EXPERIENCING REMOTE CLASSICAL MUSIC PERFORMANCE OVER LONG DISTANCE: A JACKTRIP CONCERT BETWEEN TWO CONTINENTS DURING THE PANDEMIC

MARINA BOSI[‡], **ANTONIO SERVETTI**[†], **CHRIS CHAFE**[‡], **AND CRISTINA ROTTONDI**[◇*]
(mab@ccrma.stanford.edu) (antonio.servetti@polito.it) (cc@ccrma.stanford.edu) (cristina.rottondi@polito.it)

[‡]*Center for Computer Research in Music and Acoustics (CCRMA), Stanford University, Stanford, California*
[†]*Dept. of Control and Computer Engineering, Politecnico di Torino, Turin, Italy*
[◇]*Dept. of Electronics and Telecommunications, Politecnico di Torino, Turin, Italy*

The recent lockdown restrictions imposed by the Sars-CoV-2 pandemic have heightened the need for new forms of remote collaboration for music schools, conservatories, musician ensembles and artists, each of which would benefit from being provided with adequate tools to make high-quality, live collaborative music in a distributed fashion. This paper demonstrates the usage of the Networked Music Performance (NMP) software JackTrip to support a distributed classical concert involving singers and musicians from four different locations in two continents, using readily available Hardware/Software (HW/SW) solutions and internet connections, while guaranteeing high-fidelity audio quality. We provide a description of the technical setup together with a numerical analysis of the achieved mouth-to-ear latency and an assessment of the music-making experience as perceived by the performers.

0 INTRODUCTION

The social distancing measures adopted to mitigate the propagation of the SARS-CoV-2 pandemic have reinforced the need to develop multimedia tools to support remote music performances and teaching of music-related subjects. A recent questionnaire about teaching methods adopted during the lockdown by 23 music schools belonging to the European Music School Union [1] highlights the difficulties encountered in delivering group lessons and supporting rehearsal sessions using standard e-learning technologies.

On the other hand, what is known as Networked Music Performance (NMP) software - and that has been used for many years mainly for research purposes and experimental music performances [2] - is designed specifically for real-time interactions in order to allow musicians to perform over the Internet, from different physical locations, as if they were in the same room. An important contribution to the spreading of this technology has been the availability of high-speed networks, at first through backbone networks reserved to research and educational institutions (such as Internet2) and then through the improvements of customer Internet connections with fibre-optic technology. However, although modern technology may satisfy the bit-rate requirements for NMP, further optimizations are still needed to improve the overall system (network & software) real-timeliness, i.e., end-to-end delay, to allow musicians to play from different locations with the least possible impact on their performance.

Indeed, the software solutions that have been mostly used during the lockdown for NMP, i.e., general-purpose teleconferencing applications currently available on the market, do not allow for real-time musical interactions and do not guarantee adequate quality of the audio signal. The primary design goal of such software is to maximize the intelligibility of the voice signal: it adopts low sampling frequencies and highly efficient compression schemes in order to limit bandwidth requirements, at the price of causing distortions of the audio signal generated by musical instruments [3]. In general, the encoding/decoding process optimizes overall data rate versus distortion and it usually introduces latencies that, though acceptable for a number of applications including teleconferencing (which typically tolerate end-to-end delays up to 150 ms [4]), would not be acceptable for real-time distributed musical interactions.

[*] To whom correspondence should be addressed e-mail: cristina.rottondi@polito.it

A few codecs have been specifically designed/adapted for ultra-low delay (i.e., an algorithmic delay of less than 10 ms) musical applications, e.g., [5], [6], [7]. Given its ultra-low delay and open source availability, many of the current NMP systems (e.g., [8], [9],[10]) utilize [7].

In the NMP context, during the last 20 years, the Center for Computer Research in Music and Acoustic (CCRMA) of Stanford University has been developing JackTrip [11], an open source software for ultra-low latency audio streaming. Assuming the availability of commensurate network bandwidth, computational resources and hardware provisions, it can theoretically support any number of channels of bidirectional, high quality, uncompressed audio signal streaming (e.g., digital audio streams at 48 kHz sampling rate / 16 bits per sample precision or higher) with no additional distortion or latency with respect to stereo CD audio quality. It also allows the user to control a number of parameters including the packet and buffer sizes, in order to optimize the trade-off between audio artifacts, bandwidth occupation and packetization/playback buffering latency, depending on the current network conditions. Moreover, it implements an extremely lightweight protocol stack (leveraging the User Datagram Protocol, UDP, as transport protocol) to minimize data processing delay on general-purpose processors.

This paper demonstrates the utilization of JackTrip to support audio streaming in a distributed classical music concert that took place in November 2020 within the events program of "Biennale Tecnologia," a biannual public festival organized by Politecnico di Torino and offered to the citizenry of Turin (Italy) that focuses on technology's decisive impact on every aspect of human life[1]. While JackTrip has been utilized in the past for distributed concerts with modern music repertoire and improvisation, experiences specifically dealing with classical repertoire are much less frequent, due to the intrinsic characteristics and interplay requirements of such musical genre, which makes it extremely challenging to be executed in presence of end-to-end delays above a few tens milliseconds [12]. Indeed, this classical music concert highlighted both the potential and the technical challenges of tackling such demanding, high-synchronization music selection in a distributed network performance with transnational, and even intercontinental connections. The musical program included pieces whose performance involved intricate rhythmic coordination which would be difficult given the longer-latency scenarios expected. Rubato and expressive timings were of particular interest as they might be ambiguous for musicians, who might confuse whether micro-timing inflections were intentional or artifacts of network delay. The concert involved twelve musicians (six singers and six instrumentalists), displaced in four different geographical locations: Politecnico di Torino (Italy), Ludwig-Maximilian Universität in Munich (Germany), two different households in the surroundings of Stanford University (California) and a household in the surrounding of New Haven (Connecticut). The concert program included a set of classical music pieces for singers and piano trio (piano, violin, and cello), instrumental pieces for violin/cello and piano and an improvisation for dilruba on top of a renaissance vocal music piece. Due to the lockdown restrictions in force at the time of the festival to counteract the spreading of the SARS-CoV-2 virus, no audience was allowed to enter the concert rooms in Turin and Munich and the event was streamed online. For video capturing and streaming, a commercial video-conferencing software was adopted (with muted audio): the audio streams transmitted via JackTrip and the video streams transmitted by the video-conferencing software were resynchronized prior to broadcasting to the audience. During rehearsals and the concert, extensive measurements on the experienced network and mouth-to-ear delay were collected, as well as subjective ratings and opinions of the performance experience from involved artists, in order to understand up to which extent they were able to cope with latency issues and what kind of delay compensation techniques they adopted. Such insights shed light on the musicians' preferences and can guide NMP software developers in devising future enhancements and new features, as well as provide suggestions to practitioners dealing with similar NMP setups.

