M. Bosi, A. Servetti, C. Chafe, and C. Rottondi, "Experiencing Remote Classical Music Performance Over Long Distance: A JackTrip Concert Between Two Continents During the Pandemic"
J. Audio Eng. Soc., vol. 69, no. 12, pp. 934–945, (2021 December).
DOI: https://doi.org/10.17743/jaes.2021.0056

Experiencing Remote Classical Music Performance Over Long Distance: A JackTrip Concert Between Two Continents During the Pandemic

MARINA BOSI,[1] *AES Fellow*,
(mab@ccrma.stanford.edu)

ANTONIO SERVETTI,[2] *AES Associate Member*, **CHRIS CHAFE,**[1] **AND CRISTINA ROTTONDI**[3]
(antonio.servetti@polito.it) (cc@ccrma.stanford.edu) (cristina.rottondi@polito.it)

[1] *Center for Computer Research in Music and Acoustics, Stanford University, Stanford, California*
[2] *Department of Control and Computer Engineering, Politecnico di Torino, Turin, Italy*
[3] *Department of Electronics and Telecommunications, Politecnico di Torino, Turin, Italy*

The recent lockdown restrictions imposed by the severe acute respiratory syndrome coronavirus 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 software JackTrip to support a distributed classical concert involving singers and musicians from four different locations in two continents, using readily available hardware/software solutions and internet connections while guaranteeing high-fidelity audio quality. This paper provides a description of the technical setup with a numerical analysis of the achieved mouth-to-ear latency and assessment of the music-making experience as perceived by the performers.

0 INTRODUCTION

The social-distancing measures adopted to mitigate the propagation of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic have reinforced the need to develop multimedia tools to support remote music performances and the 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 fiber-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 and 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 or 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. 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., [5–7]. Given its ultra-low delay and open-source availability, many of the current NMP systems (e.g., [8–10]) utilize [7].

In the NMP context, during the last 20 years, the Center for Computer Research in Music and Acoustics (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 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 because of the intrinsic characteristics and interplay requirements of such a musical genre, which makes it extremely challenging to be executed in the presence of end-to-end delays above a few tens of milliseconds [12]. Indeed, this classical music concert highlighted both the potential and 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 performances involved intricate rhythmic coordination that would be difficult given the longer-latency scenarios expected. Rubato and expressive timings were of particular interest because they might be ambiguous for musicians, who might confuse whether micro-timing inflections were intentional or artifacts of network delay. The concert involved 12 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.

Because of the lockdown restrictions in force at the time of the festival, to counteract the spreading of SARS-CoV-2, 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 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 to what 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, can guide NMP software developers in devising future enhancements and new features, and can provide suggestions to practitioners dealing with similar NMP setups.

The remainder of the paper is organized as follows. Sec. 1 briefly reviews the related literature and some existing solutions for NMP, whereas Sec. 2 provides insights on the impact of latency on networked musical interactions and JackTrip software. Sec. 3 includes a detailed description of the technical setup for audio-video streaming adopted during the performance, and Sec. 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 Secs. 5 and 6, which conclude the paper.

1 RELATED WORK

1.1 Hardware/Software 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 [8–10, 13–16] 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 a 48-kHz sampling rate and 16-bit sample

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

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; HW, hardware; ISP, internet service provider; NMP, Networked Music Performance; SW, software.

precision[2]) using off-the-shelf, minimal hardware requirements (see also Sec. 3.2). Of the systems listed in Table 1, LOLA [15] allows for uncompressed-only media signal 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/performing 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 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 sixty-two 1-h sessions exploring different performance concepts (including free improvisation, graphical scores, text-based scores, and sound painting). Over 20 guest musicians and visual artists have participated from their homes in various countries, including the United States, Canada, Ireland, Germany, Lithuania, Australia, and United Kingdom. Free and open-source technologies were utilized for these sessions, most notably 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 Secs. 3.3 and 3.4).

Fig. 1. Quarantine Session at Stanford's Center for Computer Research in Music and Acoustics (CCRMA).

