Pristine Quality Networked Music Performance System for the Web

DOI: 10.61782/fa.2023.0556

forum acusticum 2023

PRISTINE QUALITY NETWORKED MUSIC PERFORMANCE SYSTEM FOR THE WEB

Matteo Sacchetto[1] **Leonardo Severi**[1] **Cristina Rottondi**[1] **Louena Shtrepi**[2]
Marco Masoero[2] **Carlo Barbagallo**[3] **Andrea Valle**[4] **Antonio Servetti**[5*]
[1] Department of Electronics, Politecnico di Torino, Turin, Italy
[2] Department of Energy, Politecnico di Torino, Turin, Italy
[3] SMET / Conservatorio G. Verdi di Torino, Turin, Italy
[4] StudiUm/CIRMA, Università di Torino, Turin, Italy
[5] Department of Control and Computer Engineering, Politecnico di Torino, Turin, Italy

ABSTRACT

The urge for high-speed Internet connectivity, that we have witnessed since the origins of the web, has now have paved the way for pristine quality audio communications, i.e., without artifacts arising from audio compression, and, whenever possible, with low acquisition and processing latency. In this work, we present a software application (part of the HiFiReM project) that provides a unified environment, both native/embedded and web-based, for high-fidelity interactive multi-user musical performances across the Internet. The system is shown to outperform popular video-conferencing applications that adopt state-of-the-art audio compression techniques in terms of perceived audio quality. Subjective quality assessments confirm that, given adequate network conditions, users involved in remote audio communications experience performative conditions close to those of in-presence musical interactions, even if at several miles of distance.

Keywords: *networked music performance, web application, WebRTC*

1. INTRODUCTION

The COVID-19 pandemic has forced a paradigm shift in the way musicians interact and perform with each other. With the enforcement of lockdowns and social distancing protocols, traditional musical performances and collaborations have been severely constrained. Consequently, the interest in remote musical interactions (namely Networked Music Performances - NMP) has escalated, leading to the development of various software solutions catering to this domain.

However, despite the existence of several experimental and commercial NMP solutions, their adoption outside highly specialized user communities has been, up to now, quite limited. The main reason is that these solutions often require a high level of technical-IT knowledge, making them inaccessible to non-expert users. In contrast, widely used videoconferencing tools have become the go-to platforms for remote music teaching and collaboration, despite their apparent limitations in terms of performance and audio quality.

With the development of a hybrid web-based and native application, the research described in this paper aims to demonstrate that the user's preference for less sophisticated solutions can be addressed without compromising the overall quality of experience in terms of audio quality, thus catering to the high-fidelity audio demands of professionals while still being effortlessly operable through a web interface.

After a brief overview of the state of the art in Section 2, the description of the architecture of the proposed solution follows in Section 3, while the audio quality results obtained after a subjective assessment phase are reported in Section 4. Section 5 concludes the paper.

*__Corresponding author__: antonio.servetti@polito.it.

2. STATE OF THE ART

During the COVID-19 pandemic, prevalent software tools for NMP and remote music teaching included videoconferencing applications such as Zoom [1], Skype [2], and Google Meet [3]. Though specific NMP applications with dedicated functionalities and far better audio quality already existed [4–8], most users opted for well-known videoconferencing software, clearly preferring ease of use over superior audio performance [9].

Furthermore, as it will be demonstrated in the subsequent sections of the paper, the audio quality attained through Zoom and Skype is generally deemed superior to that achievable through Google Meet. Nevertheless, the latter platform has also garnered a substantial user base, even among audiophiles. One plausible explanation for this phenomenon is the clear distinction that exists between these tools. Whereas the two former ones are native applications that necessitate downloading, installation, and configuration on local computers, the latter is a web-based application that operates on web browsers, thereby obviating any need for installation and configuration procedures.

The trend toward web-based software solutions for remote musical interactions has emerged also in the aftermath of the pandemic. A good number of the new solutions that appeared on the market in the last two years adopted the form of web-based services that can be accessed directly through the web [10–13], thereby enabling even those with limited computer proficiency to utilize them without encountering any skill-based impediments.

