor provides 3DOF for translation and 3DOF for rotation at a 240 Hz sampling rate, with translation and rotation static accuracies of 0.75mm and 0.15 degrees RMS respectively within a range of 4m of distance to the source. This method uses two sensors per tracked instrument: one wired sensor fixed on the back plate of the violin in order to track its position and orientation, and another sensor fixed on the bow near the frog. Each of

the sensors tracks a moving coordinate system, and interest points in the instruments (i.e., string ends or hair ribbon ends) are inferred as relative to such coordinate system. More details on the measurement procedure can be found in [8].

The second motion capture system used in our experiment is developed by Qualisys Inc. It is composed of 32 Oqus 400 infrared cameras, each one capable of 3 MegaPixels of resolution and of a maximum of 485 frames per second. We spatially arranged the cameras in a way that minimized marker occlusion through trial and error. The focus and aperture of each camera lens was manually adjusted, together with the exposure time, to match the specific condition of each camera due to the non-uniform lighting of the room where the system was installed. The infrared camera-based system configured to also work at a sample rate of 240Hz and used to capture both instrumental gestures as well as upper body movement. Regarding the instrumental gestures we used the marker setup reported in [14, 15, 9, 10]; the reason behind using two different motion capture systems simultaneously to acquire instrumental gestures was to perform future analysis on their comparative accuracies, and to provide complementary measurements. Regarding upper body movement, we used the standard marker placement protocol proposed in [16].

3.3 Synchronization

Several different devices were used to record the acquired data. A laptop computer with an external audio interface (RME Fireface 800) was used for both audio as well as wired motion capture data, a desktop computer was used for optical motion capture data, and a video camera for the high-quality video. In order to minimize the effects of drift and/or jitter in the digital sampling clocks of the various recording devices, a "master" external clock signal source (Rosendahl Nanosyncs HD) was sent to both computers to control their respective sampling clocks. Additionally, a linear timecode signal (SMPTE) was sent to the two computers as well as the video camera, in order to associate each captured frame to a common time reference for latter alignment and trimming.

The Qualysis system (used for optical motion tracking) allows for synchronous operation from an external clock and timecode source. However, the Polhemus Liberty system (used for wired motion tracking) only offers the possibility of detecting incoming TTL pulses (fed into its sync-in pin) and inserting a flag in the most recent data frame. Using custom recording software developed in-house, the audio interface connected to the laptop computer would periodically send an audio pulse to a custom-built analog conversion board that transforms it to a TTL pulse, which was then fed into the Polhemus sync-in pin. Audio-Polhemus synchronization was carried out in real time, measuring the delay between the two signal acquisition devices by counting the received audio samples from the time the pulse was sent by the audio I/O interface, until the time the converted TTL pulse is detected as a data flag within the incoming Polhemus data frames. Fluctuations in the delay were compensated for by either discarding or repeating incoming data frames. Qualysis 6DOF data and Polhemus 6DOF data are then processed and stored with a common reference time.

4. Performance Description

4.1 Audio description

For each of the individual pickup audio signals, as well as for the ambient audio signal, a number of audio descriptors were extracted by using the Essentia Library for audio analysis [17]. Essentia is an open-source C++ library for audio analysis that provides routines to compute more than 200 audio descriptors by both spectral and temporal domain analysis. By means of python binders and a dedicated processing module embedded in our data archiving system, the analysis of every audio stream led to the computation of a total of 23 associated audio descriptors of different nature: first, a number of low-level descriptors providing numerical estimations for instantaneous energy, fundamental frequency, dissonance, and various spectral shape descriptors; second, a number of tonal descriptors; finally, various tempo-related descriptors such as beat estimation or note onset times. For further reference on the exact audio descriptors extracted or on the actual implementation, please refer to the Quartet Dataset (Section 5) or the Essentia documentation.

4.2 Motion and gesture description

A first complete set of continuous bowing motion descriptors of each instrument were acquired using the two 6DOF sensors of the Polhemus system. As introduced before, the first sensor was attached to the violin back plate, in the upper bout, at the neck edge. The second one was attached to the bow wood, close to its center of gravity. From the data provided by these sensors, a set of motion descriptors were extracted as outlined in this section. For a more detailed description of the procedure for obtaining relevant bowing motion parameter streams, refer to [8].

