Deepgram

WHITE PAPER

How Deepgram Works

Table of contents

Executive Summary

As more businesses differentiate on customer experiences, better voice experiences and deeper insights from your voice channels represent a strategic need. In the 2023 State of Voice Technology report, 99% of respondents said that voice-enabled experiences were a critical part of their future plans. Regardless of whether you are evaluating automatic speech recognition (ASR) solutions to get more value out of your call center data, build the next game-changing voice feature, or are just looking to save money on speech transcription, Deepgram is the platform to get you there.

Deepgram is a developer-centric speech-to-text platform built with the enterprise in mind. Unlike alternative speech recognition solutions, Deepgram uses a 100% deep learning solution that is faster, more accurate, more reliable, more flexible, and more scalable than any ASR on the market—and it runs on premises or in the cloud.

What Makes Deepgram Different

- **Frictionless Developer Experience:** Use our Python, Node.js, or .NET SDKs—or our REST API—to get up and running, typically in less than 5 minutes.
- **GPU-powered:** Develop an end-to-end, GPU-powered deep learning infrastructure for processing voice data. From speech training to inference, our managed service operates with superior efficiency, scalability, and reliability for high throughput workloads.
- **Models for Specific Use Cases:** Deepgram serves hundreds of models simultaneously, rather than just the one or two permitted in traditional speech pipelines, including specific use case models for important modes of speech such as phone calls, meetings, spoken media, and more.
- **Tailored Models:** If one of our models doesn't work for you, you can tailor a model based on your provided training data. Custom speech models can be trained in weeks, not months–and boost accuracy levels even further.
- **Rapid Feature Development:** We are committed to continuous improvement and release new features and models on a regular basis to constantly improve the value of transcripts and language understanding capabilities we provide.

Value for Enterprises

1. **Maximum Accuracy:** Out-of-the-box accuracy up to 90% on typical business audio (e.g., phone calls, meeting transcriptions, etc.). See how we stack up.
2. **Time to Value:** Transcribe hour-long recordings in 30 seconds or less. Deploy high quality speech models in weeks, not months or years.
3. **Superior ROI:** Enterprise pricing options available for high throughput customers deliver the most value with the least expense.
4. **Scalability and Resilient Operations:** Process hundreds of audio streams at once. Built-in reliability and scalability thanks to our enterprise-grade infrastructure.
5. **Flexible Deployment:** Train models and deploy anywhere—on premises or in the cloud in days.
6. **Security and Privacy:** SOC2 compliant and on-prem support provides enterprise-grade security and data privacy.

The Value of Enterprise Audio

Any organization that hosts meetings or uses the phone, video conferencing, interactive voice response (IVR), voicebots, or any other audio-based solution to communicate is sitting on a wealth of untapped insights.

Here are some example business applications that benefit from ASR and their high level requirements across accuracy, speed, and customization.

Applications	Examples	Accuracy Required	Speed Required	Customization Required
Accessibility	• Expand your customer base with captioning • Meet regulatory requirements • Increase productivity in blended learning environments • Take notes automatically for the deaf and hard of hearing	High	High	Med/High
Analytics	• Analyze customer experiences • Find new product and service ideas • Determine the appeal of marketing campaigns • Increase telemarketing productivity • Detection topics including products, companies, and competitors	High	Med	High

Applications	Examples	Accuracy Required	Speed Required	Customization Required
Coaching	• Live coaching and training • Perform post-call reviews	High	Med/High	High
Compliance	• Ensure all agents have provided correct compliance language • Track compliance with human resources guidelines • Provide alerts for out-of-compliance language	High	Med	Med/High
Content Moderation	• Find and eliminate hate speech or radicalization • Prevent bullying, harassment, and child grooming • Eliminate audio scams and spam • Remove sexual audio content	High	Med	High
Enablement	• Provide solutions during real-time conversations • Alert telemarketing reps on buying signals • Recommend upsell opportunities from conversations • Determine sentiment and likelihood of customer churn	High	High	Med/High
Monitoring	• Check mental wellness of patients • Monitor environment for escalating danger situations • Alert leaders to tense situations via police body cam audio • Determine emotions or sentiment of patients and customers	High	High	High
Transcription	• Summarize meetings and action items automatically • Provide full transcription for review • Chart doctor-patient interactions • Analyze recruiting interviews	High	Med	Med/High
Voicebots IVR	• Understand customer voice requests clearly • Enable more conversational interactions • Improve customer experience	High	High	High

Deepgram Product Overview

Speech recognition is hard. We make it easy.

