VAD & DTX - Tuning Cheat Sheet

Fora Soft Learn

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 x 50 = 3,000 bytes/sec.

Idle (DTX): ~3 byte descriptor every 400 ms = 3 x 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.