Initially, a calibration of the string and hair ribbon ends is performed. The exact position of the eight string ends (four at the nut and four at the bridge) can be tracked by previously annotating (during calibration) their position relative to the coordinate system defined by the position and orientation of the 6DOF sensor placed in the violin. In a similar manner, the positions of the two hair ribbon ends (at the frog and at the tip) are estimated from the position and orientation of the 6DOF sensor placed on the bow. Both the violin plane and the bow plane (respectively defined by their normal vectors) are estimated. The former is estimated from the eight string ends, and the latter is estimated from both the sensor position and from the two hair ribbon ends. From the violin and bow planes and tracked positions, we estimate the bow transversal position, the bow-bridge distance, and bow transversal velocity as described in [8]. The force that the bow exerts on the string is inferred indirectly, by estimating the deflection of the hair ribbon as described in [18]. A second complete set of bowing motion descriptors, containing the exact same time series as described above, can be computed

following the same procedure but starting from the solid bodies defined by wireless reflective markers attached to the instrument bodies and the bows. More details can be found in [15]. Each of the extracted bowing motion descriptor was represented as a time series, as well as the position of the string endpoints and the hair ribbon endpoints.

With respect to the musician body motion, 3D position of every body marker was extracted and stored as a set of three time series. In addition, bone links were defined to join human body relevant upper body parts[3].

4.3 Segmentation and annotation

For each of the four musicians, note-by-note alignment between the performed score and the recorded data streams was carried out semi-automatically via a two-step process. First, audio-only descriptors (fundamental frequency and chroma vectors as extracted from the pickup audio signals) were used to perform a first alignment by employing the online dynamic time warping technique described in [19], and using a synthesized audio track as a target. The synthesized audio track was generated from the score-defined pitch and the body admittance impulse responses as measured from the instruments by means of a laser doppler vibrometer prior to the recordings. Then, a second alignment step was carried out by means of a Hidden Markov Model in which transition costs were computed by attending both to audio descriptors (energy, fundamental frequency, and harmonic-to-noise ratio) and to bowing gesture descriptors (bow velocity and bow force profiles). Finally, segmentations were manually revised in order to correct errors. To each note segment defined by its onset and offset times as extracted from the alignment procedure, an annotation was created in which the score pitch, bow direction, bowing articulation type, and performance-specific score annotations (see Section 5) were added as symbolic data. Further details on segmentation and annotation can be found in [9, 10, 8].

5. The Quartet Dataset

As outlined in Section 2, we aimed to capture a classical string quartet performing a series of musical tasks requiring the establishment of interdependence in order to achieve a shared goal. These tasks are divided in two groups: (i) exercises, where the shared musical goal (e.g. a common crescendo for all musicians) is explicitly stated along with the description of the exercise and the realization of that goal is feasible due to the simplicity of the exercise, and (ii) excerpts of musical pieces, where the shared musical goal is not stated but rather shaped throughout the rehearsals of the ensemble. The subjects in the main experiment were all members of a professional string quartet ensemble based in Montreal, Canada.

Regarding the exercises, materials were drawn from the handbook "Exercises for String Quartet" by Mogens Heimann [20]. We chose six sets of short exercises, with each set concentrating on one aspect of the performance at a time (intonation, dynamic shading, rhythm, timbre, unity of execution, and phrasing); nine exercises were recorded in total. Each exercise was accompanied by a set of instructions from the composer that were presented as the shared goal of the exercise. Additionally, a number of short excerpts of pieces by well-known western classical composers were selected and categorized according to the same six categories used for the exercises. Ten piece excerpts were recorded in total, selected from the recorded quartet's existing repertoire: Beethoven's *String quartet n.4 in C minor* (opus 18), Haydn's *String quartet n.3 in Eb major* (opus 71), and Borodin's *String quartet n.2 in D major*.

Recordings were centered around two conditions: *solo*, where each musician performs alone, and *ensemble*, where all musicians perform together as a group following a short rehearsal. In the solo condition, each musician performed their part alone using a stripped-down version of the score; meaning they had no access to the scores of the other musicians nor the textual instructions or annotations therein. Prior to each recording, we provided the musicians with four bars of a metronome click as an indication of the desired tempo, as well as a tuning reference of A=440Hz. Following the solo condition, the musicians were given the complete quartet score along with the instructions of the composer in terms of the shared goal. They were then left to rehearse the exercise together for a short period (10-15 min) or until they felt they were achieving the shared goal satisfyingly. We did not interfere with the musicians during this rehearsal period, or attempted to control the rehearsal progress. Finally, following rehearsal, the musicians were recorded in the ensemble condition; similar to the solo condition, they were provided with four bars of a metronome click immediately before recording as well as a tuning reference.