However, the actual improvement of these new services mainly focuses on the implementation of new features, such as additional audio controls and mixing options, to meet the requirements of musicians. On the contrary, in terms of audio quality, their implemented improvements are marginal and do not go beyond the ones achievable with a fine-tuning of the options provided by Skype or other web-based applications like Jitsi (Google Meet, unfortunately, does not allow for such customizations). Moreover, the communication latency remains within the range of interactive speech communications, which typically falls between 100-300 ms [14]. This latency is far beyond the ultra-low delay necessary for networked music performance, which is below 30 ms [15].

The reason for such a limitation is that, as far as we know, all these web services were implemented using the Web Audio [16] and WebRTC [17] standards, that do not permit transmitting audio data with low delay and without perceptual audio compression. Even with high bit rates, such compression hinders the audio quality and adds algorithmic delay during the encoding process [18].

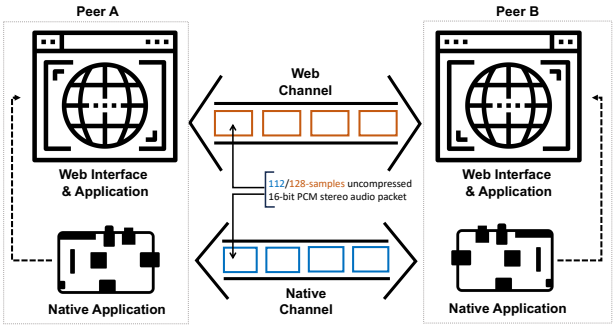


Figure 1: The proposed application web interface is shared between the native and the web implementation. Users can musically interact with other peers leveraging either the web or the native channel, using the same audio representation.

In addition, audio processing algorithms designed for voice communications can introduce sound artifacts due to automatic gain adjustment mechanisms that change sound levels. Furthermore, noise-canceling algorithms tend to dampen sustained stationary sounds, while algorithms for adapting playback speed to channel bandwidth fluctuations can alter the pitch and timbre of sounds.

This project aims to overcome the aforementioned limitations by providing a unified solution, both web and native, for the transmission of pristine (uncompressed) stereo audio. This approach attempts to find a compromise between technical sophistication and ease of use in designing NMP solutions that cater to a broad range of users.

3. SOFTWARE ARCHITECTURE

The service architecture leveraged in this paper is an enhanced version of the one presented in [19]. The service is structured around a client component which has two implementations, a web-based one and a native one. Independently of the implementation, in both versions of the client, the user interacts via the same Graphical User Interface (GUI) accessible through the web browser.

As shown in Fig. 1 the user can choose to activate the interactive music communication either through the web browser itself or through the native implementation that

runs on a dedicated device.

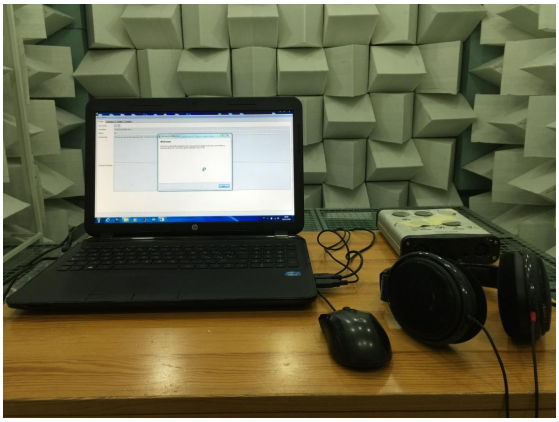
In short, the rationale behind opting for a hardware component over a personal computer is that it eliminates the need for software installation and the possibility of conflicts with other software. Additionally, it allows for the implementation of the application to suit the specific hardware, thereby optimizing computational performance and audio input/output chain. In fact, direct access and configuration of the audio card (something that is not available in the web context) permit to reduce the audio latency of the NMP system to its minimum. The native implementation is based on a Raspberry Pi 4B, it runs on top of a custom Linux Os, and it reaches a latency in the order of 10-15 ms. Additional details on the performance of the native implementation can be found in [19].

On the other side, the server component of the system is limited to managing the signaling functionality, which enables the discovery of the various clients and the initialization of the communication.

