

DISTRIBUTED INTERNET REVERBERATION FOR AUDIO COLLABORATION

CHRIS CHAFE

Center for Computer Research in Music and Acoustics (CCRMA), Stanford University, California, USA
cc@ccrma.stanford.edu

Low-latency, high-quality audio transmission over next-generation Internet is a reality. Bidirectional, multichannel flows over continental distances have been demonstrated in musical jam sessions and other experimental situations. The dominating factors affecting delay are no longer system issues, but the transmission time bounded by lightspeed. This paper addresses a method for creating shared acoustical spaces by “echo construction.” Where delays in bidirectional paths are sufficiently short and “room-sized,” they can be used to advantage as components in synthetic, composite reverberation.

INTRODUCTION

The project involves setting up two collaborating audio hosts separated by short round trip delay times (e.g., between San Francisco and Seattle, $RTT = 20ms$). Monitoring on both ends includes a composite reverberation in which the round-trip delay is used to construct multipath echoes, corresponding to multiple “rays” in a composite room.

The first implementation involves simulation of two identical rooms with identical monitoring (mic and speaker locations). For simplicity, the rooms can be thought of as small, 3 m on a side. Using the technique described, a composite room is heard which incorporates the 10 ms (one-way) network delay in a synthetic reverberation circuit running in software as part of the audio transmission system. The added 10 ms roughly corresponds to an additional 3 m inserted between the monitoring locations. The listeners have the impression of communicating with each other in the same 9 m room.

The paper describes the audio transmission techniques, reverberation circuit, its extension to multichannel monitoring and initial evaluation of this “echo construction” method.

1. PRO AUDIO OVER THE INTERNET

Several successful trials have demonstrated concert-quality, interactive audio streaming technology. The first experiments were unidirectional [1], followed shortly after with bidirectional tests [2]. Audio quality presently features:

- High resolution (e.g., 24 bit, 96 kHz)
- Multi-channel (including ambisonic format)
- Minimum latency (when interactive)

Experiments have utilized internets with “fat pipes” such as Internet2’s *Abilene* network in the USA and Canarie’s *CA*net4* in Canada. Both provide backbones (nominally 10 Gb/s at present) connecting large consortia of subscribing universities and research organizations. These

testbeds have allowed groups at McGill University and Stanford to gain experience with jam sessions and teleconferencing concentrating on audio quality. Both teams have been developing on Linux-based commodity hardware (with standard networking components) [3]. The variety of low-latency streaming applications being explored includes low-latency, high-resolution video (at McGill’s Centre for Intelligent Machines) and audio-based techniques for network monitoring (at Stanford [4]).

This is the early stage. In the not-too-distant future, common Internet access to home and business will be sufficiently provisioned to carry the kinds of flows seen in our recent high-quality audio collaborations.

1.1. Latency Issues

Real-time audio signals don’t arrive immediately at the receiver. Several sources of latency can delay the arrival, though not by much. If the overall acoustic round trip (ART) is short enough, the impression of hearing one’s own sound return is the same as hearing it bounce back from a nearby wall or object. At roughly 3 ms/m, reflections off nearby network hosts and nearby walls contribute somewhat the same effect to the original sound.

In a simpler world ART delay would only be able to be blamed on network transmission time, itself only a function of distance. The actual world in which we’ve been developing this technology introduces other delays which, on the network side, are attributable to best-effort policies for packet routing and delivery. And since our world also includes time-shared operating systems, the applications themselves must be designed with rugged internal audio guarantees.

Fig. 1 shows a 12-hop path from Stanford to Princeton mapped using the *traceroute* utility. As can be seen from the first 5 hops, intermediate routers (network junctions) contribute only minimal delay when forwarding the signal ($< 1ms$). Transmission time is largely a function of distance for this Internet2 path.

