# Media is not Data: The Meaning of Fairness for Competing Multimedia Flows

Timothy B. Terriberry
<tterriberry@mozilla.com>

**Abstract**

For typical data-oriented networking applications, the notion of "fairness" asserts that each flow should receive a "fair share" of the available bandwidth. With little a priori information about the data in a flow, "fair" typically means "equal". However, "fairness" for media is better expressed in terms of the perceptual quality delivered to the far end. Under this definition, the relative share given to each flow by a "fair" allocation can change drastically as the total channel capacity changes or even the contents of the media itself changes.

There are a number of different factors which affect the perceived quality of audio and video data. For video, simple objective metrics like the Peak Signal-to-Noise Ratio (PSNR) or the Structural SIMilarity index (SSIM) can measure the degradation introduced by lossy compression. These are the first things that video engineers think of when someone says "quality". The relationship between bitrate and quality is not linear. Most compression algorithms have a "knee" in the quality/bitrate curve, beyond which small reductions in bitrate produce drastic reductions in quality.

The bitrate required to achieve a given PSNR or SSIM score can also vary wildly, by as much as an order of magnitude, depending on the content being coded. Something as simple as the background noise level can have a significant effect. A camera that pans between speakers or a very animated speaker gesturing with their arms requires much more bitrate to achieve a given quality than an idle listener in front of a static background. Even if these events are only transitory, perception of quality is often dominated by the worst moments in a call, rather than the average [XZ06]. Therefore a congestion control algorithm that partitions the available bandwidth equally to each flow (or even into unequal, but consistent fractions of the total) results in considerably suboptimal quality.

PSNR and SSIM, useful as they are, can only measure quality against a known reference with the same characteristics. An encoder might also be able to change the framerate or resolution. If this is done by modifying the capture parameters in the camera, then there may be no adequate reference to compare

to. For example, increasing the capture interval from 40 ms to 66.7 ms (reducing the framerate from 25 fps to 15 fps) clearly reduces quality (and there is subjective data to back this up [Mee01]), but there isn't a good way to measure this reduction with PSNR. For a fixed bitrate, reducing the framerate increases the number of bits available per frame, which may produce increased PSNR (measured against the reduced-framerate input)[1].

Similarly, for a fixed bitrate, a reduction in resolution increases the bits available per pixel, leading to increased PSNR against the uncompressed (but still reduced-resolution) original. If the high-resolution video was available, one could measure PSNR against that by resampling the reduced-resolution sequence, but resampling introduces its own degradation, and the resampling algorithm used at the receiver (if any) is not generally known. Which resolution gives perceptually better quality depends on the bitrate. A low-resolution video may be perceptually better than a high-resolution one once the bits per pixel of the high-resolution video drops low enough that the lossy compression algorithm starts to break down.

Reductions in framerate and resolution can help to reduce the effect of the knee in the quality/bitrate curve, but a particular codec may only have a limited ability to take advantage of them. For example, a "scalable" codec may only have a few available resolutions, with fairly large jumps between them. Switching between them will produce noticeably jarring changes in quality. Other codecs like VP8 can only change resolutions at a keyframe. Sending keyframes is quite expensive, and generally done rarely in low-latency applications because of the increased bitrate they require.

The situation with audio is similar. Objective metrics like PEAQ (Perceptual Evaluation of Audio Quality) and PESQ (Perceptual Evaluation of Speech Quality) play a similar role, though they are much more complex than their video counterparts (out of necessity: simple metrics such as the Signal-to-Noise Ratio (SNR) are mostly useless). Fortunately, the bitrate required to achieve a given audio quality varies much less with the content being encoded than with video (by a factor of two or three at most, rather than an order of magnitude). The one exception to this, however, is silence (or near silence), which requires only an occasional packet to be sent (called Discontinuous Transmission, or DTX). The audio is simply replaced by comfort noise. A congestion control algorithm must be prepared to deal with the sudden onset of packets from a stream which had been idle, or it will drop the start of a person's sentence.