The remainder of the manuscript is organized as follows. Section 1 briefly reviews the related literature and some existing solutions for NMP, whereas Section 2 provides insights on the impact of latency on networked musical interactions and on the JackTrip software. Section 3 includes a detailed description of the technical setup for audio-video streaming adopted during the performance and Section 4 discusses numerical measurement of network latency, jitter buffer sizing, and subjective rating of the performance experience. Recent JackTrip features and open technical challenges in NMP are reported in Section 5 and Section 6 concludes the paper.

1 RELATED WORK

1.1 HW/SW solutions for NMP

A number of hardware/software-based solutions for low-latency audio streaming aimed at networked applications are currently available. Table 1 compares several at the experimental or commercial stage. The interested reader can consult [13, 14, 8, 15, 16, 10, 9] for more details. One of the main advantages of using the JackTrip platform resides in the fact that it allows for ultra-low latency, uncompressed audio transmission within the boundaries of commonly accessible internet bandwidth (less than 800 kbps per audio channel at 48 kHz sampling rate and 16-bit sample precision[2]) using off-the-shelf, minimal hardware requirements (see also Section 3.2). Of the systems listed in Table 1, LOLA [15] allows for uncompressed only media signal

[1] Some excerpts of the concert are available at the link `https://www.youtube.com/watch?v=fgmc4Sdx1Mk`

[2] Note that bandwidth occupation is computed without considering possible overheads introduced by the adoption of Error Correction (FEC) techniques.

Table 1: Feature comparison for some of the currently available HW/SW solutions for NMP

	ELK Aloha [13]	Digital Stage* [14]	Jamulus [8]	LOLA [15]	JamKazam [16]	Soundjack [10]	JackTrip [11]	SonoBus [9]
Embedded systems support	✓	×	×	✓	(✓)	✓	✓	✓
Uncompressed audio	×	(✓)	×	✓	×	✓	✓	✓
Video streaming support	×	✓	×	✓	×	✓	×	×
Native broadcast functionality to external audience	×	(✓)	×	(✓)	(✓)	(✓)	×	×
Supported by commodity ISP	✓	✓	✓	(✓)	✓	✓	✓	✓

✓=supported; (✓)= partially supported; ×= not supported; * = under construction.

transmission, whereas Soundjack (developed on top of the basic component Soundjack Core) [10] and SonoBus [9] support both uncompressed and compressed audio streaming. LOLA has to be supported by very high-speed internet connection (between a minimum of 100 Mbps up to more than 2 Gbps, depending on the video specifications) and very strict Audio/Video (A/V) hardware requirements. However, most musicians working/rehearsing/preforming from home during the pandemic do not have access to such hardware equipment and stable, high-speed connections.

The benefit of uncompressed audio transmission over the internet is twofold. First, having uncompressed audio between the performers and the hub allows for high-quality audio interactions. It also provides a convenient method of avoiding cascading another codec in the last stage of the broadcast streaming process. Cascading codecs, i.e., the transmission channel applying a second low bitrate audio coding scheme to the audio signal, generally causes audible deterioration in the quality of the broadcasted signal (see, for example, [17]) and should be prevented, if at all possible. Second, the elimination of the encoding/decoding phases helps to reduce the audio processing delay and thus contributes to minimizing latency overheads on top of the unavoidable propagation delay.

1.2 Recent NMP demonstrations leveraging Jacktrip

A recent example of NMP can be found in the Quarantine Concert Sessions. The Quarantine Sessions are a series of telematic concerts of experimental electroacoustic improvisation that started at Stanford University's CCRMA in March 2020 during the first lockdown due to the Sars-CoV-2 pandemic (see Fig. 1). As part of this series, CCRMA has thus far live-streamed 62 one-hour sessions exploring different performance concepts (including free improvisation, graphical scores, text-based scores, and sound painting). Over twenty guest musicians and visual artists have participated from their homes in various countries, including the US, Canada, Ireland, Germany, Lithuania, Australia, and the UK. Free and open source technologies were utilized for these sessions, most notably, the use of JackTrip for the audio transmission of uncompressed, low-latency audio between the performers, Jitsi [18] for the corresponding video, and Open Broadcaster Software (OBS) [19] for combining and synchronizing both streams for live broadcast streaming (see also Sections 3.3 and 3.4).

Additional pre-pandemic examples of distributed musical performances are mentioned in [12]. Recent studies have also investigated the use of metronome-based solutions to help distant musicians remain synchronized during remote performances, e.g., [20, 21]. While the use of a global metronome seems to have helped the players' synchronicity in these studies, the musicians (for example, the cellist at Stanford and the violinist in New Haven) experiencing the highest latencies due to the physical distance from the hub in Torino, preferred to synchronize directly with the pianist in Torino without the help of a global metronome. They described the achieved synchronicity as a "learned skill" (both of them had previous experience performing chamber music over the internet). During the preliminary trials they learned to adapt and anticipate the beat by one eighth/sixteenth depending on the metronome marking of the piece and they played with the beat and

Fig. 1: Quarantine Session at Stanford's CCRMA.

music completely internalized while following the signal received from the hub.

2 BACKGROUND

2.1 The Impact of Latency on Networked Musical Performances

To allow musicians to maintain the necessary synchronism to play together, an extremely low end-to-end delay is required, ideally below 20-30 ms, which corresponds to the time required by sound propagation to cover approximately 7-10 m. This distance is typically considered as the maximum acceptable physical separation between musicians that allows synchronous musical interplay when performing in the same location, in the absence of further tempo references such as, for example, those provided by the gestures of a conductor. Above such threshold, latency becomes perceivable by the players and impacts the musical performance, typically leading to a tendency to tempo deceleration [12]. The tolerance level to the perceived delay may vary considerably, depending on the onset density and on the rhythmic complexity of the musical piece being performed [22, 12, 23]. For latencies above 60 ms, tempo deceleration becomes significant and may lead to unacceptable performance conditions. To counteract the effect of end-to-end delays above 30 ms, musicians must therefore develop and apply suitable delay compensation strategies, which highly depend on the tightness of the synchronization required by the performance and on the personal attitudes and roles (leader/follower) of the players within the ensemble.