Despite the differences in the communication protocols, both the communication channels, i.e., the one initiated by the web application and the one initiated by the native application, carry the same audio format; thus, they are equivalent in terms of audio quality. The audio data is transmitted as plain uncompressed 16-bit PCM [20], which ensures unaltered audio quality at the receiver. While sending uncompressed audio is straightforward in a native application, the relevant contribution of this work is the achievement of the same result also on a peer-to-peer communication between web browsers.

In fact, the proposed system avoids using the conventional WebRTC components for the management of audio communications, i.e., the MediaStream and RTCPeerConnection. Instead, a new implementation (initially presented in [21]) has been tested, which, after the audio acquisition utilizing an AudioWorklet, extracts the audio samples from the MediaStream and redirects them to an RTCDataChannel for transmission over the network. This implementation allows for controlling the audio processing and communication channel at a very low level, i.e., bypassing the speech-oriented algorithms built into the RTCPeerConnection that alter the audio signal. In addition, the SharedArrayBuffer object used for the data transfer between the AudioWorklet and the main thread showed to be extremely efficient, i.e., with negligible latency and processing overhead. The system shows latencies as low as 40 ms on Firefox on macOS, and in general achieves a latency reduction of 10 to 50% w.r.t. the conventional WebRTC setup. Further details on the architecture of the system and its performance can be found in [18].

(a) Anechoic room setup



(b) Listening test hardware

Figure 2: Listening test performed in the anechoic chamber of Politecnico di Torino [22].

The combination of the AudioWorklet, SharedArrayBuffer, and RTCDataChannel proved to be able to efficiently handle the strict communication requirements of NMP, both in terms of bitrate and frames per second, without any noticeable performance reduction with respect to the conventional RTCPeerConnection implementation.

4. AUDIO QUALITY EVALUATION

From the point of view of the system's audio quality, the performance of this software implementation is compared to that of state-of-the-art NMP software.

To evaluate the best performance achievable by the different applications while excluding the impact of net-

work conditions (that are out of our control of the software engineer), the experiments have been carried out on a wired local network operating at 1 Gb/s, where the impact of transmission bandwidth, latency, and network losses are negligible. In this context, where network delay variability can be assumed more consistent than on the public Internet, a small de-jitter buffer has been used, i.e., a buffer introduced at the receiver side to prevent losses due to late packets. The size of the de-jitter buffer is roughly comparable among the different applications being considered (in the order of 5-9 ms for the native solutions and 20 ms for the web solutions).

4.1 Subjective audio assessment without packet losses

A subjective audio test has been performed to compare the proposed solution with state-of-the art solutions used for remote music interaction during the COVID-19 pandemic.

The test was carried out in an anechoic chamber with a background noise level in accordance with the recommendations of ITU-R BS.1116-1 by 26 subjects who self-declared to have non-impaired hearing conditions and who evaluated four stimuli corresponding to different types of audio selected from the material already used in a previous study [9]. For each stimulus, the uncompressed audio transmitted by the proposed solution was compared to the compressed audio transmitted by the benchmark applications through an ABX-type test [23] presented to each participant through headphones (Sennheiser 600HD).

In this procedure, three stimuli are presented to the listener: stimulus A and stimulus B, which have a known difference, and stimulus X. The listener's task is to identify whether X equals A or B. If there is no audible difference between the two signals, the listener's responses should be binomially distributed, such that the probability of responding $X = A$ equals the probability of responding $X = B$, i.e., 50%. This score is interpreted as an indication of the absence of perceptual differences between A and B.

The minimum number of correct answers needed to indicate a perceptual difference can be given by the inverse cumulative probability of a binomial distribution based on the number of trials, confidence level and probability of correct answer. For the conditions of the test performed, the minimum number of correct answers necessary to indicate a statistically significant perceptual difference is 17.

The four stimuli presented to the subjects represented different audio material: i) a bowed violin playing pitch B6 for 6.7 seconds (Violin), ii) a cello playing pitch C6 for 6.9 seconds (Cello), iii) a female sex soprano singing a C5 for 2.8 seconds (Female NV), and iv) a female sex soprano singing a G4 with operatic vibrato production (Female V). Each stimulus was then recorded after the processing and transmission with each one of the four considered applications: i) the proposed solution, ii) Zoom with "original sound" option turned on, iii) Jitsi with gain control, noise reduction, and echo cancellation disabled, and iv) Google Meet with its default configuration (since it is impossible to customize its audio processing features).