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 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

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

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 1/8 or 1/16 depending on the metronome marking of the piece, and they played with the beat and 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 the maximum acceptable physical separation between musicians that permits 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 a 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 rhythmic complexity of the musical piece being performed [12, 22, 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.

Because of 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. 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. Delay introduced by the audio acquisition/reproduction, processing, and packetization/depacketization at sender/receiver sides;
3. Pure propagation delay over the physical transmission medium;
4. Data processing delay introduced by the intermediate network nodes traversed by the audio data packets along their path from source to destination; and
5. Playback buffering that 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 1,000 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.

2.2 The JackTrip Software

JackTrip is a multi-machine technology that supports bidirectional flows of uncompressed audio over the internet while minimizing latency. Developed in the early 2000s, 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 information technology 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 is required for ultra-low–latency rhythmic synchronization.

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 (see JackTrip WebRTC [26]), and Pure Data [27] versions. The newly established JackTrip Foundation [28] is a nonprofit 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 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 because of the physical distance between the two sites (about 10,000 km,

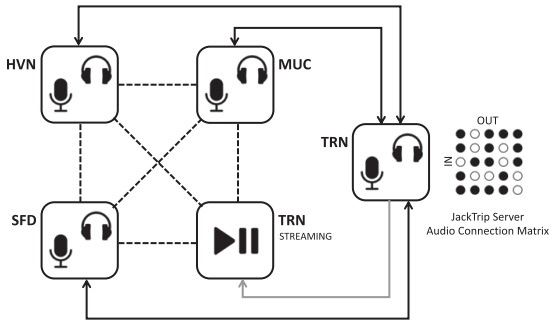


Fig. 2. Technical setup of the performance between Turin (TRN), Stanford (SFD), Munich (MUC), and New Haven (HVN). Solid lines correspond to client-server stereo audio connections, and dashed lines correspond to peer-to-peer video connections. On the JackTrip server in Turin the audio is mixed and routed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

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 a 44.1-kHz sampling rate, measured network round trip delay of around 166 ms, and overall systems delay of 190 ms. Based on the outcomes of such trials and 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 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 software (Ableton Live) that takes care of the overall mixing and 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 prevent peaking by means of the Waves Linear Phase Multiband Compressor plugin (the threshold was set to –12 dB, the ratio was set 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 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 a 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 makes the video even detrimental to the musicians.

3.4 Broadcast Streaming Setup

The broadcast streaming setup is based on OBS [19], which captures the A/V inputs, performs the A/V mix, encodes audio and 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 MPEG Advanced Audio Coding-Low Complexity (AAC LC) [32] stream at 320 kbps and the video as a 1,920 x 1,080 30 fps H.264 stream at 9 Mbps.

The video feed is exported from the videoconferencing software to OBS as a Network Device Interface [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 Jack-

Fig. 3. Distributed concert; Turin-Munich session.

Fig. 4. Distributed concert; New Haven–Turin session.

Trip. 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 string players of the piano trio preferred to not 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, and 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.

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 geographical locations from which the artists were performing. Overall the concert program accounted for around 45 min of live music.

4.1 Network and Mouth-to-Ear Latency

During the concert rehearsals, some quantitative tests were performed to set up the software configuration for the specific scenario. After some preliminary trials, the JackTrip packet size was set to 512 samples and buffer queue size to four packets. A relatively large size was needed to increase the robustness of the connections with Stanford (distance from Torino to Stanford is about 10,000 km) and New Haven (distance Torino to New Haven is about 6,000 km). Tests with Munich (distance from Torino to Munich is 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 ping utility across the Atlantic Ocean reported an average network round-trip latency of 184 ms with Stanford and about 122 ms with New Haven, while the same measure with Munich reported about 27 ms.