Due to the different stages of the audio signal transmission [12], there are several contributions that add up to the end-to-end latency experienced by remotely located musicians:

1. the air propagation delay of sound waves from the emission source to the audio acquisition device (e.g., microphone) and from the sound reproduction device (e.g., loudspeaker) to the listener's ear;
2. the delay introduced by the audio acquisition/reproduction, processing, and packetization/depacketization at sender/receiver sides;
3. the pure propagation delay over the physical transmission medium;
4. the data processing delay introduced by the intermediate network nodes traversed by the audio data packets along their path from source to destination;
5. the playback buffering which might be required to compensate the effects of jitter in order to provide sufficiently low packet losses to ensure a target audio quality level.

The delay due to pure signal propagation cannot be avoided and it can be quantified at around 5 ms every 1000 km in optical fibers and copper twisted pairs. In order to reduce the overall latency impact as much as possible, however, it is possible to devise technological solutions capable of diminishing all the remaining delay contributions.

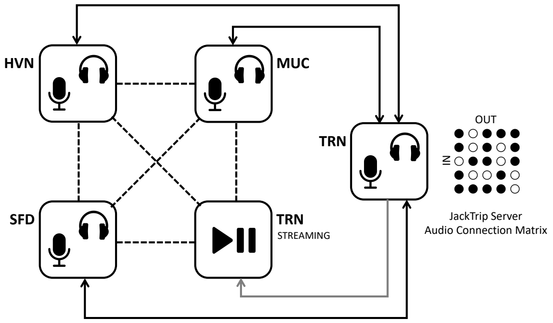

Fig. 2: Technical setup of the performance between Turin (TRN), Stanford (SFD), Munich (MUC), and New Haven (HVN). Solid lines correspond to client-server stereo audio connections, dashed lines correspond to peer-to-peer video connections. On the JackTrip server in Turin the audio is mixed and routed back to the participating nodes.

2.2 The JackTrip Software

JackTrip is a multi-machine technology which supports bi-directional flows of uncompressed audio over the internet while minimizing latency. Developed in the early 2000's, it was used in intercontinental telematic music concerts and a variety of musical experiments using high-speed research networks. Its ability to carry hundreds of channels simultaneously and its lightweight architecture led to a range of applications from IT for concert halls to small embedded systems. The pandemic has ushered in a new phase of development driven by musicians seeking solutions during lockdown. Major improvements have focused on ease of use and the ability to scale across worldwide cloud infrastructure. With orchestral-sized ensembles urgently in need of ways to rehearse on the network and most participants running their systems over commodity connections, this ”new reality” runs counter to what's required for ultra-low-latency rhythmic syncronization. Many developers and musical practitioners have joined in the cause of finding adequate solutions. JackTrip, which has generally been run as a native software application, is now complemented by dedicated solutions including numerous Raspberry Pi-based systems (see, e.g., [24]), graphical interfaces [25], standalone physical web devices, browser-based Web Real-Time Communication (WebRTC), see JackTrip WebRTC [26], and Pure Data [27] versions.

The newly established JackTrip Foundation [28] is a non-profit clearing house for open-source development, training, and support of partners and affiliates providing their own roll-outs of the technology.

3 TECHNICAL SETUP

The concert setup includes two separated environments - one for audio and one for video transmission - whose outputs are joined only before streaming to the passive audience, i.e., the concert listeners. Fig. 2 represents the A/V connections between the different geographical sites (i.e., Turin, Munich, Stanford and New Haven).

3.1 Preliminary trials

A set of preliminary rehearsal sessions took place a few months before the official concert date between Stanford/New Haven-Turin and Munich-Turin, with the twofold aim of making the musicians acquainted with the performance setup and for allowing the technical staff to explore and test different parameter configurations for the JackTrip software. For example, the Stanford-Turin synchronization offered a particularly challenging experiment due to the physical distance between the two sites (about 10000 km, corresponding to a 100 ms contribution to the round-trip time due to signal propagation over optical fiber). In the first set of rehearsals between Stanford and Turin, JackTrip was run with a buffer packet of 128 audio samples at 44.1 kHz sampling rate, a measured network round trip delay of around 166 ms and a overall systems delay of 190 ms. Based on the outcomes of such trials and on statistical measurements of the experienced end-to-end packet delay and jitter, the technical setup was finalized as described in the following subsections.

Moreover, during the rehearsals the musicians devised a range of latency-coping strategies, which varied across the performances. Such techniques involved some form of pre-agreed anticipation w.r.t. the perceived audio feedback for the musicians performing from Munich and Stanford/New Haven, to ensure that synchronism was maintained from the standpoint of the musicians performing in Turin, where the JackTrip server was located and from where audio and video signals were broadcasted to the audience.

3.2 Audio setup

The audio setup is based on a JackTrip instance that, at each location, is run on a local workstation and is connected to the instance running in Turin. In Turin JackTrip is run in *hub mode* as a server, while all the other remote locations are connected to it as clients. To limit the amount of bandwidth, an initial mix is performed at each local workstation in order to send to the hub only two audio channels (stereo mode). On the server, each pair of channels is mapped to a virtual audio device and processed by a Digital Audio Workstation (DAW) software (Ableton Live) that takes care of the overall mixing and of the audio effects (compression and reverb). Depending on the music piece, the impulse response of a small (chapel) or large (concert) hall has been applied to thicken up and add space to the recording by means of an Audio Ease Altiverb convolutional reverb plugin. Compression has been used to control the dynamic range and to prevent peaking by means of the Waves Linear Phase Multiband Compressor plugin (the threshold was set to -12 dB, the ratio to 3:1, and slow attack and release times were used) coupled with an Izotope Ozone Maximizer plugin with a target value of -14 LUFS (Loudness Units to Full Scale). The channel connection/mapping between the system channels and the JackTrip channels is performed by the *QjackCtl* [29] tool and exported by means of the *jmess* tool [30]. All the tools use JACK server [31] as their host audio server. From the JackTrip server in Turin, each remote location receives back a stereo mix of the audio signals gathered from the other locations. An additional JackTrip client is then used in Turin for streaming purposes: it is connected to the server and receives a single stereo mix of the audio signals gathered from the four locations.

3.3 Video setup

The video streaming is managed by a videoconferencing software. Each location has a local workstation that handles the local video cameras by means of an hardware video mixer. From that, a single webcam at a time is shared with the other participants. A simpler setup with just one webcam is used in the two households. The viewport captures the musicians' videos mainly for the aim of tailoring the final video streaming to be delivered to the audience. In fact, the huge latency introduced by the off-the-shelf videoconferencing software and its lack of synchronization with the audio make the video even detrimental to the musicians.