Audio compression has the same knee in the quality/bitrate curve that video does, but like video, there are other parameters which can be varied to affect bitrate, such as sampling rate and channel count. Like video, the effect is not well-captured by objective metrics. Higher sampling rates (and thus higher audio bandwidth) generally give improved subjective quality, and as with video resolution, there is a bitrate below which switching to a lower sampling sounds better [RT11]. Stereo gives an increased sense of presence when available [DARV10],

[1] Certain coding tools such as motion compensation become less effective as the distance between frames grows larger, so an increase is not guaranteed.

though it requires approximately twice the bitrate of mono[2].

The Opus audio codec is one of the only codecs which can seamlessly switch between different audio bandwidths and between mono and stereo. Other codecs have a much more limited ability to use these kind of changes to adapt to the channel capacity. A stream could switch between the AMR family of codecs (AMR-NB, AMR-WB, etc.) by changing the RTP payload type, but this switch would not be seamless. As an extreme example, G.711 (still commonly used for interoperability) cannot even change its bitrate, requiring a fixed 64 kbps with an 8 kHz sampling rate. The knee in its curve is in fact a cliff.

For interactive applications, network conditions also play an important role in determining perceived quality. Latency and packet loss, are the key parameters which are influenced by the congestion control algorithm. The effect of packet loss is difficult to judge, as it depends on the quality of the concealment available at the receiver. In general it is much more damaging for audio than video due to the reduced temporal sensitivity of the human visual system [vdBL96]: it is sometimes possible to send an RTCP Slice Loss Indication (SLI) message back to the sender and receive an updated, undamaged frame before the viewer notices. Such techniques are impossible with audio: there is no way to go back and fill in a missing word.

The effect of all of this is that it becomes very difficult to allocate bandwidth in a perceptually optimal manner. An algorithm must handle sudden increases in bitrate (from quick motion in video or DTX in audio) without introducing excess buffering (which would significantly increase latency). Even in the steady state, the knee in the quality curve means that a reduction of the channel capacity by half does not necessarily mean each flow should have its rate cut in half. Since this knee may fall in different places for different streams, this could cause one of them to fall off a cliff, but not the other. A more fair allocation would let the first stream keep more bits, but common mechanisms such as priorities, traffic classes, and lower bound reservations fail to capture the dynamic nature of this problem. A real solution requires real-time knowledge from the encoder, not only of the location of the bitrate/quality knee but of the other parameters which can be adapted (such as resolution, framerate, sampling rate, etc.) and the limits a particular codec imposes on them. How much of this can be incorporated into a practical congestion control algorithm remains to be seen.

[2]Even when using stereo coupling. The objective quality of each channel must be higher than an equivalent mono stream because stereo unmasking effects make artifacts more perceptible.

# References

[DARV10] Christina Dicke, Viljakaisa Aaltonen, Anssi Rämö, and Miikka Vilermo. Talk to me: The influence of audio quality on the perception of social presence. In *Proc. 24$^{th}$ British Computer Society Interaction Specialist Group Conference (HCI 2010)*, pages 309–318, Dundee, Scotland, September 2010.

[Mee01] Michael J. Meehan. *Physiological Reaction as an Objective Measure of Presence in Virtual Environments.* PhD thesis, University of North Carolina at Chapel Hill, Department of Computer Science, 2001.

[RT11] Anssi Rämö and Henri Toukomaa. Voice quality characterization of IETF Opus codec. In *Proc. 12$^{th}$ Annual Conference of the International Speech Communication Association (Interspeech 2011)*, Florence, Italy, August 2011.

[vdBL96] Christian J. van den Branden Lambrecht. *Perceptual Models and Architectures for Video Coding Applications.* PhD thesis, École Polytechnique Fédérale de Lausanne, 1996.

[XZ06] Bo Xie and Wenjun Zeng. A sequence-based rate control framework for consistent quality real-time video. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(1):56–71, January 2006.