The above described recordings have been compiled into what is referred to here as the QUARTET dataset; a collection of 96 multimodal recordings (9 exercises and 10 pieces, in *solo* and *ensemble* experimental conditions) with a total duration of approximately 85 minutes, featuring audio, movement, and video streams and extracted descriptors and annotations for each performance. For a comprehensive description of the experimental conditions, contents, and structure of the dataset, the reader is referred to [10, 21].

6. The Repovizz System

As a response to our own research needs and to disseminate our data we developed Repovizz, a remote hosting platform and data archival protocol through which data of different modalities can be stored, visualized, annotated, and selectively retrieved via a web interface and a dedicated API. The Repovizz system is composed of an API endpoint, a Web server, a Transcoding module, a Processing module, and an Access control module, all of them connected to a central unit. Users can interact either programmatically through the API endpoint, or using the Web app served through the Web server. This is illustrated in Figure 3. An overview of what these modules enable will be presented next.

Repovizz provides a cloud-based environment where multimodal, time-aligned data streams can be easily indexed and archived into a database of numeric or symbolic data as stream collections, or *datapacks.* Each datapack usually holds streams corresponding to a single recording, and is represented in the resource database as a tree document. The tree is formed by nodes that can organize data, hold textual descriptions, or hold pointers to data files of different types.

An access control system ensures that data marked as closed can only be accessed by collaborators, while data published as open has no restrictions. There is a transcoding module that can automatically transcode most data streams using tools such as FFmpeg into lightweight formats and cut for a quick previsualization over the network. This is an especially relevant point, since multimodal data is often uploaded in very high quality formats that allow detailed analysis but make it unfeasible to easily browse the data even if it is available locally.

This transcoding engine powers a web server to which users can connect using any browser and use a visualization tool that allows previewing the vast collection of multimodal data in the repository from any web browser, downloading then only the required segments of multiple data streams in full quality. This is made possible by real-time web protocols such as Websockets. From within the web browser, users can easily download data from entire recordings if they know the specific recording they need; they can also open a transcoded preview, visualize its contents and then select the specific streams they wish to obtain. These streams, if properly time-aligned, can be automatically cut to within a selected region. A screenshot of the web client interface showcasing the data visualization and exploration capabilities of Repovizz for one of the experimental recordings can be seen in Figure 4.

The Repovizz browser-based data navigation and preview engine allows users to arrange and interact with a number of data visualization and media streaming widgets (e.g., multi-track audio waveform visualization, video playback, 3D motion capture rendering, etc.) purposely developed as HTML5 snippets delivered by the web server. Users can visually inspect data streams and edit and add metadata in-place in the form of time-stamped annotations. This way, preliminary data analyses can be performed without the need of downloading any data

nor locally installing additional software. The metadata formatting scheme, which is based on tabulated plain text, is used in the score note-level segmentations included in the Quartet Dataset (see Figure 4).