3.4 Broadcast Streaming setup

The broadcast streaming setup is based on OBS [19], that captures the A/V inputs, performs the A/V mix, encodes audio end video, and then uploads the stream to a streaming server hosted on TOP-IX, the TOrino Piemonte - Internet eXchange point. For streaming purposes the audio is encoded as a two-channel Advanced Audio Coding-Low Complexity (AAC LC) [32] stream at 320 kbps and the video as a 1920x1080 30 fps H.264 stream at 9 Mbps.

The video feed is exported from the videoconferencing software to OBS as a Network Device Interface (NDI) [33] stream. This allows to independently identify and acquire each video feed by means of a specific identifier that corresponds to the videoconferencing software user account. Other solutions can be used if the videoconferencing software is web based: for example, while using Jitsi Meet [18], a Google Chrome plugin [34] can be used to ”pop-out” each single camera in a separate window for OBS grabbing.

On the streaming workstation, the JackTrip audio feed is imported into OBS as an audio input via the *QjackCtl* tool. In OBS, a different *scene* can be defined for each musical piece where only the musicians that are currently performing are displayed with the proper graphical overlay designed for the musical event (see Figs. 3- 4).

Finally, the most challenging setup to be performed on the streaming workstation regards the A/V synchronization, since the video streams are received from the videoconferencing software with a delay much higher than the audio streams. This difference needs to be compensated by OBS adding a *sync offset* in the advanced audio properties of the auxiliary audio input used to inject the audio from JackTrip. Such an offset needs to be estimated before the beginning of the streaming of the performance as explained in Sec. 4.1.

3.5 Performance setup

When performing as soloists, the instrumentalists wore headphones (see Fig. 4), whereas the singers preferred to get acoustic feedback in only one ear, to better control their voice loudness (see Fig. 3, right). Thanks to this specific setup, singers were capable of monitoring both live sound from other local sources (e.g., in the duet) and audio signals received from JackTrip, thus experiencing different conditions compared to those experienced by the instrumentalists. For ensemble pieces, the members of the vocal group opted for the same acoustical setup, whereas the two strings players of the piano trio preferred not to wear headphones and to rely indirectly on the acoustic feedback received by the pianist by adjusting their playing to his tempo and dynamics (see Fig. 3, left). The audio capture setup in Turin and Munich was as follows. In Turin, two coincident microphones were placed at the center of the stage in front of the musicians for stereo recording, then each instrument was recorded with a directional condenser microphone placed in front of or above it. In Munich, the vocalists were arranged in a semicircle and three spaced microphones were placed in the center; soloists took advantage of the central microphone by stepping in front of it during their performance.

Concerning visual feedback, the piano trio in Turin considered them unnecessary and preferred not to take advantage of the videoconferencing software run in parallel to JackTrip. This choice was in part a consequence of the physical placement of the three players on the stage, with the two string players sitting in the front row and pianist located behind them (see left side of Fig. 3), which made it impossible to place a screen in a position visually accessible to the three of them at the same time. The three soloists at Stanford and New Haven received the video on their local laptop or smartphone but relied mostly on the JackTrip audio mix for synchronization. Also the singers in Munich had a large screen in front of them, which they could look at both during solos and ensemble pieces.

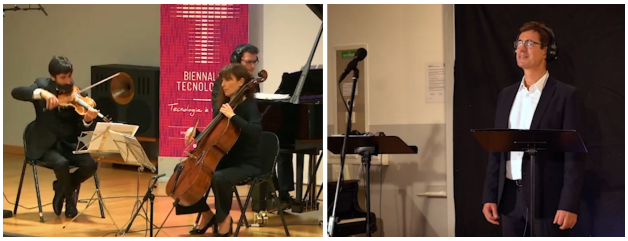

Fig. 3: Distributed concert, Turin-Munich session.

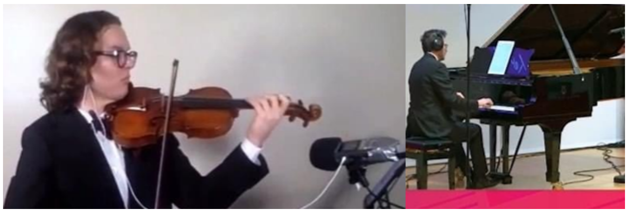

Fig. 4: Distributed concert, New Haven-Turin session.

4 MEASUREMENTS AND RATING OF THE PERFORMANCE EXPERIENCE

The concert repertoire consisted in a number of classical music pieces listed in Table 2. The Table also reports the instruments and voices involved, and the geographical locations the artists were performing from. Overall, the concert program accounted for around 45 minutes of live music.

4.1 Network and mouth-to-ear latency

During the concert rehearsals we performed some quantitative tests to setup the software configuration for the specific scenario.

After some preliminary trials, we set the JackTrip packet size to 512 samples and the buffer queue size to 4 packets. A relatively large size was needed to increase the robustness of the connections with Stanford (the distance Torino-Stanford being about 10000 km) and New Haven (distance Torino-New Haven about 6000 km). Tests with Munich (distance Torino-Munich about 600 km) showed that shorter packet sizes could have been used, e.g. 256 or 128 samples, were it the only connection to be handled during the performance. Network tests by means of the Internet Control Message Protocol (ICMP) ping utility across the Atlantic Ocean reported an average network round trip latency of 184 ms with Stanford, about 122 ms with New Haven, while the same measure with Munich reported about 27 ms.

At the same time, during the ping test, we measured the round-trip audio latency (i.e., the My Mouth to My Ear - MM2ME - latency) by means of the *jack_iodelay* tool [35] that emits an audio signal with several tones and computes the phase difference between that signal and the same signal after the round trip. To perform the test, we connected the *jack_iodelay* output to the JackTrip client input and the JackTrip client output back to the *jack_iodelay* input. Since the JackTrip server was looping back its own audio signal to the client, the *jack_iodelay* tool could compute the overall round trip latency (an integer multiple of the JACK server frame size, i.e., for a packet of 512 samples it corresponds to 11.6 ms at 44.1 kHz sampling rate).

Fig. 5 shows an excerpt of both the ping and audio round trip latency measurements between Turin and Stanford captured during the rehearsals. The minimum round trip audio latency value is 243.81 ms, thus a one way delay that corresponds to 121.90 ms, if the connection is considered symmetric. Lower latency was measured with Munich and New Haven, with a minimum round trip audio latency of 81.27 ms and 185.76 ms respectively. During the test reported in Fig. 5, JackTrip experienced different buffering delays as investigated in the following, thus we notice a variation of the MM2ME audio latency: 243.81, 255.42, 267.03, and 278.64 ms were reported respectively in the 31%, 2%, 21%, and 46% of the measurements.