Deepgram uses cutting-edge, proprietary methods for model creation, training, data labeling, and deployment. These innovative approaches—based on the latest advances in deep learning—generate transcription and understanding with higher accuracy and flexibility than our competitors. For difficult audio, accents, industry terms, noise, or crosstalk, we can train a customized speech model for your application and specific audio characteristics within weeks for even higher accuracy.

Developer Focused

Deepgram is wholly focused on speech and provides rich developer resources and a high-touch customer experience. We're serious about being ***the*** speech company, and we want our customers to succeed with voice.

View full documentation

Within the Deepgram Console, you can upload your audio to get immediate transcripts. You can also upload audio via our easy-to-use REST API, or use one of our SDKs. For companies that do not have training-ready datasets, contact us to use our in-house transcription team to convert your audio into labeled data.

Enterprise Speech-to-Text API Features

TRANSCRIPTION Get accurate, readable transcriptions back in seconds, not minutes. See how we stack up	**REAL-TIME STREAMING** Keep the conversation flowing by transcribing phone and meeting conversations as they happen, with <300 millisecond latency.	**BATCH TRANSCRIPTION** Transcribe a backlog of audio files at up to 120 times normal audio speed (i.e., transcribe one hour of audio in less than 30 seconds).
MULTI-LANGUAGE SUPPORT Transcribe audio across more than 20 languages and dialects.	**DIARIZATION** Identify up to 10 different speakers at one time, charged by the second not by the number of speakers.	**DEEP SEARCH BY PHONETICS** Accurately identify top terms or phrases in your audio with acoustic pattern matching instead of text search.

REDACTION Automatically redact sensitive data such as private health information or credit card information from transcripts.	**PUNCTUATION AND CAPITALIZATION** Use punctuation or capitalization in your transcripts to make them easier for humans to read.	**KEYWORD BOOSTING** Boost industry terms, unique product names, and company names to increase transcription confidence.
PROFANITY FILTERING Filter any profanity from transcripts.	**MULTI-CHANNEL SUPPORT** Reliably identify speaker changes across single and multi-channel audio.	**MULTI-AUDIO TYPES** Support over 40 different audio formats including WAV, MP3, FLAC, and AAC. No need to create different jobs for different file extensions.
AUDIO TIMESTAMPS All words include an associated timestamp. Drill into audio snippets with specific start and end times.	**CONFIDENCE %** Every word—and the entire transcript—is rated on the model's confidence that it's correct.	**CUSTOMIZABLE** Each model is tuned to the audio you care about. This is done through state-of-the-art data labeling and model training.
UTTERANCE DETECTION Segment speech into meaningful semantic units to interact more naturally and effectively with spontaneous speech patterns.	**NAMED ENTITY RECOGNITION (NER)** Turn strings of letters and words into important acronyms and numbers (e.g., "1aq347" from "one a q three four seven").	**VOICE ACTIVITY DETECTION** Monitor incoming audio to detect when a sufficiently long pause is present to stop transcription.

View full feature list

Multiple Speech Models

Deepgram doesn't believe in a one-size-fits-all speech model. Because everyone speaks differently, with different accents, slang, and jargon, how can one speech model be accurate for all of these differences? **It can't.**

Deepgram has a library of speech models to unlock your audio data. Choose from various Base speech models and languages to get started. Choose our Enhanced model for long tail words not normally spoken in general conversation or for higher accuracy in certain applications, like compliance or voicebots. Go further with our Tailored models to identify industry topics, difficult, noisy, crosstalk audio and audio unique to your business.

Base Model	Enhanced Model	Tailored Model
For companies looking for a higher level of accuracy over an existing vendor or homegrown solutions. You can access our various use case-specific speech models as well.	For companies that have long tail words, words not commonly used in general audio, or need very high accuracy, our Enhanced models offer the highest accuracy out of the box.	For companies that have difficult audio and specific vocabulary needs. Starting with a base or enhanced model, Deepgram uses labeled audio training data to produce a Tailored model suited to your unique needs.