In the first row of Fig. 3, the time and frequency representation of an audio track is presented, consisting of a bowed violin playing the pitch B6 quietly (-12 dB). The subsequent rows display the audio impairments produced by various applications, including, from top to bottom, the proposed solution, Zoom, Jitsi, and Google Meet. For each application, the Signal-to-Noise Ratio (SNR) [24] and the Log-Spectral Distance (LSD) [25] were computed and shown in the figure respectively in the time and frequency domain plot. It is acknowledged that the time domain difference and SNR computation are not significant for audio distortion. In fact, due to the involvement of an analog section in the communication path, a slight misalignment of even a portion of a sample period leads to low SNR, not justified by the listening and by the spectral difference evaluation. The spectral difference representation yields instead significant results, as further demonstrated by subjective listening tests, indicating that the proposed implementation can achieve almost transparent quality even if implemented as a web application. A quality that is clearly superior to the one offered by similar software such as Jitsi and Google Meet. Although there is some distortion at low frequencies in the Zoom plot, the subjective tests evaluated its audio as nearly transparent. It is highly likely that the spectral difference is due to components of the sound that are masked by the original signal and, therefore, imperceptible.

Figure 4 shows the overall results of the subjective experiment. In the case of audio transmitted by the proposed system, only 12-14 subjects out of 26 correctly identified the non-original audio in the various listening tests, while in the case of audio transmitted by a third-party web application, 13-25 out of 26 people correctly identified the non-original audio. As mentioned in the previous paragraph, the results of correct answers above 17 have been considered as statistically significant in order to confirm the perceived differences. The results indicate that the audio of