Table 2: Concert Repertoire

Composer	Title	Voices (IDs)	Instruments (IDs)	Locations
Ludwig van Beethoven	Op. WoO 158a n° 16 "Schöne Minka, ich muss scheiden"	Soprano (S), Tenor (T1)	Piano (P), Violin (V1), Cello (C1)	Turin (instrumental trio), Munich (singers)
	Op. WoO 158a n°17 "Lilla Carl, sov sött i frid"	Soprano (S)		
	Op. WoO 158a n°23 "Da brava, Catina"	Tenor (T1)		
	Op. 108 n°2 "Sunset"	Tenor (T1)		
	Op. 108 n° 6 "Dim, dim is my eye"	Soprano (S)		
Arvo Pärt	Spiegel im Spiegel	-	Piano (P), Violin (V2)	Turin (pianist), New Haven (violinist)
Dmítrij Šostakóvič	Sonata For Cello and Piano, Op.40 - 3. Largo	-	Piano (P), Cello (C2)	Turin (pianist), Stanford (cellist)
Giovanni Pierluigi da Palestrina, Christopher Chafe	Improvisation on Lamentationes Jeremiae Prophetae (Lectionis In Feria Sexta Parasceve, Lectio II)	Soprano (S), 3 Tenors (T1,T2,T3), Baritone (B1), Bass (B2)	Dilruba (D)	Munich (vocal ensemble), Stanford (dilruba player)

To further investigate the different contributions that sum up to the final audio delay, we traced the length of the JackTrip queues in the audio round trip path: the transmission queue of the client, the reception and transmission queue of the server, and the reception queue of the client. The sum of the number of frames awaiting in each queue, as measured *after* a frame is extracted, can inform us about the overall buffering delay, i.e., the time that a frame spends in the queue. We sample the queue length value every time an audio frame is read, thus every 11.6 ms. If b is the number of frames in the buffer, we can assume that the buffering delay t_b is between $(b) \cdot t_f$ and $(b+1) \cdot t_f$, where t_f is the duration of a frame, thus $((b+1)+b)/2 \cdot t_f$ on average.

Fig. 6 shows the length of the client and server reception queues during the same test of Fig. 5. The dots represent the measured values at each sampling time, while the line is the moving average calculated over a sliding window of one second. The length of the transmission queues is not shown in the plot because it is always zero *after* the current audio frame is read and extracted. Measurements of the queue delay confirm that the delay introduced by buffering varies in time, from 0 to the maximum JackTrip queue size, because of variable packet delays and the clock skew between the client and the server instances. To better

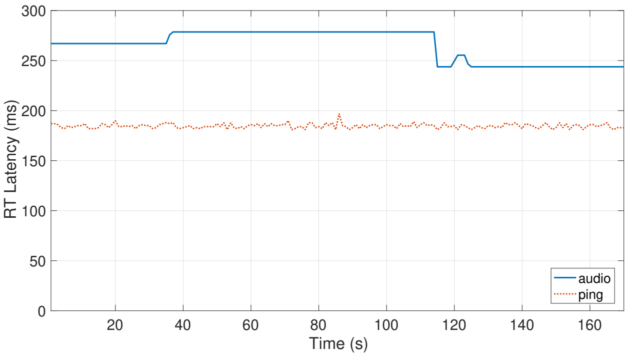

Fig. 5: Audio and network round trip latency measured with *jack_iodelay* and ping during a JackTrip connection between Turin and Stanford.

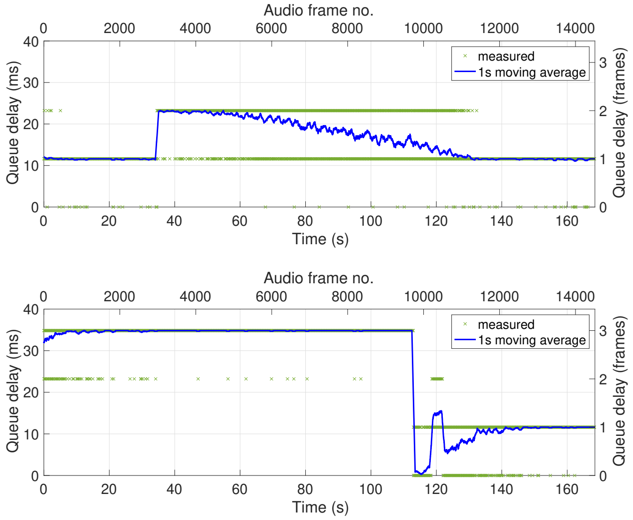

Fig. 6: Server (top) and client (bottom) JackTrip reception queue size/delay after an audio frame is read and extracted (i.e., every 11.6 ms). Measured value and moving average over a one second sliding window during a test connection between Stanford and Turin.

understand the impact of clock drifting on the buffer queue length, the interested reader is referred to [36].

Since the concert audio had to be recorded and streamed with the video of the musicians, we had to take care of the synchronization of both audio and video that were received with different delays, as they were transmitted with different software applications. The video, sent by means of a commercial videoconferencing tool, suffered from a higher delay than the audio, so we used OBS to add an additional delay to the audio stream. We measured the A/V delay with a test video that emits a beep every second and moves a bar on the screen, from left to right, over a series of equally spaced ticks that represent the time difference from the beep in steps of 0.1 s. The videoconferencing tool frame rate was 30 fps, thus we were able to measure the A/V delay with a precision of about $1/30$ of a second. On the connection between Munich and Turin we performed a one-minute test and measured an A/V delay in the range of 3-5 frames. This result is compatible with both the varying audio delay that we measured with JackTrip and the varying video delay that we assumed the videoconferencing software could introduce.

4.2 Rating of the performance experience

Rehearsals were organized in four sessions (Munich-Turin, Munich-Stanford, New Haven-Turin, Stanford-Turin), depending on the geographical locations of the artists performing the repertoire's pieces. Similarly to the approach adopted in [37], after each rehearsal session all the singers and instrumentalists were asked to fill in a questionnaire to rate their experience on a 5-point Likert scale (with two positive, one neutral and two negative options) in terms of:

Q1 (1-very bad, 5-excellent) quality of the audio signal received via JackTrip[3]
Q2 (1-unacceptable, 5-unnoticeable) impact of MM2ME audio delay;
Q3 (1-totally useless, 5-very useful) usefulness of audio feedback from the remote counterpart;
Q4 (1-very bad, 5-excellent) quality of musical interaction.