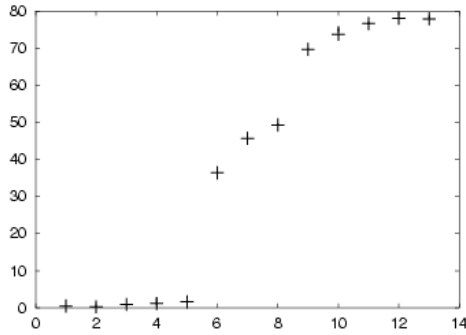


Figure 1: Round trip delay in milliseconds between Stanford and Princeton measured by pinging intervening routers. The bulk of the delay incurred is distance-related. Routers near the edges are in close temporal proximity to one another.

The *ping* utility sends an ICMP echo request (a small packet to be echoed back from the target host) and measures the reply’s time of arrival. An important footnote to Fig. 1 is that traceroute can only summarize conditions existing on the tested path. Each router has been pinged separately by the originating host computer, rather than in relay fashion. The full path exhibits a slight amount of jitter (variance) under rapid pinging as shown in Fig. 2.

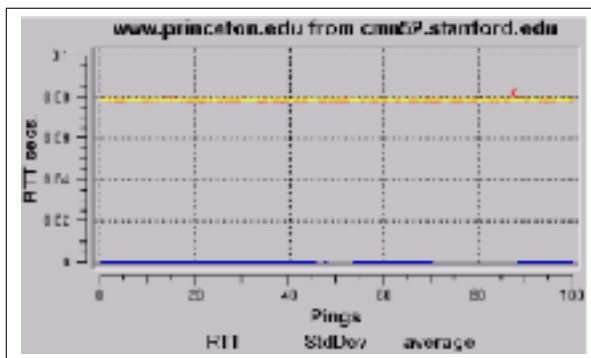


Figure 2: Ping sequence to Princeton. Red curve is RTT per ping, yellow is windowed average, and blue windowed standard deviation (jitter), pings launched at $RTT + 10ms$ intervals.

1.2. ART in Practice

An application buffers against network jitter by maintaining a so-called playback buffer. The buffer smooths out delivery but introduces latency. At its lowest delay setting it is tuned to accommodate the maximum expected jitter.

Operating system dependent buffering also contributes to overall audio delay. The present setups (with non-real-time OS’s) depend on a combination of priority assign-

ment and further application buffering to accommodate those brief intervals when the OS’s scheduler needs to process someone else’s task (e.g. disk access, network access, etc.). Again, in a different world, many of these uncertainties would not be present, for example with the use of dedicated hardware and / or isochronous networks. In practice, $ART \approx RTT + 20ms$ on the present systems.

1.3. Network Zones and QoS

Latency zones encompass a network host in onion-like layers radiating outward. The closest zone lies within the host itself, where packets stream between the application and the network device and latency is dominated by OS-dependent quantities. Next in line, moving outward toward the network “cloud,” is the local area network, dominated by jitter. Finally, wide area network latency depends on distance as already described, as well as the quality of service (QoS) provided by the networks traversed.

As can be seen from the path in Fig. 1, transit times at the edges of the long-haul segment contribute relatively little to the overall latency. This would not be the case were our flows in contention with other traffic, with one or more of the routers heavily loaded.

1.4. Recirculating Audio

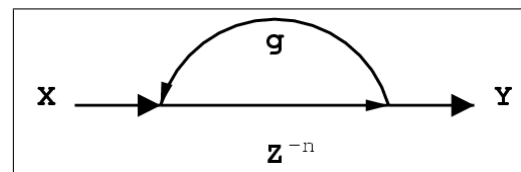


Figure 3: Comb filter.

Comb filters are closed-loop circuits with feedback around a delay line. The signal flow graph of Fig. 3 is adopted from Moorer’s important early paper on synthetic reverberation [5], where X indicates input, Y output, Z^{-n} delay of n audio samples, and arcs are unity gain paths unless specified.

