Real-time Music Collaboration and TSN for Service Providers

Norman Finn, Huawei Technologies Co. Ltd
April 19, 2021

In March, 2020, hundreds of thousands of performing ensembles around the world suddenly had to stop rehearsing in person. Singing in a choir, for example, is one of the most efficient means known to spread an air-borne virus. Many of these groups have ceased to function entirely, and many have resorted to time-tested recording techniques to continue operation. But, many have starting using a new class of Internet applications, real-time music collaboration, that have a real potential for sparking a new class of internet services.

Since the invention of earphones, musicians have used sound-on-sound recording to make music, by combining multiple recorded and/or live performances into a new recording. As computers have blossomed, non-real-time applications have emerged, both on-line and locally-hosted, for collaborative music-making in a similar manner. Examples are BandLab (http://www.bandlab.com/) and Audacity (https://www.audacityteam.org). With these tools, a timing track ("click track"), which can be audio, video, or both, is first recorded. Each performer records a performance while viewing and/or listening to the click track. Afterwards, an editor combines the separate tracks (without the click track), adjusting the individual performances using sound and/or video editing tools, into a final performance.

While such tools are useful for a musical ensemble to record a virtual performance, they are useless for developing, learning, or rehearsing new works, which have always required the whole group to be in the same room, working together. At this writing, in-person rehearsals have been inadvisable, or even illegal, for almost a year.

Most groups have fallen back to virtual meetings, using on-line meeting software such as Zoom (http://zoom.us), Cisco Webex (http://webex.com), or Google Hangouts (http://meet.google.com/meet). On-line meeting tools allow the group to talk and share presentations and performances, but not to rehearse. The reason is high, and variable, latency. We define "latency" here as the time between one person making a sound and another person hearing it.

Let us take a 50-person choir as an example. When rehearsing together, the group stands or sits within a physical space with a radius of perhaps 5-10 meters. Given that the speed of sound is approximately three milliseconds per meter (3 ms/m), the singers have an average latency of approximately 20 ms (7 times 3). A smaller group, such as a five-person band, can have latency as low as 3 to 5 ms. Performers have, for thousands of years, been accustomed to these typical latencies. They are trained to generate sound in synchrony, by singing or playing

at the right time with respect to the tempo of the work, and not to wait until they hear the other musicians. In other words, they are trained to adjust for some amount of latency.

When an ensemble tries to rehearse using on-line meeting tools (Zoom, Webex, Hangouts), the result ranges from humorous to tragic, depending on one's perspective, but it is never satisfactory. The reason is that the mouth-to-ear latency is much larger, typically 100 - 500 ms, and varies widely across different pairs of participants; it even varies with time for the same pair. Such latency is not a serious problem for meetings, where one person speaks at a time. It is disastrous when all try to make music at the same time.

So, a new range of real-time music collaboration tools have become widely used since March, 2019. Such tools include Jamulus (http://jamulus.io), Jamkazam (http://jamkazam.com) and JackTrip (https://www.jacktrip.org). The intent of these tools is to reduce mouth-to-ear latency over the internet to a time equal to that of a rehearsal room. All of these tools require the participants to use wired connections within their homes, instead of Wi-Fi; Wi-Fi alone adds tens of milliseconds delay, with considerable variation. All use short (1 to 10 ms) analog-to-digital buffers, and transmit the packets of the stream independently (UDP) to a server, which mixes the sound and transmits the mix back to the participants. Across a large metropolitan area, the best-case internet latency is only a few milliseconds, which means that total mouth-to-ear latency can be as low as 3-5 ms. But best-case is not what matters.

In order to prevent drop-outs due to variability in the latency across the internet, the receiving computer must buffer some number of milliseconds of data before presenting it to the earphones. Even if 80% of the packets have 3 ms latency across the internet, if there is no buffering delay at the receiver, that will mean that 20% of the packets arrive after it is time to present them to the earphones, which results in silence 20% of the time, randomly placed in 5 ms segments. This choppy sound is unusable. It is quite common for a few percent of the packets to be delivered with a latency three or four times larger than the best-case packets. Therefore, either the user or the software has to adjust the receive buffer size to strike a balance between latency and drop rate. 20 to 50 ms round-trip latency is considered good by these tools, at this time.

The latency variation vs. packet loss tradeoff is, of course, typical of applications that make real-time demands on best-effort services, and it is precisely the kind of problem for which the Time-Sensitive Networking (TSN) Task Group of IEEE 802.1 is developing standards (http://1.ieee802.org). TSN standards provide for making a reservation across a network for a stream of data. This reservation allocates resources such as bandwidth and buffer space to that stream along a particular unicast or multicast path from a talker to one or more listeners across a single management domain. The network can then guarantee zero congestion loss and a consequent worst-case latency for any packet in that stream for the duration of the reservation.

An additional need is for a one-way video for the director. Of course, it would be nice for people to be able to see each other, but this is generally not necessary. But, the visual clues

given by the director of the performance group are critical to producing music, whether at rehearsal or at a performance. This is not a trivial task, as the time to collect a video frame at the typical 30 frames/second is 33 ms. Equivalent to 11 meters of sound distance, this is beyond the sound latency requirement. Higher frame rates are clearly necessary, perhaps at the expense of latency.

A rehearsal lasts for one to several hours, with a relatively stable set of participants, generally within a metropolitan area. (Light travels roughly 300 km/ms, so trans-continental rehearsals are only marginally possible.) This makes the reservation process perfectly feasible. The service provided by TSN, zero congestion loss and worst-case end-to-end latency, allows the application to minimize its receive buffers to exactly that delay. In fact, all participants can have their buffering delay adjusted to match the latency of the least-fortunate participant, so that all are synchronized. This application can command a fee—the cost of the rehearsal room is saved. Overall latency can be brought down to in-person speed-of-sound values. In short, real-time music collaboration is a perfect match for IEEE 802.1DF TSN Profile for Service Provider Networks (https://1.ieee802.org/tsn/802-1df).