Artists involved in multiple sessions provided a different set of ratings for each session. Of the artists involved, two declined to answer the questionnaire (P and C2) while the dilruba player (D) was excluded from the survey, being an author of this paper and one of the main developers of the JackTrip software. Note that none of the components of the piano trio (P, V1, C1) nor any of the singers (S, T1, T2, T3, B1, B2) had ever had any prior experience with NMP, whereas the players from Stanford (D, C2) and New Haven (V2) had significant background and had already performed remotely on multiple occasions. It is important to remark that the collected ratings do not ensure statistical significance due to the limited number of musicians involved and to the heterogeneous performance conditions. Nevertheless, they provide interesting insights on the outcome of the performance experience (PE) from the point of view of the artists. Averaged results are reported in Figs. 7, 8, and 9, to facilitate the reading, whereas Fig. 10 shows the overall results of the survey.

Ratings provided for the Munich-Stanford and Munich-Turin sessions are reported in Figs. 7 and 8, respectively, while the ratings for the New Haven-Turin session are reported in 9. Note that the string players of the piano trio (V1, C1), who did not have any direct audio feedback from the hub during the performance as they were following the pianist, were excluded from the survey. Ratings provided by the two soloists (S and T1) are reported in Fig. 7. As per Table 2, the results plotted in Figs. 7-9 were assessed for different types of musical interactions. For example, the Turin-Munich and New Haven-Turin sessions involve a set of musical pieces where the singers/violinist perform as soloists, whereas the Munich-Stanford session consists in a choral piece with 6 singers performing in the same environment, with remote improvisation. Note that the instrumentalists in the New Haven-Turin session, who were isolated, fully monitoring network audio, and performing with network accompaniment from a similarly isolated, remote musician, experienced arguably the most common Jacktrip use-case conditions.

Results show that the perceived quality of the audio stream delivered by JackTrip was rated as good (4) or excellent (5) by the majority of the players.

In general, the impact of MM2ME delay is evaluated between tolerable (3) and barely tolerable (2).

Though delayed, the audio feedback from the hub is rated as quite useful (4) or useful (5) in the majority of the cases. For example, in the Munich-Stanford performance, four singers out of six rated the audio feedback as quite useful (4), one as neither useful nor useless (3) and one as completely useless (1).

The overall quality of interaction in the performance was rated as good (4) or average (3) by four musicians and quite low (2) by five musicians. Finally, some of the general comments we received include: ”We didn’t have much musical interaction this time, so we cannot comment a lot on the quality. The sound was quite good”; ”Interesting situation for musicians. Probably it needs a lot of practice to get to the point of making music together”; ”Still I look forward to get more experience with the program”; ”It is a very interesting and useful experience”; ”After training and practicing with the Jacktrip software, this was a very enjoyable performance experience”.

5 RECENT JACKTRIP FEATURES AND OPEN TECHNICAL CHALLENGES

Though a number of NMP solutions, either experimental or commercial, are already being used to support remote musical interactions, several technical challenges remain still open to bring NMP technology to a large scale use.

[3] In this context, the ratings refer mostly to the impact of potential artifacts and glitches caused by packet losses, rather than an evaluation of quality of the PCM audio stream exchanged between the players through Jacktrip.

In the months that have elapsed since the concert, JackTrip has been further modified to improve jitter handling as well as to support a second set of audio outputs with an additional buffer whose longer queue length better avoids jitter-induced drops. With this broadcast feature, uplinks and recordings can have significantly improved quality and allow the primary real-time set of outputs to be pushed at lower latencies, something which would have been especially useful for this concert with strongly-synchronized chamber music works.

Currently envisioned technological advancements include:

- Development of dedicated hardware-based solutions (relying for example on field programmable gate array processors) to further decrease audio acquisition and processing times.
- Integration in NMP hardware of 5G wireless transmission technologies, which promise to ensure access delay to the backbone telecommunication infrastructure below 10 ms, while getting rid of wired Ethernet connections currently required for local area network connectivity (WiFi wireless connections cannot be leveraged as they introduce too high jitter) and thus providing additional flexibility in the equipment setup.
- Fostering the involvement of Internet service providers and network operators to develop commercial solutions for NMP services, where traffic generated by audio streams could be prioritized to reduce queuing delays occurring at intermediate nodes of the telecommunication infrastructures.
- Leveraging machine learning approaches to predict how the audio signal will evolve in a future time period: such predictions could be used to replace portions of the audio streams that may get lost or arrive too late to be reproduced, due to transmission errors or delay fluctuations [38].

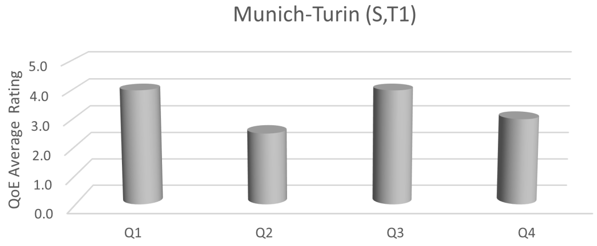

Fig. 7: PE ratings for the Turin-Munich session (S, T1).

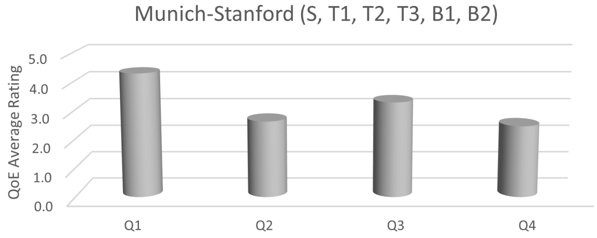

Fig. 8: PE ratings for the Munich-Stanford session (S, T1, T2, T3, B1, B2).

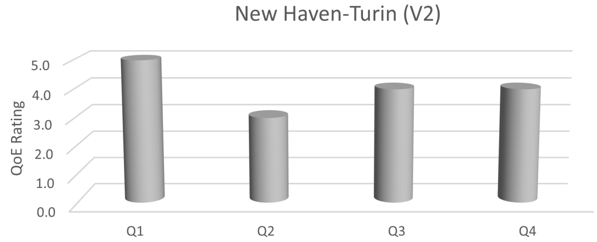

Fig. 9: PE ratings for the New Haven-Turin session.

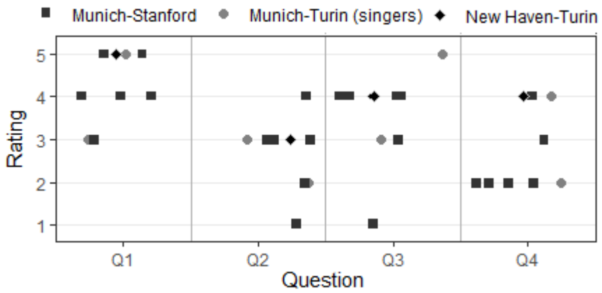

Fig. 10: Overall survey data.