The circuit can be modified to use a bidirectional network stream involving two interacting audio hosts, by splitting the delay Z_{comb}^{-n} into the network’s round trip lag [6]. The modification is shown in Fig. 4.

1.5. Perceptual Effects of Latency

As delay is increased, the filter’s effect varies through a familiar perceptual continuum from flange to pitch to echo (F-P-E). Comparing the above categories of network RTT to the F-P-E continuum produces Table 1, in which comb filters implemented over different network paths have a different type of sound.

NETWORK	delay (ms)	freq (Hz)	PERCEPT
localhost	< .05	> 20k	ultrasound
LAN	.25	4k	highest pitch
WAN (1)	25	40	lowest pitch
WAN (2)	50	20	infrasound
WAN (3)	100	10	echo
WAN (4)	200 - 500	5 - .5	rhythm

Table 1: RTT and *echo percept* for various network types: localhost internal to the computer, local area network (LAN), and wide area network (WAN) over next-generation (NGI) backbones from Stanford to 1) Seattle 2) Dallas 3) Montreal or over 4) commodity Internet to global sites.

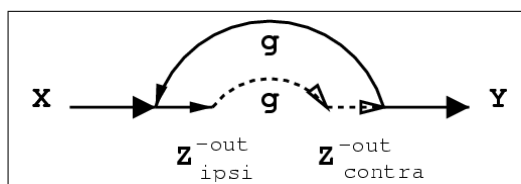


Figure 4: Network comb filter from the point of view of either host. Outgoing delay across the network to the opposite host is Z^{-out} . Portions running on the ipsilateral host are marked by solid lines, and dotted for contralateral.

Longer echoes plague telephony and VoIP, making echo cancellation techniques essential for these longer circuits. The present work concentrates on the region of sub-50 ms round trip delay where reflection times are too short to be perceived as echo. In this region, comb filter effects create a “ringy” tone superimposed on signals flowing across the path. The impulse response of such a comb filter (with attenuating values of g) is shown in the signals observed in Fig. 5. As with further simulations included below, $RTT = 50ms$ and both ipsilateral and contralateral outputs are shown.

2. ECHO CONSTRUCTION

Comb filters are a basic ingredient of the Schroeder-style reverberators described in detail by Moorer [5]. The following method uses several network comb filters (Fig. 4) to implement a distributed version of this class of reverberator. The result is a more complex recirculating audio path which eliminates the problem of ringiness in a closed-loop monitoring situation. Diagrammed in Fig. 6, $NRev$ is the representative Schroeder-style reverberator that was chosen for the initial implementation. It’s a circuit which has been around since the 80’s and has been implemented in numerous open-source [7] and commercial packages.

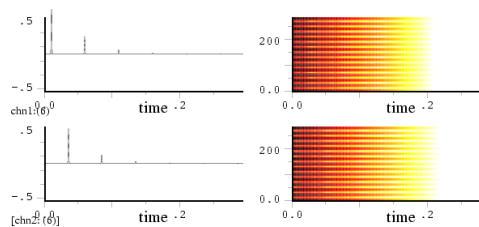


Figure 5: Impulse responses of ipsilateral and contralateral hosts resulting from an impulse into the ipsilateral side (top). RTT of 50 ms creates a comb filter effect resulting in spectrograms showing a pitched, harmonic response (the comb teeth are spaced 20 Hz apart).

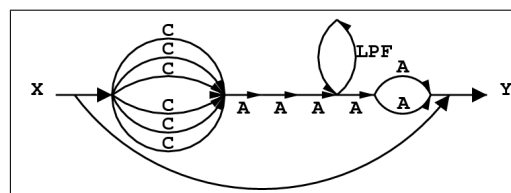


Figure 6: $Nrev$, an example of a classic Schroeder-style reverberator composed of 6 parallel comb filters, 6 all-pass delays, a one-pole low-pass filter, and direct signal path.

