

Pyannote Production Setup Checklist

Speaker diarization — who spoke when — on your own infrastructure

1 · The three-stage pipeline

#	Stage	Model	What it does
1	Segmentation	segmentation-3.0	Find speech + overlap in a 10 s sliding window (powerset)
2	Embedding	ResNet / WeSpeaker	Turn each solo voice into a numeric fingerprint
3	Clustering	agglomerative	Group matching fingerprints; speaker count discovered
-	Aggregation	—	Assemble a who-spoke-when timeline (RTTM)

2 · Self-host pyannote vs a hosted service

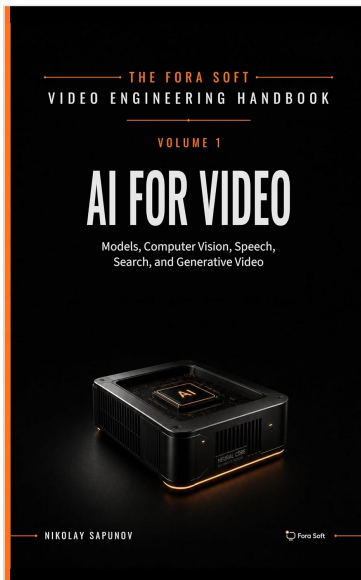
Criterion	Self-hosted pyannote	Hosted service
Software cost	Free (MIT license)	Per-hour / per-minute fee
Accuracy (DER)	Strong; tunable on your data	Often higher out of the box
Speed	~40× real time on one GPU	Vendor-managed, often faster
Data location	Stays on your servers	Sent to vendor cloud
Setup	GPU, models, gated token, tuning	An API key
Best when	On-prem; steady volume; can tune	Ship fast; top accuracy, no ops

3 · Pre-flight checklist before you ship

- ☐ Confirmed you actually need **speaker labels** — if you only need the words, use ASR, not diarization.
- ☐ Created a **Hugging Face token** and accepted user conditions for **both** segmentation-3.0 and speaker-diarization-3.1.
- ☐ Extracted clean **mono audio at 16 kHz** — the input the pipeline expects.
- ☐ Passed the **speaker count** (num_speakers) or a range (min/max_speakers) wherever the product knows it.
- ☐ Checked **data residency**: can the audio legally leave your servers, or must it stay on-prem?
- ☐ Reported DER honestly: stated the **collar** and whether **overlapped speech** was scored; compared like for like.
- ☐ Set expectations that **overlap and short backchannels** are the hardest cases and may be mislabelled.
- ☐ Built a **human-correction path** for records where a wrong speaker label has real cost (medical, legal).
- ☐ Considered **tuning** (clustering threshold δ) or **fine-tuning** the segmentation model on domain audio.
- ☐ Modelled cost at **real batch volume** (GPU-hours = audio-hours / realtime-factor) vs a hosted service.

Companion to: 'Speaker Diarization With Pyannote In Production'

www.forasoft.com/learn/ai-for-video-engineering/articles-ai/pyannote-speaker-diarization-production



GO DEEPER · VOLUME 6 OF 7

AI for Video: A Complete Guide to Intelligence in Video Products

From Models and Inference to Computer Vision, Speech, Search, and Generative Video

By Nikolay Sapunov, CEO at Fora Soft · 1,444 pages · Kindle & paperback on Amazon

[Read it on Amazon →](#)
amazon.com/dp/B0HH6TKPZ1

This checklist covers the pyannote pipeline, the self-host vs hosted choice and ten checks before shipping diarization. Volume 6 goes deeper into speech in video products: ASR, TTS and diarization, model artifacts and deployment topologies, and search over a transcribed video archive.

WHAT YOU WILL FIND INSIDE

- ASR and speaker diarization in production
- ML basics for build vs buy decisions
- Model artifacts and deployment topologies
- Search and RAG over transcribed video

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

Need reliable who-spoke-when labels on your own servers?

We will deploy diarization next to your ASR, tune it on your audio and report DER honestly, with a cost per audio hour. 250+ video products shipped since 2005.

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