6 CONCLUSION

This paper demonstrates the usage of the NMP software JackTrip to support a distributed classical concert involving musicians from four different locations (Politecnico di Torino in Turin, Italy; Ludwig-Maximilian Universität in Munich, Germany; Stanford University, California; and New Haven, Connecticut) and using readily available HW/SW solutions and internet connections. To the best of the authors' knowledge, this was the first documented attempt to perform a classical music repertoire in a distributed concert setting where physical separation between locations is in the order of 10,000 km, while ensuring high-fidelity audio quality: the need to cope with audio transmission delays above 100 ms (one-way) required preliminary training for the musicians, to devise specific compensation mechanisms.

The aim of the project was twofold. On one side, the involvement of first-class professional artists provided feedback while testing JackTrip over international and even intercontinental connections which allowed researchers and technical staff to better identify features and requirements that should be satisfied to improve performance conditions. On the other side, the concert aimed at raising the attention of the audience to the fact that research and current technology is moving beyond the first steps in making remote, real-time musical interactions possible, but also reaching out to the research community and industrial stakeholders to make them aware of the technological challenges that still need to be addressed in order to achieve conditions closer to those of traditional in-person music playing.

7 ACKNOWLEDGMENTS

The authors thank the instrumentalists Alexander Goldberg, Francesca Gosio, Stephen Harrison, Piergiorgio Rosso, Antonio Valentino and the singers Christian Meister, Claudia Reinhard, Marcus Schmidl, Jakob Steiner, Manuel Warwitz, Markus Zapp for their participation to the project, as well as the computer engineers Christian Riepl, Kostantinos Rigas and the audio engineers Carlo Barbagallo, Giovanni Corgiat, Zach Miley, Konrad Zinner for their precious assistance in the technical setup of the musical performance and prof. Hartmut Schick for the artistic co-direction of the event.

8 REFERENCES

[1] "European music schools in times of Coronavirus," http://www.musicschoolunion.eu/wp-content/uploads/2020/05/EMU-Survey-Coronavirus.pdf.

[2] P. Oliveros, S. Weaver, M. Dresser, J. Pitcher, J. Braasch, C. Chafe, "Telematic music: six perspectives," *Leonardo Music Journal*, vol. 19 (2009).

[3] I. Howell, K. Gautereaux, J. Glasner, N. Perna, C. Ballantyne, T. Nestorova, "Preliminary Report: Comparing the Audio Quality of Classical Music Lessons Over Zoom, Microsoft Teams, VoiceLessonsApp, and Apple FaceTime," *Ian Howell, DMA* (2020).

[4] A. Badr, A. Khisti, W.-T. Tan, J. Apostolopoulos, "Perfecting protection for interactive multimedia: A survey of forward error correction for low-delay interactive applications," *IEEE Signal Processing Magazine*, vol. 34, no. 2, pp. 95–113 (2017).

[5] J. Hirschfeld, J. Klier, U. Kraemer, G. Schuller, S. Wabnik, "Ultra Low Delay audio coding with constant Bit Rate," *Proceedings of the 117th AES Convention* (2004).

[6] U. Kraemer, H. Jens, G. Schuller, S. Wabnik, A. Carôt, C. Werner, "Network Music Performance with Ultra-Low-Delay audio coding under unreliable network conditions," *Proceedings of the 123rd AES Convention* (2007).

[7] J. Valin, G. Maxwell, T. Terriberry, V. Kos, "High-quality, low-delay music coding in the Opus codec," *Proceedings of the 135th AES Convention* (2013).

[8] "Jamulus," https://jamulus.io/it/.

[9] "SonoBus," https://sonobus.net/.

[10] A. Carôt, C. Hoene, H. Busse, C. Kuhr, "Results of the fast-music project—five contributions to the domain of distributed music," *IEEE Access*, vol. 8, pp. 47925–47951 (2020).

[11] J.-P. Cáceres, C. Chafe, "JackTrip: Under the hood of an engine for network audio," *Journal of New Music Research*, vol. 39, no. 3, pp. 183–187 (2010).

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

[13] "Aloha by Elk," https://elk.audio/aloha/.

[14] "Digital Stage," https://digital-stage.org/?lang=en.

[15] C. Drioli, C. Allocchio, N. Buso, "Networked performances and natural interaction via LOLA: Low latency high quality A/V streaming system," presented at the *International Conference on Information Technologies for Performing Arts, Media Access, and Entertainment*, pp. 240–250 (2013).

[16] "JamKazam," https://jamkazam.com/.

[17] L. B. Nielsen, "Subjective assessment of audio codecs and bitrates for broadcast purposes," *Proceedings of the 100th AES Convention* (1996).

[18] "Jitsi Meet," https://meet.jit.si/.

[19] "Open Broadcaster Software Studio," https://obsproject.com/.

[20] R. Hupke, L. Beyer, M. Nophut, S. Preihs, J. Peissig, "Effect of a global metronome on ensemble accuracy in networked music performance," presented at the *Audio Engineering Society Convention 147* (2019).

[21] R. Battello, L. Comanducci, F. Antonacci, A. Sarti, S. Delle Monache, G. Cospito, E. Pietrocola, F. Berbenni, "An Adaptive Metronome Technique for Mitigating the Impact of Latency in Networked Music Performances," presented at the *2020 27th Conference of Open Innovations Association (FRUCT)*, pp. 10–17 (2020).

[22] C. Rottondi, M. Buccoli, M. Zanoni, D. Garao, G. Verticale, A. Sarti, "Feature-Based Analysis of the Effects of Packet Delay on Networked Musical Interactions," *Journal of the Audio Engineering Society*, vol. 63, no. 11, pp. 864–875 (2015).

[23] A. Carôt, C. Werner, T. Fischinger, "Towards a comprehensive cognitive analysis of delay-influenced rhythmical interaction," presented at the *ICMC* (2009).

[24] H. von Coler, N. Tonnätt, V. Kather, C. Chafe, "SPRAWL: a Network System for Enhanced Interaction in Musical Ensembles," presented at the *Proceedings of the 18th Linux Audio Conference (LAC-20)* (2020).

[25] "QJackTrip," https://www.psi-borg.org/other-dev.html.

[26] F. Cretti, L. Morino, M. Liuni, S. Gervasoni, A. Agostini, A. Servetti, "Web Wall Whispers: an interactive web-based sound work," presented at the *4th Intl. Web Audio Conference*, WAC '18 (2018 September).