At the same time, during the ping test, the round-trip audio latency [i.e., the My Mouth to My Ear (MM2ME) latency] was measured 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, the *jack_iodelay* output was connected to the JackTrip client input and 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 a 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, and is 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, and thus a variation of the MM2ME audio latency was noticed: 243.81, 255.42, 267.03, and 278.64 ms were reported in the 31%, 2%, 21%, and 46% of the measurements respectively.

Table 2. Concert repertoire.

Composer	Title	Voices (IDs)	Instruments (IDs)	Locations
Ludwig van Beethoven	Op. WoO 158a no. 16 “Schöne Minka, ich muss scheiden”	Soprano (S), Tenor (T1)	Piano (P), Violin (V1), Cello (C1)	Turin (instrumental trio), Munich (singers)
	Op. WoO 158a no. 17 “Lilla Carl, sov sött i frid”	Soprano (S)		
	Op. WoO 158a no. 23 “Da brava, Catina”	Tenor (T1)		
	Op. 108 no. 2 “Sunset”	Tenor (T1)		
	Op. 108 no. 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: III. 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)

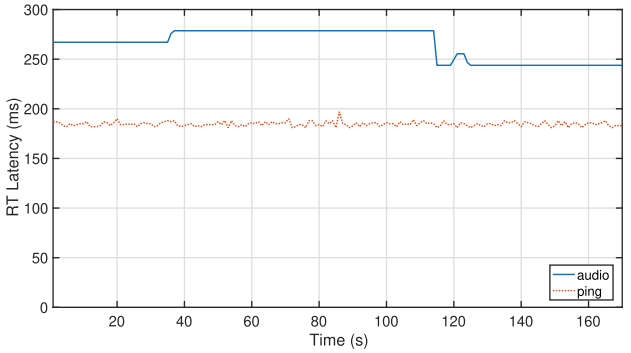


Fig. 5. Audio and network round-trip latency measured with *jack_iodelay* and ping during a JackTrip connection between Turin and Stanford.

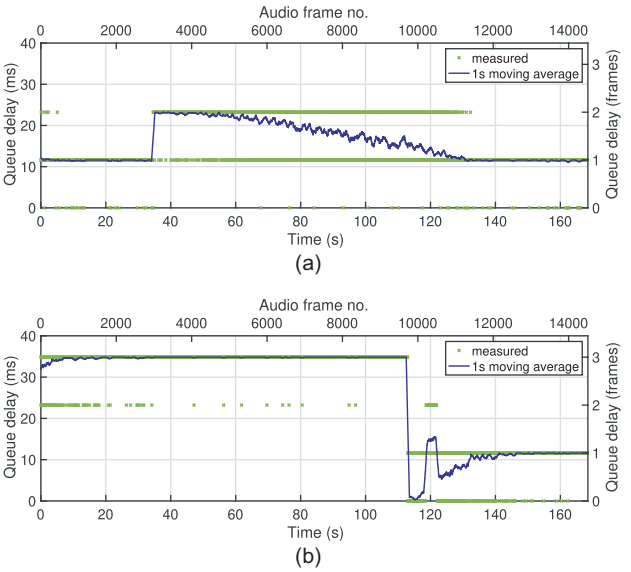


Fig. 6. Server (a) and client (b) JackTrip reception queue size/delay after an audio frame is read and extracted (i.e., every 11.6 ms). Measured value and moving average over a 1-s sliding window during a test connection between Stanford and Turin are shown.

To further investigate the different contributions that sum up to the final audio delay, the length of the JackTrip queues in the audio round-trip path was traced: the transmission queue of the client, reception and transmission queue of the server, and reception queue of the client. The sum of the number of frames waiting in each queue, as measured *after* a frame is extracted, can provide information about the overall buffering delay, i.e., the time that a frame spends in the queue. The queue length value was sampled every time an audio frame was read, thus every 11.6 ms. If b is the number of frames in the buffer, it can be assumed 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 as Fig. 5. The dots represent the measured values at each sampling time, and the line is the moving average calculated over a sliding window of 1 s. 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 zero to the maximum JackTrip queue size, because of variable packet delays and the clock skew between the client and server instances. To better 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, the authors had to take care of the audio/video synchronization, since audio and video were transmitted using different software applica-