2.1. Implementation of a Network Reverberator

$Nrev$ components include a parallel bank of comb filters, feeding a cascade of all-pass delays and a low-pass filter. To convert the patch into a network circuit, combs are split to produce the bidirectional circuit shown in Fig. 7. Two conjoined instances of $Nrev$ reverberators are used. One runs on the ipsilateral host and employs 6 network comb filters in parallel, each of which requires a bidirectional audio channel connected to the opposite host. The contralateral duplicate shares the same comb structure. From the ipsilateral point of view (shown), the contralateral host inserts and taps off the bank of combs at its midpoints. The resulting reverberator is identical from either perspective. All comb filter channels traverse the network loop in multichannel sample frames and are kept sample synchronous with each other.

$Nrev$ ’s comb and all-pass delay times are mutually prime to avoid coincident resonances and to reach the desired overall resonance complexity. Unmodified, all network comb delays would experience the same RTT and would exhibit the same resonances, a problem which requires that they be individually lengthened. After lengthening, the signal flow for such a comb unit consists of 4 portions:

$$Z_{ipsi}^{-out} \quad Z_{contra}^{-out} \quad Z_{ipsi}^{-ext} \quad Z_{contra}^{-ext}$$

A graph of the complete network comb filter with extensions is shown in Fig. 8 along with locations of input

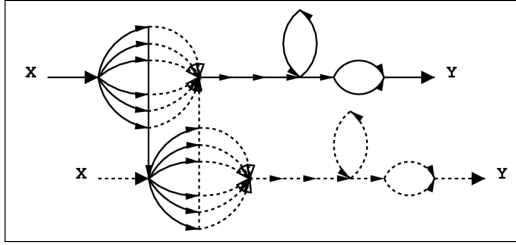


Figure 7: Conjoined Nrev reverberators. A bank of network comb filters (like the one in Fig. 4) is shared by ipsilateral (top) and contralateral reverberators.

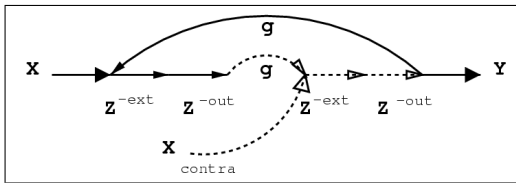


Figure 8: Network comb filter showing extension delays for tuning to desired length (when $RTT < Z_{comb}^{-n}$).

insertion points for the ipsilateral and contralateral sides. Typical settings for each of the 6 comb delays in a stand-alone version of Nrev are listed in the following table, along with the values recasting them into network portions and complementary non-network extensions. The second set of values constitute the delays for one side of the shared comb filter bank shown in Fig. 7: (in ms)

Nrev	55	62	72	80	87	93
Z_{ipsi}^{-out}	25	25	25	25	25	25
Z_{ipsi}^{-ext}	2.5	6	11	15	18.5	21.5

An additional bidirectional channel (not shown) provides the direct signal path. It runs parallel to the reverberator and provides the “dry” side of the “dry / wet” mix which is controlled in the usual fashion. The only difference is that here, 100% dry corresponds to the recirculating, ringy loop described above. Fig. 9 is an impulse response of the entire system. The impulsive excitation is injected into the ipsilateral side and, observed from the contralateral side, can be seen arriving ahead of the reverberation (via the separate channel added for the direct signal path). The delayed direct signal arrives after $Z_{ipsi}^{-out} = 25ms$, and in effect creates a pre-echo ahead of the reverberation (which begins with the first reflection at $Z_{ipsi}^{-out} + Z_{contra}^{-ext} = 27.5ms$). As for the ipsilateral side, the effect is the same as if it were a stand-alone version of Nrev (with direct signal also mixed, undelayed, to the output).