[27] M. Puckette, *et al.*, "Pure Data: another integrated computer music environment," *Proceedings of the second intercollege computer music concerts*, pp. 37–41 (1996).

[28] "JackTrip Foundation," https://www.jacktrip.org.

[29] "QjackCtl," https://qjackctl.sourceforge.io/qjackctl-index.html#Intro.

[30] J.-P. Cáceres, "JMess - A utility to save your audio connections (mess)," https://github.com/jacktrip/jmess-jack.

[31] "JACK Audio Connection Kit," https://jackaudio.org/.

[32] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre,

G. Davidson, Y. Oikawa, "ISO/IEC MPEG-2 Advanced Audio Coding," *Journal of the Audio Eng. Soc.*, vol. 45, no. 10, pp. 789–814 (1997).

[33] NDI, "Network Device Interface (NDI)," `https://www.ndi.tv/`.

[34] "Pop-up videos," `https://github.com/Jip-Hop/pop-up-videos`.

[35] F. Adriaensen, "jack_iodelay: JACK toolkit client to measure roundtrip latency," `https://github.com/jackaudio/tools/blob/master/iodelay.c`.

[36] P. Ferguson, C. Chafe, S. Gapp, "Trans-Europe Express Audio: testing 1000 mile low-latency uncompressed audio between Edinburgh and Berlin using GPS-derived word clock, first with jacktrip then with Dante." presented at the *Audio Engineering Society Convention 148* (2020).

[37] K. Tsioutas, G. Xylomenos, I. Doumanis, C. Angelou, "Quality of Musicians' Experience in Network Music Performance: A Subjective Evaluation," presented at the *Audio Engineering Society Convention 148* (2020).

[38] P. Verma, A. I. Mezza, C. Chafe, C. Rottondi, "A Deep Learning Approach for Low-Latency Packet Loss Concealment of Audio Signals in Networked Music Performance Applications," presented at the *2020 27th Conference of Open Innovations Association (FRUCT)*, pp. 268–275 (2020), doi:10.23919/FRUCT49677.2020.9210988.

THE AUTHORS

Marina Bosi

Antonio Servetti

Chris Chafe

Cristina Rottondi

Marina Bosi, a pioneer in the development of digital audio coding, received her degrees in Physics and in Music from the University of Florence (Italy) and from the National Conservatory of Music (Florence, Italy), respectively, and completed her thesis as a Chargèe de Recherche at IRCAM in Paris, France. After working with the composer Luciano Berio at Tempo Reale, Florence, Marina came to the USA and became a Consulting Professor (since 1998) at Stanford University's Computer Center for Research in Music and Acoustics (CCRMA) and in the Electrical Engineering department (2004-2009). An experienced industry leader, Marina co-founded (with L. Chiariglione, MPEG Chair) the Digital Media Project and was the CTO of MPEG LA, Vice President - Technology at DTS, Project Engineer at Dolby Laboratories, and DSP Engineer at Digidesign (Avid). A Fellow and Past President of the AES and a Senior Member of IEEE, Marina's awards include the AES Silver Medal in recognition of "outstanding achievements in the development and standardization of audio and secure digital rights management", twice the AES Board of Governors Award (1995 and 2000), and the ISO/IEC Editor award for her work on ISO/IEC 13818-7, MPEG-2 Advanced Audio Coding (AAC). A graduate of Stanford Business School's "Stanford Executive Program", Marina holds several patents and scientific publications, and she is author of the acclaimed textbook "Introduction to Digital Audio Coding and Standards" (Kluwer/Springer December 2002). Marina is currently Treasurer and a Board member of the AES, a founding Director of the Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) and Chair of MPAI Context-based Audio Enhancement Development Committee.

●

Antonio Servetti is Assistant Professor with the Department of Control and Computer Engineering of the Politecnico di Torino (Italy) since 2007. He received the MS degree and the PhD degree in Computer Engineering from the Politecnico di Torino in 1999 and 2004 respectively. In 2003 Dr. Servetti was a Visiting Scholar supervised by prof. J.D. Gibson at the Signal Compression Laboratory of the University of California, Santa Barbara, where he worked on selective encryption for speech transmission over packet networks. His research focuses on speech/audio processing, multimedia communications over wired and wireless packet networks, and real-time multimedia network protocols. With the advent of video and audio support in HTML5, his interests include also multimedia Web applications, WebRTC, Web Audio, and HTTP adaptive streaming.

●

Chris Chafe is a composer, improvisor, and cellist, developing much of his music alongside computer-based research. He is Director of Stanford University's Center for Computer Research in Music and Acoustics (CCRMA). In 2019, he was International Visiting Research Scholar at the Peter Wall Institute for Advanced Studies The University of British Columbia, Visiting Professor at the Politecnico di Torino, and Edgard-Varèse Guest Professor at the Technical University of Berlin. At IRCAM (Paris) and The Banff Centre (Alberta), he has pursued methods for digital synthesis, music performance and real-time internet collaboration. CCRMA's JackTrip project involves live concertizing with musicians the world over. Online collaboration software and research into latency factors continue to evolve. An active performer either on the net or physically present, his music reaches audiences in sometimes novel venues. An early network project was a simultaneous five-country concert was hosted at the United Nations in 2009. Chafe's works include gallery and museum music installations which are now into their second decade with "musifications" resulting from collaborations with artists, scientists and MD's. Recent work includes the Earth Symphony, the Brain Stethoscope project (Gnosisong), PolarTide for the 2013 Venice Biennale, Tomato Quintet for the transLife:media Festival at the National Art Museum of China and Sun Shot played by the horns of large ships in the port of St. Johns, Newfoundland.

●

Cristina Rottondi is Assistant Professor with the Department of Electronics and Telecommunications of Politecnico di Torino (Italy). Her research interests include optical networks planning and networked music performance. She received both Bachelor and Master Degrees "cum laude" in Telecommunications Engineering and a PhD in Information Engineering from Politecnico di Milano (Italy) in 2008, 2010 and 2014 respectively. From 2015 to 2018 she had a research appointment at the Dalle Molle Institute for Artificial Intelligence (IDSIA) in Lugano, Switzerland. She is co-author of more than 80 scientific publications in international journals and conferences. She served as Associate Editor for IEEE Access from 2016 to 2020 and is currently Associate Editor of the IEEE/OSA Journal of Optical Communications and Networking. She is co-recipient of the 2020 Charles Kao Award, of three best paper awards (FRUCT-IWIS 2020, DRCN 2017, GreenCom 2014) and of one excellent paper award (ICUFN2017).