Recording & Transcription Pipeline - Field Guide

Fora Soft Learn

The three recording branches, the ASR audio path, and what to ask before audio leaves the call.

Three recording branches at a glance

CRITERION	CLIENT	PER-PARTICIPANT	MIXED
Where capture happens	User device	Server (SFU)	Server (MCU-style)
Server CPU cost	None	Low (decode)	High (decode+mix)
Storage shape	1 per device	1 per speaker	1 combined file
Knows who said what	This user only	Yes, built in	No, must guess
Survives tab close	No	Yes	Yes
Plays in any player	Yes	Not directly	Yes
Best for transcription	Weak	Strongest	Weakest
Re-mix / redact later	No	Yes	No

The ASR audio path (sound -> features, one way)

Resample	Opus runs at 48 kHz; ASR drops to 16 kHz - speech lives below 8 kHz.
Frame	Cut into 25 ms windows that slide 10 ms at a time (they overlap).
Log-mel	Each frame -> 80 mel bands, loudness log-compressed. Whisper uses 80 bins.
One-way	30 s -> an 80x3000 grid (240,000 numbers). Cannot be reversed to audio.

Tools you'll meet

MediaRecorder	Client-side recorder (W3C). timeslice -> dataavailable chunks; WebM/Opus.
LiveKit Egress	Server recording: hidden participant; room-composite = headless Chrome+GStreamer.
Whisper	ASR input is log-mel, not your recording. 16 kHz, 80 bands, 30 s chunks.
Diarization	pyannote/NeMo -> RTTM (start, duration, speaker); WhisperX merges word timings.
Live captions	Partials in 200-500 ms; finals after the pause. Speed vs steadiness trade-off.

Before audio leaves your call

- [ ] A live call and its recording are different jobs - tune for one and the other suffers.
- [ ] Mixing is a one-way door: you lose 'who said what' and the ability to redact one speaker.
- [ ] Per-participant recording is the strongest input for transcription and cleanest for compliance.
- [ ] DTX silence + a naive mixer = recordings that cut in and out. Hold level / add comfort noise.
- [ ] ASR throws the sound away: 16 kHz, 80 log-mel bands, one-way - a transcript is not a recording.
- [ ] Per-participant means no diarization problem - each file is already one speaker.
- [ ] Feed the captioner a clean decoded track, not DTX-thinned live packets, when accuracy matters.
- [ ] Stored voice is regulated data: consent (1-party vs all-party), GDPR delete, HIPAA encrypt + 6 yr.