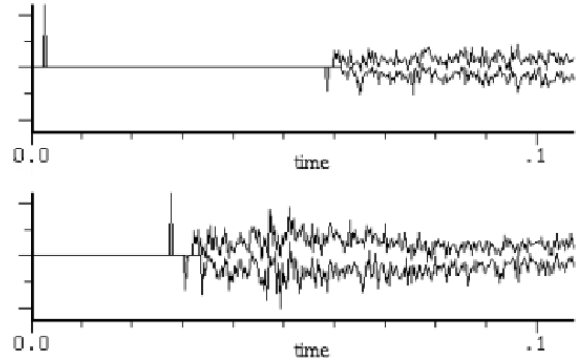


Figure 9: Impulse response of Fig. 7 plus a direct signal path. The impulse is introduced at the ipsilateral side (top) and first appears as a pre-echo via the direct signal path to the contralateral output (bottom).

3. COMPARED TO WHAT?

A distributed internet reverberator (for audio collaboration) — DIRAC — simulates two sources enclosed in a room, each with its own perspective on the geometry of sources and reflections. The Nrev-based implementation of Fig. 7 is useful for analysis of DIRAC’s basic features. Suppose, for example, two sources are placed symmetrically near the opposite ends of a 9 m room, and for simplicity the floor and ceiling reflections are ignored. An exchange of direct signals would follow the same time course measured in the impulse response of Fig. 9. A real world signal arriving at the contralateral side would receive early reflections following the direct signal and then late reverberation.

The room also feeds back reverberation to the source emitting the sound, an effect which Nrev only simulates with late reverberation (since it has no provision for early reflections). The present implementation intentionally ignores any ipsilateral reflections during the “priming” period of the comb filter bank, opting instead to allow actual, live reflections from the originating studio to fill in this interval.

Now imagine two studio booths isolating two players who monitor one another via mics and loudspeakers in order to play together. In this case, isolation kills the direct path. DIRAC’s feed-forward signal provides a substitute direct path but with a delay of Z_{ipsi}^{-out} .

Early reflections from the musician’s booth dominate the ipsilateral ambience (created by surfaces surrounding the location in the real room e.g., objects, nearby walls, etc.). These fill in the priming period up to the moment in which the synthetic reverberation arrives and mixes with the real.

3.1. Real Rooms Networked Together

In the example, identical synthetic reverberation is added to the actual room ambience on both sides (both are running instances of the same Nrev tunings). Yet, in neither case is the sense of nearby real space diluted, since the added portion is late reverberation (and doesn't compete with the perceptually more important early reflections which are arriving from surrounding walls and objects). The intended effect is a shared synthetic acoustical space within which the immediate locales can retain their real room cues.

Ultimately, it will be advantageous to explore calibrated ways of grafting the synthetic room onto the real room. Synthetic early reflections could be added to Nrev by passing the incoming contralateral signal through a number of delays tuned according to the geometry of the listening space. Calibration would be carried out to match synthetic early reflection times and spatial locations to real early reflections measured at the (ipsilateral) listening position.

The two synthetic reverberators would usually differ. Only in the case of identical rooms with identical listening positions would they be the same.

3.2. Multichannel Extension

The present implementation is single-channel (Nrev only provides one input and one output). Multichannel reverberation, for example Nrev's cousin *Freeverb*, provides more than one input point into a structure feeding more than one output. Freeverb utilizes a separate bank of parallel combs for each input, each with slightly decorrelated delay times (8 combs and 4 all-pass filters per input). The number of bidirectional audio channels for a network version of Freeverb would scale with the number of inputs, $N_{chans} = N_{inputs} * (N_{combs} + direct)$.

What's the limit? Practically speaking, probably quite high. An 8-input reverberator created ala Freeverb, would require 72 bidirectional channels. At ≈ 1.2 Mb/s per channel for 24 bit / 48 kHz signals, the network load (each direction) would be 83 Mb/s, well within the capability of present networks and network hardware.

3.3. Other Reverberator Structures

Constructing different implementations of DIRAC should be possible based on alternative types of digital reverberation. The only requirement is that the method have separable delays which can be broken out and replaced by the network latency.