tions and thus received with different delays. The video, sent by means of a commercial videoconferencing tool, suffered from a higher delay than the audio, so OBS was used to add an additional delay to the audio stream. The A/V delay was measured with a test video that emitted a beep every second and moved a bar on the screen from left to right over a series of equally spaced ticks that represented the time difference from the beep in steps of 0.1 s. The videoconferencing tool frame rate was 30 fps, and thus the A/V delay could be measured with a precision of about 1/30 s. On the connection between Munich and Turin, a 1-min test was performed to measure the A/V delay, the results of which showed a range of 3–5 frames. This result is compatible with both the varying audio delay that was measured with JackTrip and varying video delay that the videoconferencing software was assumed to introduce.

4.2 Rating of the Performance Experience

Rehearsals were organized into four sessions (Munich-Turin, Munich-Stanford, New Haven–Turin, and 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 Quality of the audio signal received via JackTrip[3] (1: very bad; 5 excellent);
- Q2 Impact of MM2ME audio delay (1: unacceptable; 5: unnoticeable);
- Q3 Usefulness of audio feedback from the remote counterpart (1: totally useless; 5: very useful); and
- Q4 Quality of musical interaction (1: very bad; 5: excellent).

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 by 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 because of the limited number of musicians involved and heterogeneous performance conditions. Nevertheless they provide interesting insights on the outcome of the performance experience from the point of view of the artists. Averaged results are reported in Figs. 7–9 to facilitate the reading, whereas Fig. 10 shows the overall results of the survey.

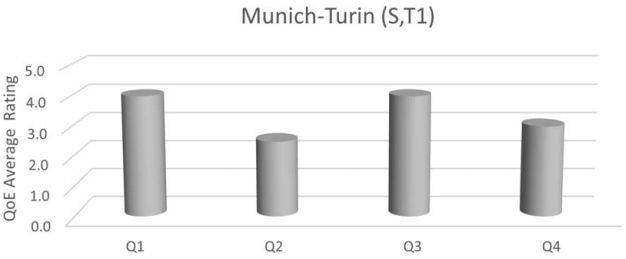


Fig. 7. Performance experience ratings for the Turin-Munich session (S, T1). S, soprano; T1, tenor 1.

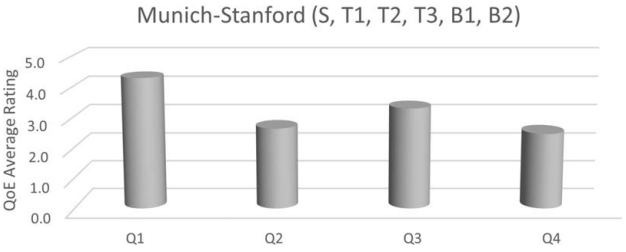


Fig. 8. Performance experience ratings for the Munich-Stanford session (S, T1, T2, T3, B1, B2). B1, baritone; B2, bass; S, soprano; T1, tenor 1; T2, tenor 2; T3, tenor 3.

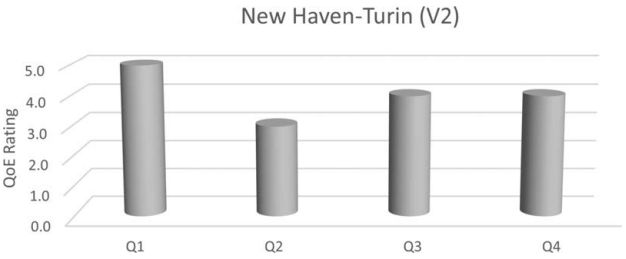


Fig. 9. Performance experience ratings for the New Haven–Turin session. V2, violin 2.