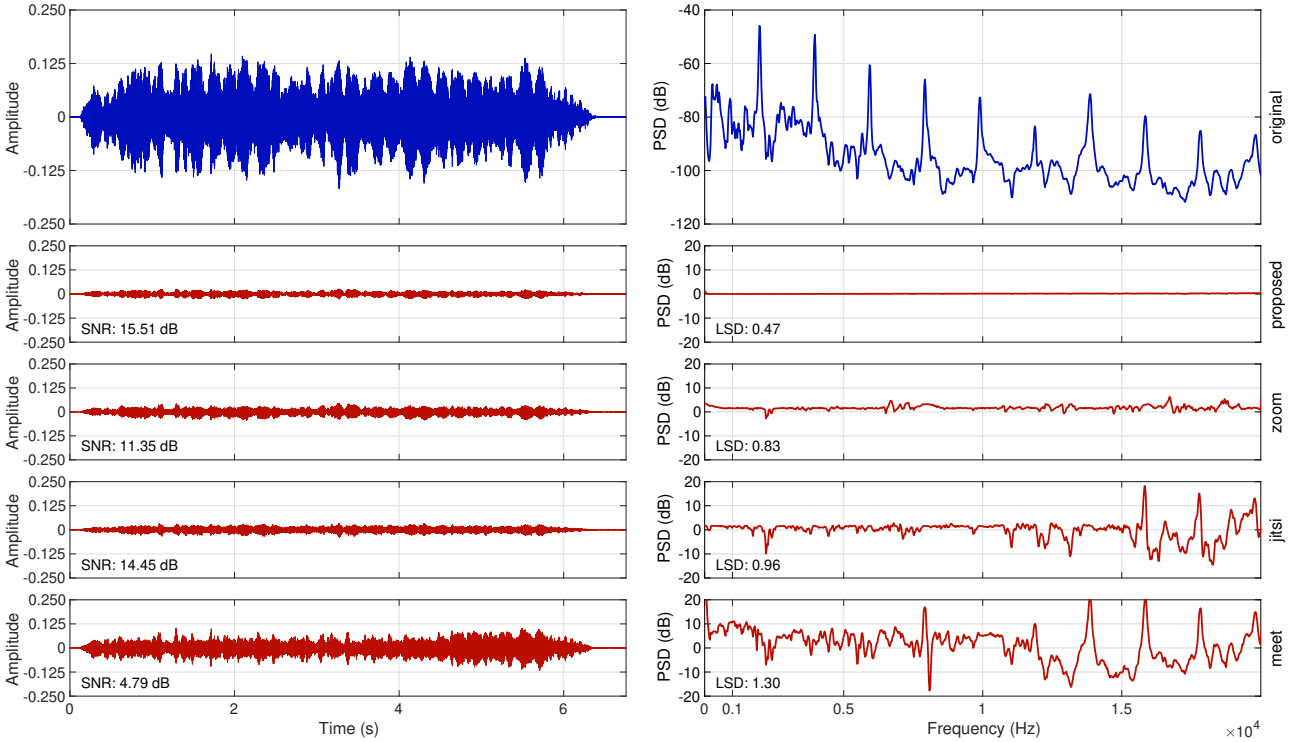


Figure 3: The first row shows the time and frequency domain representation of an audio segment (bowed violin). The following rows show the distortions introduced by the proposed audio transmission framework and by three state-of-the-art videoconferencing solutions, mainly due to perceptual codec compression and automatic digital signal processing applied to the original signal. The plots also indicate the Signal-to-Noise Ratio (SNR) and the Logarithmic Spectral Distortion (LSD) computed between the original and the processed signal for the proposed framework and then for Zoom (with the original audio feature turned on), Skype (with audio processing disabled) and Google Meet.

the proposed system does not show any significant difference for all the stimuli, conversely, Google Meet presents significant differences independently on the type of stimuli. On the other hand, Zoom shows significant differences for the instrumental stimuli, i.e. Cello and Violin, while for Jitsi, a significant difference emerges for the Female singing a C5 (Female NV) only.

4.2 Subjective audio assessment with losses

For the proposed solution, an additional simulation and a corresponding subjective test have been performed to evaluate the effect of audio packet losses on the audio quality perceived by the receiving user. Unfortunately, the same experiment could not be performed on the other software because they are not open source and thus their source code cannot be accessed to simulate the effect of packet losses and their concealment.

Whenever an audio packet is unavailable for playback at the receiver in due time, either because it was dropped by an intermediate network node or because it was excessively delayed, a gap may occur in the playout. Since the retransmission of the data from the sender is not applicable due to the strict latency constraints of NMP, Packet Loss Concealment (PLC) techniques are usually employed to reconstruct the original audio frame in order to reduce the auditory perception of an artifact or glitch [26].

In the following, we present the performance of the concealment technique described in [27], when applied to the audio playout streams generated by the NMP appli-

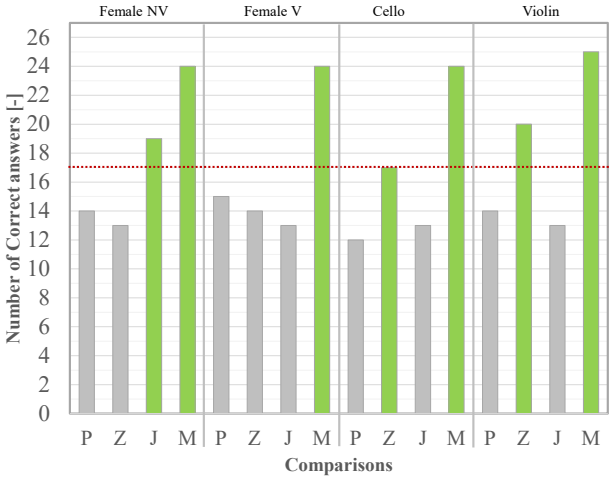


Figure 4: Results of the subjective experiments with no losses. Four stimuli – female sex soprano (Female NV), female sex soprano with vibrato (Female V), cello (Cello), and bowed violin (Violin) – processed by four different systems – the proposed one (P), Zoom (Z), Jitsi (J), and Google Meet (M) – have been presented to 26 subjects. The statistical significance threshold of correct answers is 17.

cation presented in this paper. More in detail, the missing audio frames are reconstructed by means of an autoregressive model that is capable of predicting and synthesizing the lost audio section based on its historical trend, without introducing any additional delay [28]. This technique is compared with two standard concealment algorithms, typically implemented in NMP software, that fill the audio gap with silence (silence substitution) or with the repetition of the last received audio frame (pattern replication).