The Benefits of Multiple Models

- With our lower processing costs, you can use speech recognition at scale—processing all of your audio through different models instead of just some of it.
- Deepgram is the only speech company that offers flexible Tailored models that can quickly be trained to recognize your unique branded terms, industry jargon, accents, and other features of your voice data.
- Deepgram's emphasis on processing speed means that you can get real-time transcripts in milliseconds for all our models, and unlike our major competitors, our success team will be with you at every step, accelerating your path to deployment and helping you get the most out of your voice data.

Machine learning is also capable of transfer learning. Transfer learning is a method for accelerating the creation of a new model by starting it with the weights of a model trained to do something similar. This is particularly well suited to speech models, allowing us to continuously improve all our base and enhanced models and languages to be at the forefront of speech recognition.

Continuous Improvement

Unlike traditional ASR systems, which are trained by meticulously editing sub-components of a data pipeline, our deep-learning neural network improves with each data set it receives. With Deepgram, you can continuously train your model with the voice of your customers, and it will improve identification of sounds, and subsequently the words in new audio submitted.

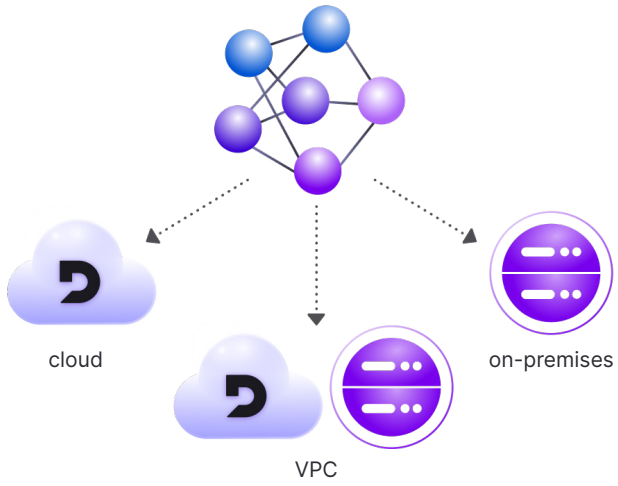
Your accuracies will continually improve the more data is processed and the more training is done on that real-world data. Our Base speech models, which are highly accurate, are not the end but the beginning of your accuracy improvements. Take for example our work with the National Aeronautics and Space Administration (NASA). They could not find nor were able to build an ASR solution that could reach 80% accuracy with the International Space Station (ISS) to Mission Control communications audio. If you've ever listened to this audio, it has static, cross-talk, various accents, and around 7500 terms and acronyms, many of which are unique to NASA.

Deepgram took our Base model, trained it for a few weeks with NASA audio, and were able to achieve over 80% accuracy on their ISS audio. With our Enhanced model and training we are up to over 85%+ accuracy. Naturally, we won that contract. Our technology along with our model training team allows us to rapidly develop better and more accurate speech models for any application—even in space.

You can view an example of how we perform on NASA audio on our ASR comparison tool.

Deployment Flexibility

Deepgram is an enterprise-grade, cloud-developer service. Our programmable API allows developers without deep data science expertise to run speech recognition models at scale. Deepgram is Kubernetes-ready with Docker images, and has pre-built VM images to enable rapid deployment to most cloud providers.



Deepgram has been built from the bottom up with performance as its highest priority. Existing frameworks and toolkits (e.g., PyTorch, TensorFlow) are not designed for this level of transcription speed, and are too slow to perform at enterprise requirements. Deepgram's patented infrastructure outperforms other deep learning frameworks by over 30%.

Key Benefits

- Remove any external latency from your process with on-premise deployments—for example, voicebots or real-time agent enablement.
- Easily deploy the Deepgram speech model on our servers or yours. We provide a step-by-

step guide for on-premise deployments with recommendations on the most cost effective GPUs to use.

- Train speech models on-premise to protect and secure your customer's audio. We have created proprietary on-premise training tools for our speech model training. No audio or transcripts will leave your environment.

The Future of Speech Recognition

As the landscape of speech recognition continually evolves, Deepgram remains at the forefront of these advancements. Our enthusiasm is fueled by the potential these technologies hold to solve real-world challenges. Our commitment is to seamlessly integrate speech recognition into your communication applications, unlocking more value and better experiences than ever before.

We're on a mission to make machines understand human speech as well as humans do and move us toward a world in which every voice is heard and understood. We hope you'll join us on this journey!