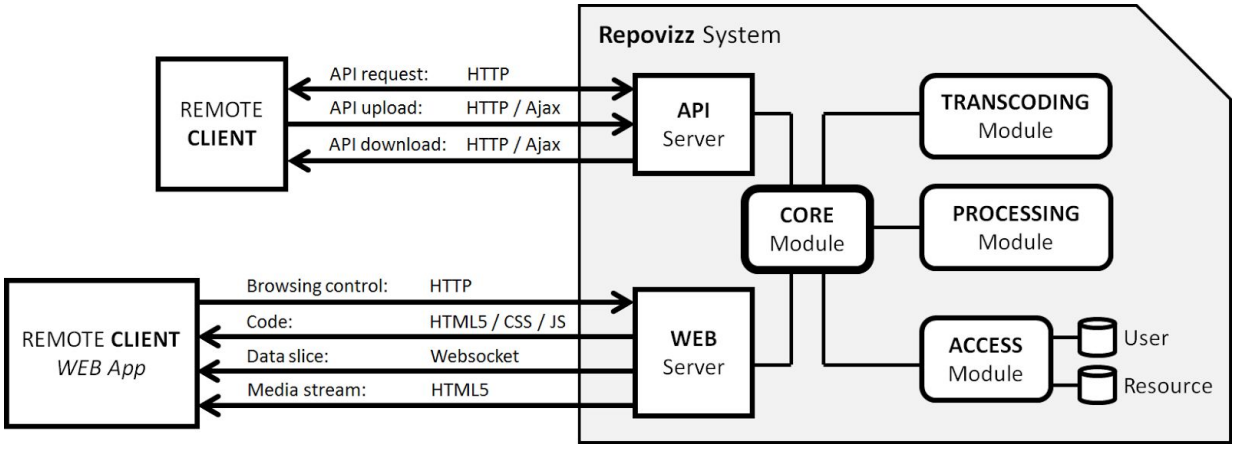

Figure 3. Schematic representation of the Repovizz system architecture

Since Repovizz is a cloud-based solution, it goes beyond on-demand data access and provides data *sharing* features, such as the generation of URLs with embedded access tokens that give access to protected resources or link to specific preview settings for a given data stream or dataset. Besides, a dedicated portal features a search engine that indexes data via diverse text identifiers, such as user-supplied tags, which aids data discovery.

To facilitate scripting for the automation of access to collections of data, Repovizz allows programmatic access to the system through its API endpoint. Using the HTTP-based REST API, clients can search for resources, fetch metadata, retrieve individual data streams, and download datasets for further inspection. This empowers Repovizz as a backend to support data-intensive applications or interactive data visualizations as companions to research articles. Furthermore, Repovizz enables server-side data processing: for data of some given modalities a processing module running feature extraction routines can automatically compute and index descriptors upon upload. As an example, all audio descriptors within the Quartet Dataset are computed automatically by Repovizz using the Essentia audio feature extraction library [17] (see Section 4).