Convolution-based (sampling) reverb has recently become a practical means for achieving high-quality reverberation. Systems include multichannel versions running multiple simultaneous convolutions. A network version would require time-shifting the sampled impulse responses by the amount of network delay. Such shifting would be

limited to cases where the network delay is less than the reverb radius. It is intriguing to think of capturing opposing impulse responses from positions in a larger space and then imprinting smaller, interacting spaces with their acoustic.

Waveguide reverb [8] is also a likely technique. 2-D waveguide meshes (computed in real time) simulate the major propagation paths between each input and output. A network version of waveguide reverb could be designed with the explicit geometry of the shared synthetic space in mind. Rather than using the more abstract signal-based approach of a Schroeder-style reverberator or the non-parametric snapshots of sampling reverb, a waveguide approach would directly model the physical aspects of the intended space allowing parametric adjustments within a physically meaningful representation.

4. CONCLUSIONS

The goal is to apply echo construction techniques as a means of increasing the quality of networked audio collaboration. The best setup one can envision would use surround loudspeaker monitoring and good multichannel microphone pickup. Given these ingredients and low-latency transmission it is possible to synthesize a shared, enclosing acoustical space which will ultimately be *geometrically consistent* with real rooms. A first attempt, to study the fundamentals of DIRAC in simulation, has been presented.

The possibility for exploring something like DIRAC actually grew out of an altogether different project involving music synthesis. Physical model simulations of instruments are constructed from elements much the same as those described above, and comb filters with pitch-length delays are essential. The SoundWIRE project [9] has researched distributed physical models that span bidirectional paths, substituting the network delay for delays typically implemented in local computer memory. The technique is useful for evaluating networks because the resulting synthesized sound displays to the ear important features of the path's QoS in a very intuitive way (i.e., pitch = RTT, vibrato = jitter, and glitches = packet loss). The idea of using the same distributed framework for reverberation suggested itself after noting that talking and playing music together over the connections worked surprisingly well but was colored due to comb filter effects. As would be expected, those situations were highly sensitive to monitoring and feedback issues. Breaking the closed-loop monitoring path by using headphones on at least one side was necessary. Either that, or pulling down the monitor levels to inconveniently low levels.

DIRAC presents a different solution by using the inherent comb filter of the monitoring path to advantage in synthetic reverberation.

REFERENCES

- [1] A. Xu and J.R. Cooperstock, "Real-Time Streaming of Multichannel Audio Data over Internet" *Proc. 108th Convention of the Audio Engineering Society*, Paris (2000).
- [2] C. Chafe, S. Wilson, R. Leistikow, D. Chisholm, and G. Scavone, "A Simplified Approach to High Quality Music and Sound Over IP" *Proc. COST-G6 Conference on Digital Audio Effects (DAFx)*, pp. 159-164, Verona (2000).
- [3] Planet CCRMA,
<http://www-ccrma.stanford.edu/planetccrma/software/>
- [4] C. Chafe, R. Leistikow, "Levels of Temporal Resolution in Sonification of Network Performance" *Proc. 2001 Intl. Conference on Auditory Display*, Helsinki (2001).
- [5] J. A. Moorer, "About This Reverberation Business" *Computer Music Journal*, vol. 3(2), pp.13-18 (1979).
- [6] C. Chafe, S. Wilson, and D. Walling, "Physical Model Synthesis with Application to Internet Acoustics" *Proc. 2002 Intl. Conference on Acoustics, Speech and Signal Processing*, Orlando (2002).
- [7] Synthesis Toolkit (STK),
<http://www-ccrma.stanford.edu/software/stk/>
- [8] J. O. Smith "A New Approach to Digital Reverberation using Closed Waveguide Networks" *Proc. Int. Computer Music Conf.* pp. 47-53, Vancouver (1985).
- [9] SoundWIRE,
<http://www-ccrma.stanford.edu/groups/soundwire/>