Approaches to Diarization

Real-time diarization is important for speech-to-text applications. Several high-performance supervised and unsupervised techniques have been devised since 2016 (since transformers seemingly).

Open-Source

1. PyAnnote (link)

Pyannote is an open-source framework for speaker diarization. It is the most popular open-source speaker diarization framework to exist.
It uses a Time-Delayed Neural Network (TDNN) of about 4.3 million parameters to detect the speaker. The neural network is small enough to be deployed on edge devices.

Aperdannier et al. (2024) performed a small-scale study on real-time diarization performance across different diarization pipelines where the DIART framework with the embedding model pyannote/embedding and the segmentation model pyannote/segmentation proved to be the best system.

> Based on PyTorch machine learning framework, it comes with state-of-the-art pretrained models and pipelines, that can be further fine-tuned to your own data for even better performance.

The models are available on huggingface as well.

2. NeMo (link)

> NVIDIA NeMo™ is an end-to-end platform for developing custom generative AI—including large language models (LLMs), vision language models (VLMs), video models, and speech AI—anywhere.

NeMo offers several tools (pipeline, trainer, pre-trained models, etc.) to build AI applications. Reference

	pyannote	Nemo
Pre-trained models available	✅	✅
Good overlapping speakers detection (multilabel segmentation)	✅	➖
Easy integration with ASR task and downstream NLP tasks	➖	✅
Possibility to specify the number of speaker as a parameter for inference	✅	✅
Automatic detection of the number of speakers	✅	✅
Models available for specific use cases (phone call, outdoor conversation, high quality,...)	❌	✅
Highly customizable pipeline	➖	✅

Proprietary

1. Google Cloud Speech-to-Text (link)

Google Cloud Speech-to-Text has two "versions": v1 and v2. Both have an option to enable diarization for speech-to-text.
The prices remain similar to Azure.
API Link

Real-time diarization is newly introduced and seems to have some issues with working effectively. More about GCP Diarization

Speech-to-Text V1 API	V1 offers data residency for multi region only. Models include short, long, phone call, and video. V1 does not include audit logging. New customers get $300 in free credits and 60 minutes for transcribing and analyzing audio free per month, not charged against your credits.	$0.024 per min
Speech-to-Text V2 API	V2 offers data residency for multi and single region. Models include short, long, telephony, video, and Chirp. V2 does include audit logging and support for customer managed encryption keys.	$0.016 per min

Pricing

	FREE	TIER 1
Cloud Speech-to-Text Dynamic Batch Recognition (Logged) Cloud Speech-to-Text Recognition (Logged) Text-to-Speech On-Device Serving Library Subscription Text-to-Speech On-Device Voice Subscription	INR 0.00 /minute Starting after: 0 minute/month	INR 2.052989999 /minute Starting after: 60 minute/month

2. Microsoft Azure: diarization-ai (link)

Azure offers a relatively simple interface for real-time speaker diarization. It takes about **1$ per hour**, after the **free-tier which is 5 audio-hours** per month.

Category		Price
Speech to Text (per second billing)	Standard	Real-time Transcription: $1 per hour Fast Transcription: N/A per hour[9] Batch Transcription: $0.18 per hour[1]
	Custom	Real-time Transcription: $1.20 per hour Batch Transcription: $0.225 per hour[1] Endpoint hosting: $0.0538 per model per hour Custom Speech Training[5]: $10 per compute hour
	Enhanced add-on features: • Continuous Language identification • Diarization • Pronunciation Assessment (prosody, grammar, vocabulary, topic)	Real-time: $0.30 per hour per feature Batch (Continuous Language identification, Diarization): Included in Standard/Custom (no extra charge)

Quickstart

3. AWS (link)

The service offered is very similar to Azure, it is very straightforward and easy to use.

Region:

US East (Ohio)

Tier	Volume (minutes/month)	Standard Streaming Transcriptions ($/minute)*
T1	First 250,000 minutes	$0.02400
T2	Next 750,000 minutes	$0.01500
T3	Next 4,000,000 minutes	$0.01020
T4	Over 5,000,000 minutes	$0.00780

4. AssemblyAI

AssemblyAI is a startup that offers Speech-to-Text options, it's API includes an option for diarization.