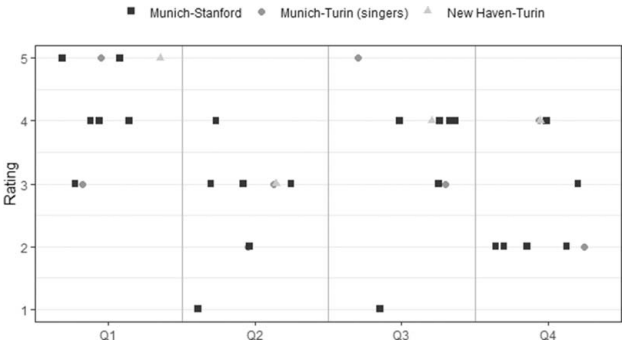


Fig. 10. Overall survey data.

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 Fig. 9. Note that the string players of the piano trio

[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 Pulse-Code Modulation audio stream exchanged between the players through JackTrip.

(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 involved a set of musical pieces in which the singers/violinist perform as soloists, whereas the Munich-Stanford session consisted of a choral piece with six singers performing in the same environment with remote improvisation. Note that the instrumentalists in the New Haven–Turin session, who were isolated and were 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 was evaluated between tolerable (3) and barely tolerable (2). Although delayed, the audio feedback from the hub was 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 rated it as neither useful nor useless (3), and one rated it 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 the authors 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”; and “After training and practicing with the JackTrip software, this was a very enjoyable performance experience.”

5 RECENT JACKTRIP FEATURES AND OPEN TECHNICAL CHALLENGES

Although a number of NMP solutions, either experimental or commercial, are already being used to support remote musical interactions, several technical challenges still remain open to bring NMP technology to a large-scale use. In the months that have elapsed since the concert, JackTrip has been further modified to improve jitter handling and 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 since 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, in which traffic generated by audio streams could be prioritized to reduce queuing delays occurring at intermediate nodes of the telecommunication infrastructures; and
- Leveraging machine-learning approaches to predict how the audio signal will evolve in a future time period, and such predictions could be used to replace portions of the audio streams that may get lost or arrive too late to be reproduced, because of transmission errors or delay fluctuations [38].

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 hardware/software 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 ACKNOWLEDGMENT

The authors thank the instrumentalists Alexander Goldberg, Francesca Gosio, Stephen Harrison, Piergiorgio Rosso, and Antonio Valentino and the singers Christian Meister, Claudia Reinhard, Marcus Schmidl, Jakob Steiner, Manuel Warwitz, and Markus Zapp for their participation to the project. The authors also thank the computer engineers Christian Riepl and Kostantinos Rigas and the audio engineers Carlo Barbagallo, Giovanni Corgiat, Zach Miley, and 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 School Union, "European Music Schools in Times of Coronavirus," (2020 May). http://www.musicschoolunion.eu/wp-content/uploads/2020/05/EMU-Survey-Coronavirus.pdf.

[2] P. Oliveros, S. Weaver, M. Dresser, et al., "Telematic Music: Six Perspectives," *Leonardo Music J.*, vol. 19, pp. 95–96 (2009 Dec.).

[3] I. Howell, K. Gautereaux, J. Glasner, et al., "Preliminary Report: Comparing the Audio Quality of Classical Music Lessons Over Zoom, Microsoft Teams, VoiceLessonsApp, and Apple FaceTime," *Special Report of the New England Conservatory of Music Voice and Sound Analysis Laboratory* (2020 Mar.).

[4] A. Badr, A. Khisti, W.-T. Tan, and J. Apostolopoulos, "Perfecting Protection for Interactive Multimedia: A Survey of Forward Error Correction for Low-Delay Interactive Applications," *IEEE Signal Process. Mag.*, vol. 34, no. 2, pp. 95–113 (2017 Mar.). https://doi.org/10.1109/MSP.2016.2639062.

[5] J. Hirschfeld, J. Klier, U. Kraemer, G. Schuller, and S. Wabnik, "Ultra Low Delay Audio Coding With Constant Bit Rate," presented at the *117th Convention of the Audio Engineering Society* (2004 Oct.), convention paper 6197.