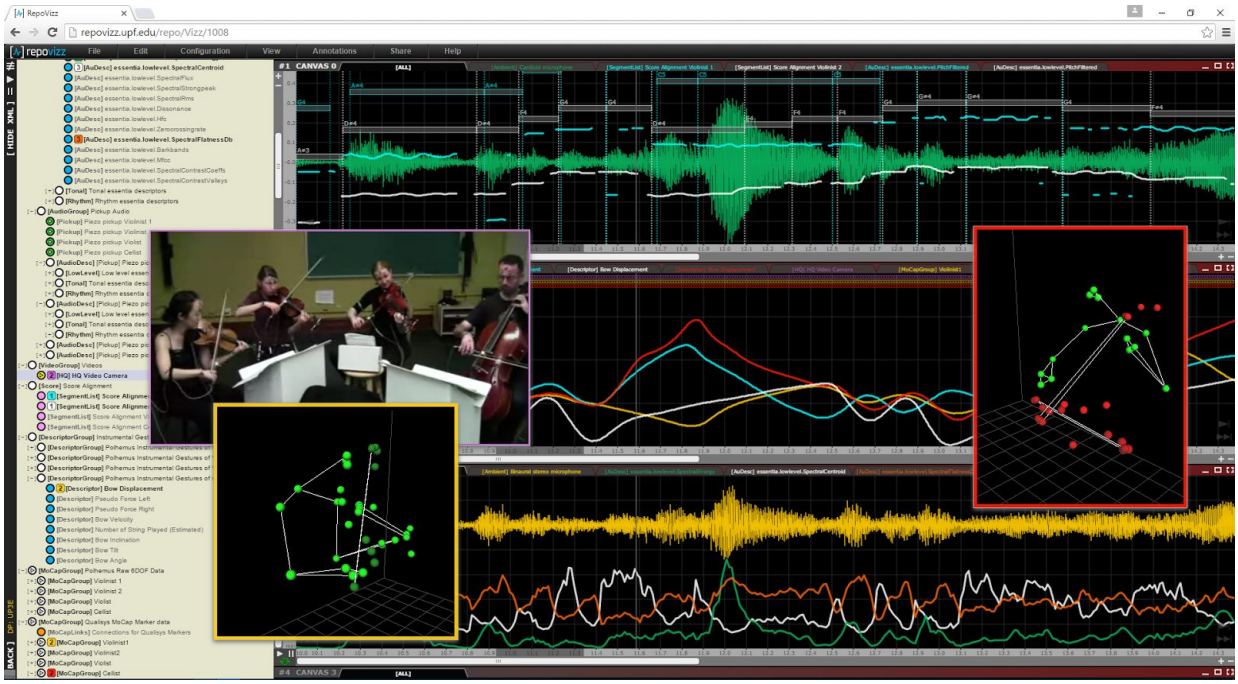

Figure 4. Screenshot of the Repovizz Web-based interface

6.1 The Quartet Dataset in Repovizz

The Quartet Dataset is a collection of multimodal representations of recorded string quartet performances, each stored as a datapack. Besides the raw streams corresponding to audio, video, and motion capture modalities (see Section 3), each datapack contains computed audio descriptors, extracted annotations and segmentations, rigid body motion descriptors, a set of derived bowing descriptors for each performer, as well as high quality video.

For each exercise and piece excerpt (see Section 5) five datapacks are provided: *Solo violin 1*, *Solo violin 2*, *Solo viola*, *Solo cello*, and *Ensemble*. Each datapack has a short-form title that serves as an identifier, with each letter or number in the title denoting a characteristic of the datapack as described in [10, 21]. For example, D2S1 corresponds to Dynamics exercise 2 Solo violin 1, and D2E corresponds to Dynamics exercise 2 Ensemble. The screenshot in Figure 4 displays the datapack UP3E, corresponding to Unity of Execution - Dynamics excerpt 3 Ensemble. Besides the short-form title, each datapack presents two text-based descriptions (about the recorded exercise/piece and about the datapack structure), and a list of keywords by which it is searchable. The data streams in each datapack are organized in 6 grouping nodes:

- Recorded audio from each of the utilized microphones (a cardioid and a binaural ambient microphones, and a pickup microphone mounted on the instrument bridge per performer), inside an *AudioGroup* node. Each audio recording is accompanied by the audio features extracted using the Essentia library in an adjacent *AudioDesc* node.
- Video stream, inside a *VideoGroup* node.

- Note onset and offset times for the audio of each instrument as they were obtained by the score-performance alignment, in a *Score* node.
- Instrumental gesture features for each instrument computed from data acquired using the wired motion capture system, in a *DescriptorGroup* node.
- Raw 6DOF (position and orientation) data acquired using the Polhemus wired motion capture system, in a *MoCapGroup* node.
- Upper body, instrument, and bow marker data acquired using the Qualisys optical motion capture system, in another separate *MoCapGroup* node.

All datapacks can be accessed online at [21] where further details are provided, or directly from the Repovizz portal. For more information please refer to [10].

7. Outlook

As the increasing affordability of data acquisition technology calls for an unfolding in the established paradigms of recording and analysis of music performances as multimodal digital artifacts, we should exploit the opportunity offered by the Internet as a medium for multimedia data delivery and exchange in both research and entertainment. This paper reports on an joint research initiative through which the acquisition, description, archival, and remote open access to multimodal music performance data has played a central role.

We first provided an overview of the research motivations and technical challenges involved in the acquisition and description of multimodal recordings of string quartet performances, leading to an annotated collection of multimodal data to which we open access by way of this paper. Both to support the access to our analyzed data and to facilitate similar data exchange initiatives by the research community and beyond, we developed the Repovizz system for cloud archival and web-based visualization and delivery of annotated multimodal representations of music performance recordings. Through Repovizz, users can upload recorded data, and navigate, playback, or edit existing datasets via a standard Internet browser.

Although the developments presented in this paper accommodated well our own collaborative research needs and are allowing us to share our data by means of a novel, cloud-based multimedia data exchange platform that we purposely developed, they only represent our attempt at expanding the possibilities for remote production-consumption of multimodal music performance data in both research and entertainment contexts. It is our hope that this work serves as an inspiration for others in proposing and developing future means for exploiting the exchange of music performance data in a digital library ecosystem that enables rich interaction with multimodal music resources and leads to devising new forms of music creation and consumption.

8. Acknowledgements

The work presented in this paper has been partly funded by the European Research Commission through the SIEMPRE FET-Open and PHENICX STREP projects, the Marie-Curie PEOPLE Program, and by the Catalan AGAUR Funding Agency through two Beatriu de Pinos Fellowships.

9. References

[1] Schedl, M., Gómez, E., and Urbano, J. (2014). Music Information Retrieval: Recent Developments and Applications. *Foundations and Trends in Information Retrieval*, 8(2– 3):127–261.

