

VAD & DTX — Tuning Cheat Sheet

Voice activity detection, discontinuous transmission, comfort noise, Opus settings, and the two failure modes.

VAD: four generations

Energy threshold	Loud = speech	Fails the moment the room is noisy
Classical (G.729-B)	4 features + adaptive noise track	ITU-T telecom reference, 1996
WebRTC VAD	6 bands + Gaussian mixture model	Aggressiveness 0-3; 10/20/30 ms frames
Silero VAD (neural)	~309K params, ~1-2 MB, v5 2024	32 ms chunks; multilingual; best accuracy

The two failure modes (+ the fix)

False negative	Real speech called silence	Clips words — users hate this
False positive	Noise called speech	Wastes bandwidth (cloud bill only)
Hangover (the fix)	Transmit a short tail after speech	Buys back most clipped-word tickets

Opus DTX — the settings that matter

- **usedtx default = 0 (off)** — Must set usedtx=1 explicitly
- **Silence: ~1 frame / 400 ms** — vs 1 frame / 20 ms speaking
- **Drop WHOLE frames only** — RTP timestamps differ by mult. of 120
- **DTX vs loss** — Receiver reads timestamp gap vs seq num
- **Do NOT add RFC 3389 CN** — Opus makes its own; pairing discouraged
- **Voice mode, not music** — DTX engages in the SILK/speech path

The bandwidth math

- Speaking: ~60 byte Opus payload every 20 ms = $60 \times 50 = 3,000$ bytes/sec.
- Idle (DTX): ~3 byte descriptor every 400 ms = $3 \times 2.5 = 7.5$ bytes/sec.
- Payload drop on the idle stream ~ 99% (smaller once RTP/UDP/IP headers are added back).
- Whole-call saving is 'tens of percent' because it depends on talk ratio and header overhead.

DTX on or off?

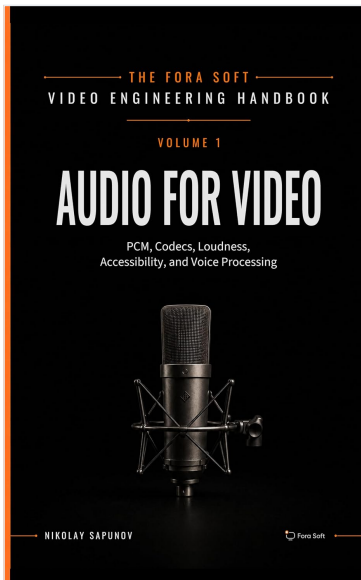
- DTX ON: conversation — meetings, telemedicine, contact centre, classrooms (most mics idle).
- DTX OFF: continuous sound — music, DJ sets, ambient installs, baby monitors (silence IS the signal).
- Gentle VAD + generous hangover: clinical / quiet-voice calls where clipped words are unacceptable.
- Aggressive DTX on audience mics: large webinars, so the media server is not flooded with idle packets.

Remember

- ☐ VAD decides speech-vs-silence per frame; DTX acts on it by stopping full transmission during silence.
- ☐ During silence send a tiny SID descriptor (level + spectral bytes), not nothing — the receiver plays comfort noise.
- ☐ Comfort noise carries NO audio, only a recipe: a level byte in dBov plus optional spectral coefficients (RFC 3389).
- ☐ The same VAD also drives transcription endpointing, recording dead-air removal, and active-speaker indicators.
- ☐ Tune DTX with noise suppression and gain control together — they are not independent in the capture chain.

Companion to: 'Voice Activity Detection (VAD) and Discontinuous Transmission (DTX)'

www.forasoft.com/learn/audio-for-video/articles-audio/vad-dtx-voice-activity-discontinuous-transmission



GO DEEPER · VOLUME 4 OF 7

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

From PCM and Codecs to Loudness, Accessibility, and Real-Time Voice Processing

By Nikolay Sapunov, CEO at Fora Soft · 1,229 pages · Kindle & paperback on Amazon

[Read it on Amazon →](#)
amazon.com/dp/B0HH6PRZH1

This sheet compares four VAD generations, the two failure modes and the Opus DTX settings that matter. Volume 4 covers the capture chain in depth: voice activity detection, noise suppression, echo cancellation and gain control working together, plus Opus and speech codecs and the loudness targets your calls should hit.

WHAT YOU WILL FIND INSIDE

- Voice activity detection and hangover
- Noise suppression, AEC and AGC together
- Opus and speech codec settings
- Loudness and LUFS for voice

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

Are your calls clipping words or wasting bandwidth on silence?

We will tune VAD, hangover and Opus DTX together with noise suppression on your real recordings, and measure the bandwidth you save. 250+ video products shipped since 2005.

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