[6] U. Kraemer, H. Jens, G. Schuller, et al., "Network Music Performance With Ultra-Low-Delay Audio Coding Under Unreliable Network Conditions," presented at the *123rd Convention of the Audio Engineering Society* (2007 Oct.), convention paper 7214.

[7] J. Valin, G. Maxwell, T. Terriberry, and V. Kos, "High-Quality, Low-Delay Music Coding in the Opus Codec," presented at the *135th Convention of the Audio Engineering Society* (2013 Oct.), convention paper .

[8] Volker Fischer, "Jamulus," https://jamulus.io/it/. (accessed April 1, 2021).

[9] Sonosaurus LLC, "SonoBus," https://sonobus.net/. (accessed April 1, 2021).

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

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

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

[13] Elk, "Aloha by Elk," https://elk.audio/aloha/. (accessed April 1, 2021).

[14] Digitale Bühne gGmbH, "Digital Stage," https://digital-stage.org/?lang=en. (accessed April 1, 2021).

[15] C. Drioli, C. Allocchio, and N. Buso, "Networked Performances and Natural Interaction via LOLA: Low Latency High Quality A/V Streaming System," in P. Nesi and R. Santucci (Eds.), *Information Technologies for Performing Arts, Media Access, and Entertainment*, pp. 240–250 (Springer-Verlag, Berlin, Germany, 2013). https://doi.org/10.1007/978-3-642-40050-6_21.

[16] JamKazam, "JamKazam," https://jamkazam.com/. (accessed April 1, 2021).

[17] L. B. Nielsen, "Subjective Assessment of Audio Codecs and Bitrates for Broadcast Purposes," presented at the *100th Convention of the Audio Engineering Society* (1996 May), convention paper 4175.

[18] 8x8 Inc.,"Jitsi Meet," https://meet.jit.si/. (accessed April 1, 2021).

[19] "Open Broadcaster Software Studio," https://obsproject.com/. (accessed April 1, 2021).

[20] R. Hupke, L. Beyer, M. Nophut, S. Preihs, and J. Peissig, "Effect of a Global Metronome on Ensemble Accuracy in Networked Music Performance," presented at the *147th Convention of the Audio Engineering Society* (2019 Oct.), convention paper 10218.

[21] R. Battello, L. Comanducci, F. Antonacci, et al., "An Adaptive Metronome Technique for Mitigating the Impact of Latency in Networked Music Performances," in *Proceedings of the 27th Conference of Open Innovations Association (FRUCT)*, pp. 10–17 (Trento, Italy) (2020 Oct.).

[22] C. Rottondi, M. Buccoli, M. Zanoni, et al., "Feature-Based Analysis of the Effects of Packet Delay on Networked Musical Interactions," *J. Audio Eng. Soc.*, vol. 63, no. 11, pp. 864–875 (2015 Nov.).

[23] A. Carôt, C. Werner, and T. Fischinger, "Towards a Comprehensive Cognitive Analysis of Delay-Influenced Rhythmical Interaction," in *Proceedings of the International Computer Music Conference (ICMC)* (Montreal, Canada) (2009).

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

[25] Aaron Wyatt, "QJackTrip," https://www.psi-borg.org/other-dev.html. (accessed April 1, 2021).

[26] M. Sacchetto, A. Servetti, C. Chafe, "JackTrip-WebRTC: Networked Music Experiments With PCM Stereo Audio in a Web Browser," presented at the *6th International Web Audio Conference (WAC)* (2021 Jul.).

[27] M. S. Puckette, "Pure Data: Another Integrated Computer Music Environment," in *Proceedings of the*

2nd Intercollege Computer Music Concerts, pp. 37–41 (Tachikawa, Japan) (1996 Aug.).

[28] "JackTrip Foundation," https://www.jacktrip.org. (accessed April 1, 2021).

[29] R. N. Capela, "QjackCtl," https://qjackctl.sourceforge.io/qjackctl-index.html#Intro. (accessed April 1, 2021).