[2] Gabrielsson, A. (2003). Music performance research at the millennium. *Psychology of music*, 31(3):221-272.

[3] Godøy, R. I., and Jensenius, A. R. (2009). Body movement in Music Information Retrieval. *In 10th International Society for Music Information Retrieval Conference (ISMIR 2009)* (pp. 45–50).

[4] Jaimovich, J., Coghlan, N., and Knapp, R. B. (2012). Emotion in motion: A study of music and affective response. *In From Sounds to Music and Emotions* (pp. 19–43).

[5] Jensenius, A., Castagné, N., Camurri, A., Maestre, E., Malloch, J., and Mc Gilvray, D. (2007). A summary of formats for streaming and storing music-related movement and gesture data. In *4th International Conference on Enactive Interfaces 2007* (pp. 125–128).

[6] De la Torre, F., Hodgins, J., Bargteil, A., Martin, X., Macey, J., Collado, A., and Beltran, P. (2008). Guide to the carnegie mellon university multimodal activity (CMU-MMAC) database. *Robotics Institute*, 135.

[7] Perez-Carrillo, A. (2009). Enhacing spectral sintesis techniques with performance gestures using the violin as a case study. PhD thesis, Department of Information and Communication Technologies, Universitat Pompeu Fabra.

[8] Maestre, E. (2009). Modeling instrumental gestures: an analysis/synthesis framework for violin bowing. PhD thesis, Department of Information and Communication Technologies, Universitat Pompeu Fabra.

[9] Marchini, M. (2014). Analysis of Ensemble Expressive Performance in String Quartets: a Statistical and Machine Learning Approach. PhD thesis, Department of Information and Communication Technologies, Universitat Pompeu Fabra.

[10] Papiotis, P. (2016). A computational approach to studying interdependence in string quartet performance. PhD thesis, Department of Information and Communication Technologies, Universitat Pompeu Fabra.

[11] Mayor, O., Llimona, Q., Marchini, M., Papiotis, P., and Maestre, E. (2013). repoVizz: a framework for remote storage, browsing, annotation, and exchange of multi-modal data. In *21st ACM International Conference on Multimedia* (pp. 415–416).

[12] Papiotis, P., Marchini, M., Perez-Carrillo, A., and Maestre, E. (2014). Measuring ensemble interdependence in a string quartet through analysis of multidimensional performance data. *Frontiers in Psychology*, 5:963.

[13] Marchini, M., Ramirez, R., Papiotis, P., and Maestre, E. (2014). The sense of ensemble: a machine learning approach to expressive performance modelling in string quartets. *Journal of New Music Research*, *43*(3), 303–317.

[14] Schoonderwaldt, E. (2009). Mechanics and acoustics of violin bowing: Freedom, constraints and control in performance. PhD thesis, KTH Royal Institute of Technology.

[15] Llimona, Q. (2014). Bowing the violin: a case study for auditory-motor pattern modelling in the context of music performance. Bachelor's thesis, Department of Information and Communication Technologies, Universitat Pompeu Fabra.

[16] Input Devices and Music Interaction Laboratory. Plug-in-Gait Marker Placement. Available at: http://www.idmil.org/mocap/Plug-in-Gait+Marker+Placement.pdf

[17] Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J., and Serra, X. (2013). ESSENTIA: an Audio Analysis Library for Music Information Retrieval. In *International Society for Music Information Retrieval Conference (ISMIR 2013)* (pp. 493–498).

[18] Marchini, M., Papiotis, P., Perez-Carrillo, A., and Maestre, E. (2011). A Hair Ribbon Deflection Model for Low-intrusiveness Measurement of Bow Force in Violin Performance. In *NIME* (pp. 481–486).

[19] Dixon, S. (2005). Live tracking of musical performances using on-line time warping. In *8th International Conference on Digital Audio Effects* (pp. 92–97).

[20] Heimann, M. (1958). Exercises for String Quartet. (H. E. Deckert and F. Marcus, Eds.) (2007 ed.). European String Teachers Association (Denmark branch), ACMP Chamber Music Network. Available at: http://www.acmp.net/media/heimann_exercises.pdf

[21] Music Technology Group, Universitat Pompeu Fabra (2016). QUARTET dataset. Available at: http://mtg.upf.edu/download/datasets/quartet-dataset