Two audio files with a duration of 5 seconds were selected from six audio categories (violin, drums, piano, guitar, songsm and generic audio) for a total of twelve audio samples. Each sample was then divided into audio frames of 128 samples, i.e., about 5.8 ms at 22'050 Hz. Packet losses were simulated by selecting twenty random frames from each sample that have been reconstructed with the three PLC techniques. During the subjective quality test, each subject was presented with twelve questions, one for each audio sample. First, each subject was asked to listen to the original audio version and then to the three reconstructed signals, in a random order, without any indication of the corresponding concealment technique. Finally, each subject was asked to select, out of the three reconstructed signals, the one which sounded more similar to the original version. This test involved a total of 25 subjects. A minimum number of 17 answers are required to indicate a statistically significant preference for the AR models PLC technique w.r.t. the traditional PLC techniques.

The results of the test are shown in Tab. 1. In most cases, the most effective PLC technique is the one based on AR models. In particular, AR models show particularly good performance with signals that contain sustained sounds, as in the case of piano, violin, and guitar, while their benefits are reduced in the case of sounds containing abrupt transients, such as drums. This is due to the fact that, since AR models predict future samples based on the past history of the audio stream, sustain-oriented signals have low variability, thus it is easier to predict their future evolution, whereas transient-oriented signals have high variability, thus their prediction is less accurate. When songs are considered, no clear preference is revealed based on the test results, since in that case, due to the presence of many audio sources, audio gaps are less noticeable, and, as a consequence, also the difference between the concealment techniques is not particularly evident to human listeners.

5. CONCLUSIONS

The development of a unified system, both web and native, for Networked Music Performances is still faced with technical challenges. Although it is relatively straightforward to transfer most applications to a web environment, the limited integration of browsers with operating systems poses a significant obstacle in meeting the high-quality and low-latency requirements for real-time applications. Nonetheless, the current implementation offers a web-based environment for real-time transmission of high-fidelity uncompressed stereo audio, suitable for music interactions, that addresses the limitations typically associated with web-based NMP systems, such as poor audio quality, changes in the perceived tempo or volume, and high latency, all of which can compromise the effectiveness of the musical interaction.

6. ACKNOWLEDGMENTS

This work has been partially supported by Fondo Integrativo Speciale per la Ricerca (FISR) of the Ministero dell'Università e della Ricerca (MUR), grant number

Table 1: Subjective test of three PLC techniques: silence substitution, pattern replication, and autoregressive models. Each column contains the number of subjects that perceived the relative PLC technique variant closer to the original version. A total of 25 subjects were involved.

Audio sample	Silence substitution	Pattern replication	Autoregressive models
Violin A	0	2	23
Violin B	0	1	24
Drums A	2	5	18
Drums B	5	3	17
Piano A	3	2	20
Piano B	0	4	21
Guitar A	1	3	21
Guitar B	2	1	22
Song A	5	7	13
Song B	7	6	12
Generic A (trumpet)	0	1	24
Generic B (voice)	1	4	20

FISR2020IP_00156, “HiFiReM - High fidelity system for remote music teaching with synchronised and collaborative live concert capabilities”.

7. REFERENCES

[1] Zoom Video Communications, Inc., “Zoom.” [online] https://www.zoom.us.

[2] Microsoft Corp., “Skype.” [online] https://www.skype.com.

[3] Google LLC, “Google Meet.” [online] https://meet.google.com/.

[4] J.-P. Cáceres and C. Chafe, “JackTrip: Under the hood of an engine for network audio,” *Journal of New Music Research*, vol. 39, pp. 183–187, Nov. 2010. https://dx.doi.org/10.1080/09298215.2010.481361.

[5] A. Carôt and C. Werner, “Distributed Network Music Workshop with Soundjack,” in *Proceedings of the 25th Tonmeistertagung*, (Leipzig, Germany), 2008.

[6] V. Fischer, “Case study: Performing band rehearsals on the Internet with Jamulus,” *URL: https://jamulus.io/PerformingBandRehearsalsontheInternetWithJamulus.pdf*, 2015.