[30] J.-P. Cáceres, "JMess - A utility to save your audio connections (mess)," https://github.com/jacktrip/jmess-jack. (accessed April 1, 2021).

[31] A. Knoth, F. Coelho, N. Arnaudov, S. Letz, "JACK Audio Connection Kit," https://jackaudio.org/. (accessed April 1, 2021).

[32] M. Bosi, K. Brandenburg, S. Quackenbush, et al., "ISO/IEC MPEG-2 Advanced Audio Coding," *J. Audio Eng. Soc.*, vol. 45, no. 10, pp. 789–814 (1997 Oct.).

[33] NDI, "Network Device Interface (NDI)," https://www.ndi.tv/. (accessed April 1, 2021).

[34] J. de Beer, "Pop-Up Videos," https://github.com/Jip-Hop/pop-up-videos. (accessed April 1, 2021).

[35] F. Adriaensen, "jack_iodelay: JACK Toolkit Client to Measure Roundtrip Latency," https://github.com/jackaudio/tools/blob/master/iodelay.c. (accessed April 1, 2021).

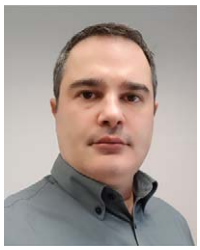
[36] P. Ferguson, C. Chafe, and S. Gapp, "Trans-Europe Express Audio: Testing 1000 Mile Low-Latency Uncompressed Audio Between Edinburgh and Berlin Using GPS-Derived Word Clock, First With JackTrip Then With Dante." presented at the *148th Convention of the Audio Engineering Society* (2020 May), convention paper 605.

[37] K. Tsioutas, G. Xylomenos, I. Doumanis, and C. Angelou, "Quality of Musicians' Experience in Network Music Performance: A Subjective Evaluation," presented at the *148th Convention of the Audio Engineering Society* (2020 May), convention paper 10357.

[38] P. Verma, A. I. Mezza, C. Chafe, and C. Rottondi, "A Deep Learning Approach for Low-Latency Packet Loss Concealment of Audio Signals in Networked Music Performance Applications," in *Proceedings of the 27th Conference of Open Innovations Association (FRUCT)*, pp. 268–275 (Trento, Italy) (2020 Oct.). https://doi.org/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, is Consulting Professor at Stanford University's Center for Computer Research in Music and Acoustics. An experienced industry leader, Marina was the Chief Technology Officer of MPEG LA, Vice President-Technology at DTS, Project Engineer at Dolby Laboratories, Digital Signal Processing Engineer at Digidesign (Avid), and a cofounder of the Digital Media Project. A Fellow and Past President of the AES, and a Senior Member of IEEE, Marina has received a number of awards including the AES Silver Medal, the AES Board of Governors Award, and the ISO/IEC Editor award for her work on ISO/IEC 13818-7, MPEG-2 Advanced Audio Coding. 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). She 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) organization; and Chair of the MPAI Context-based Audio Enhancement Development Committee.

•

Antonio Servetti has been an Assistant Professor with the Department of Control and Computer Engineering of the Politecnico di Torino (Italy) since 2007. He received M.S. and Ph.D. degrees 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 also include 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 the Institut de Recherche et Coordination Acoustique/Musique (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. 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 medical doctors. Recent work includes the Earth Symphony, the Brain Stethoscope project (Gnosisong), Polar-Tide 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. John's, 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's and Master's Degrees "cum laude" in Telecommunications Engineering and a Ph.D. in Information Engineering from Politecnico di Milano (Italy) in 2008, 2010, and 2014 respectively. From 2015–2018 she had a research appointment at the Dalle Molle Institute for Artificial Intelligence 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 2016–2020 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, three best paper awards [Conference of Open Innovations Association (FRUCT)–International Workshop on the Internet of Sounds 2020, International Conference on Design of Reliable Communication Networks 2017, and IEEE Green Computing and Communications Conference 2014], and one excellent paper award (International Conference on Ubiquitous and Future Networks 2017).