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 <u>PyTorch</u> machine learning framework, it comes with state-of-the-art <u>pretrained models and pipelines</u>, 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 Al—including large language models (LLMs), vision language models (VLMs), video models, and <u>speech Al</u>—anywhere.

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

	pyannote	Nemo
Pre-trained models available	▽	V
Good overlapping speakers detection (multilabel segmentation)	~	
Easy integration with ASR task and downstream NLP tasks		V
Possibility to specify the number of speaker as a parameter for inference	~	✓
Automatic detection of the number of speakers	▽	V
Models available for specific use cases (phone call, outdoor conversation, high quality,)	×	✓
Highly customizable pipeline		<u>~</u>

Proprietary

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

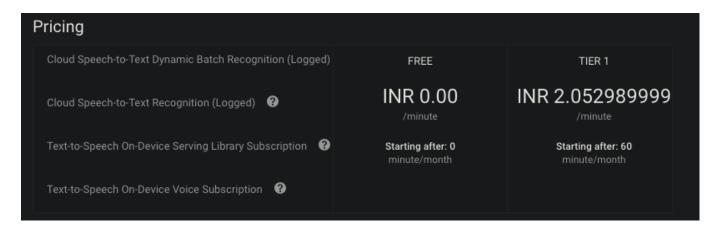
Google Cloud Speech-to-Text has two "versions": v1 and v2. Both have an option to enable diarization for speech-to-text.

The prices remain similar to Azure.

API Link

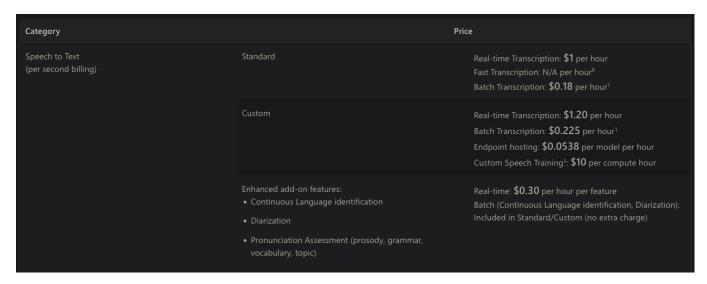
Real-time diarization is newly introduced and seems to have some issues with working effectively. <u>More about GCP Diarization</u>

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



2. Microsoft Azure: diarization-ai (link)

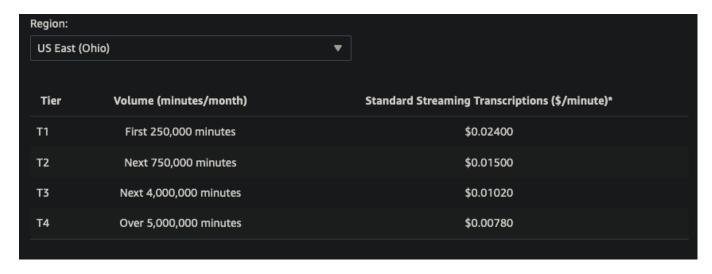
Azure offers a relatively simple interface for real-time speaker diarization. It takes about **1\$ per hour**, after the **free-tier which is 5 audio-hours** per month.



Quickstart

3. AWS (link)

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



4. AssemblyAl

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