[7] Jesse Chappell, “SonoBus.” [online] https://sonobus.net/.

[8] JamKazam Inc., “JamKazam.” [online] https://jamkazam.com/.

[9] I. Howell, K. J. Gautereaux, J. Glasner, N. Perna, C. Ballantyne, and T. Nestorova, “Preliminary Report: Comparing the Audio Quality of Classical Music Lessons Over Zoom, Microsoft Teams, VoiceLessonsApp, and Apple FaceTime.” [online] https://www.ianhowellcountertenor.com/preliminary-report-testing-video-conferencing-platforms, mag. 2020.

[10] JackTrip Foundation, “JackTrip Lab.” [online] https://www.jacktrip.com/.

[11] Elk, “Elk Audio.” [online] https://www.elk.audio/start.

[12] Syneme, “Artsmesh.” [online] https://www.artsmesh.com/.

[13] CultureHub, “Livelab.” [online] https://www.culturehub.org/livelab.

[14] K. Tsioutas and G. Xylomenos, “Audio delay in web conference tools,” in *Proceedings of the 7th International Web Audio Conference (WAC)*, (Cannes, France), Jul. 2022.

[15] C. Rottondi, C. Chafe, C. Allocchio, and A. Sarti, "An overview on networked music performance technologies," *IEEE Access*, vol. 4, pp. 8823–8843, Dec. 2016. https://doi.org/10.1109/ACCESS.2016.2628440.

[16] P. Adenot and H. Choi, "Web Audio API." W3C, June 2020. [online] https://www.w3.org/TR/webaudio/.

[17] H. Boström, J.-I. Bruaroey, and C. Jennings, "WebRTC 1.0: Real-time communication between browsers." W3C, Dec. 2020. [online] https://www.w3.org/TR/webrtc/.

[18] M. Sacchetto, P. Gastaldi, C. Chafe, C. Rottondi, and A. Servetti, "Web-based networked music performances via WebRTC: a low-latency PCM audio solution," *Journal of the Audio Engineering Society*, vol. 70, no. 11, pp. 926–937, 2022.

[19] M. Sacchetto, C. Rottondi, A. Servetti, L. Shtrepi, M. Masoero, and A. Valle, "HiFiReM: a unified platform, both web-based and native, for remote music education," in *Proc. XXIII Colloqui di Informatica Musicale*, (Ancona, Italy), Oct. 2022.

[20] S. P. Lipshitz and J. Vanderkooy, "Pulse-code modulation – an overview," *Journal of the Audio Engineering Society*, vol. 52, pp. 200–215, Mar. 2004.

[21] M. Sacchetto, A. Servetti, and C. Chafe, "JackTrip WebRTC: Networked Music Experiments in a Web Browser," in *Proc. of the 6th International Web Audio Conference (WAC)*, (Barcelona, Spain), Jul. 2021.

[22] L. Shtrepi, S. Di Blasio, and A. Astolfi, "Listeners sensitivity to different locations of diffusive surfaces in performance spaces: The case of a shoebox concert hall," *Applied Sciences*, vol. 10, no. 12, 2020.

[23] W. Munson and M. Gardner, "Standardizing Auditory Tests," *Journal of the Acoustical Society of America*, vol. 22, no. 4, pp. 675–675, 1950. https://doi.org/10.1121/1.1917190.

[24] N. S. Jayant and P. Noll, *Digital coding of waveforms. Principles and applications to speech and video*. Elsevier, 1985.

[25] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Prentice-Hall, Inc., 1993.

[26] C. Perkins, O. Hodson, and V. Hardman, "A survey of packet loss recovery techniques for streaming audio," *IEEE Network*, vol. 12, no. 5, pp. 40–48, 1998.

[27] M. Sacchetto, Y. Huang, A. Bianco, and C. Rottondi, "Using Autoregressive Models for Real-Time Packet Loss Concealment in Networked Music Performance Applications," in *Proceedings of the 17th International Audio Mostly Conference*, (New York, NY, USA), pp. 203–210, Association for Computing Machinery, 2022.

[28] G. Zhang and W. B. Kleijn, "Autoregressive model-based speech packet-loss concealment," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4797–4800, IEEE, 2008.