

Audio Frames — Cheat Sheet

Frame sizes, the samples-to-ms formula, the RTP packet, and where latency comes from.

The one rule

frame duration = samples per frame / sample rate. e.g. 1,024 / 48,000 = 21.3 ms (AAC).

A frame is the smallest chunk a codec can decode alone. It is a slice of time, measured in samples.

Frame size by codec (at 48 kHz)

Codec	Frame (samples)	Duration	Internal sub-unit
Opus	120-2,880 (you choose)	2.5 / 5 / 10 / 20 / 40 / 60 ms	none
AAC-LC	1,024	21.3 ms	8x128 short blocks on transients
MP3	1,152	24.0 ms	2 granules x 576
AC-3	1,536	32.0 ms	6 audio blocks x 256
G.711	1 (no transform)	often 20 ms packets	none

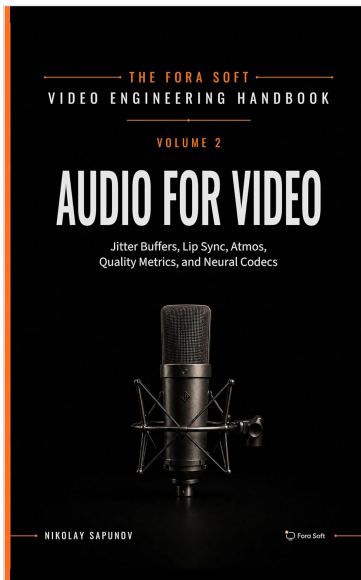
Mouth-to-ear latency stack (representative, 20 ms frame)

Stage	Time
[frame-size] capture (one 20 ms frame must fill)	20 ms
[frame-size] encoder frame + look-ahead	~25 ms
network transit (one-way, good path)	~30 ms
[frame-size] jitter buffer (holds ~2 frames)	~40 ms
decoder + output device buffering	~20 ms
total	~135 ms

tinted rows shrink if you pick a smaller frame.

Frame-and-packet checklist

- ☐ Frame size matches the latency budget (20 ms is the default sweet spot).
- ☐ RTP: one frame per packet; sequence number and timestamp present.
- ☐ Opus RTP timestamp uses the fixed 48 kHz clock (+960 per 20 ms frame).
- ☐ Stream segments rounded to a whole number of audio frames (no drift).
- ☐ Gapless playback: encoder delay / end-padding metadata set correctly.



GO DEEPER · VOLUME 5 OF 7

Audio for Video: A Complete Guide to Sound in Video Products

From Jitter Buffers and Lip Sync to Atmos, Quality Metrics, and Neural Codecs

By Nikolay Sapunov, CEO at Fora Soft · 1,288 pages · Kindle & paperback on Amazon

[Read it on Amazon →](#)
amazon.com/dp/B0HH6XPY46

This cheat sheet turns frames into milliseconds, lists frame sizes per codec and stacks up mouth-to-ear latency. Volume 5 follows those frames across the network: jitter buffers and NetEq, packet loss concealment, FEC and RED, timestamps and clock domains, and lip sync.

WHAT YOU WILL FIND INSIDE

- Jitter buffers and NetEq
- Lip sync between audio and video
- Timestamps and clock domains
- Loss concealment, FEC and RED

THE FULL HANDBOOK · 7 VOLUMES · 9,000+ PAGES · ALL ON AMAZON

Vol. 1 · Video Encoding — pixels and color to H.264, HEVC, AV1, quality metrics and AI

Vol. 2 · Video Streaming I — network transport, contribution and ingest, HLS, MPEG-DASH, CMAF

Vol. 3 · Video Streaming II — ABR, CDNs, players, WebRTC, DRM and the economics of delivery

Vol. 4 · Audio for Video I — PCM, codecs, loudness, accessibility, real-time voice processing

Vol. 5 · Audio for Video II — jitter buffers, lip sync, Atmos, quality metrics, neural codecs

Vol. 6 · AI for Video I — models and inference, computer vision, speech, search, generative video

Vol. 7 · AI for Video II — real-time pipelines, agents, serving, cost, compliance, verticals

Chasing audio latency or drift in your calls or streams?

We will measure your mouth-to-ear budget stage by stage and show which frame, buffer and clock settings to change. 250+ video products shipped since 2005.

[Book a 30-min call →](#)
[WhatsApp →](#)
[Email us →](#)