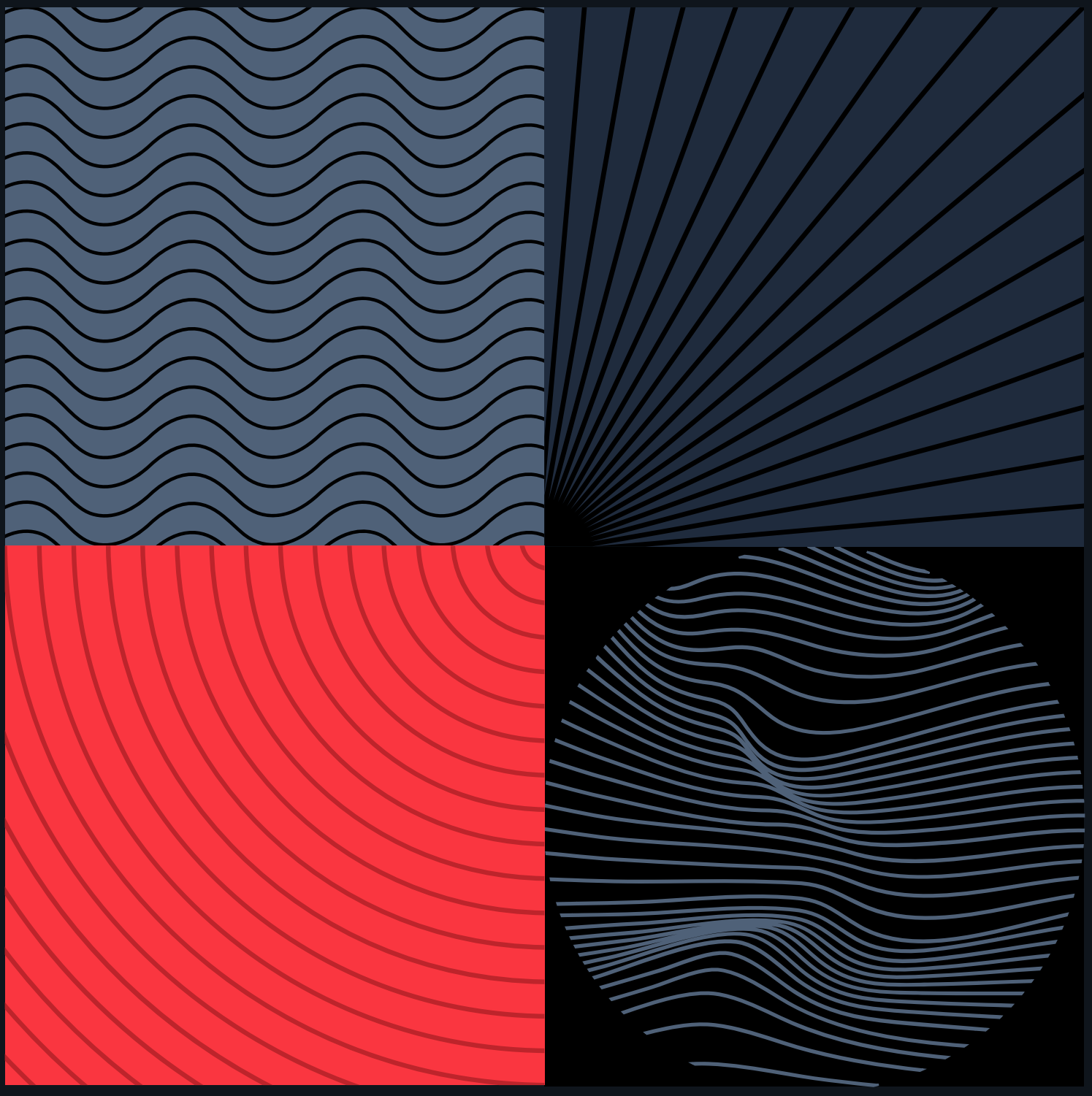
The Definitive Guide to Speech Recognition

DEEPGRAM



The Definitive Guide to Speech Recognition

DEEPGRAM

CONTENTS

What is Speech Recognition?

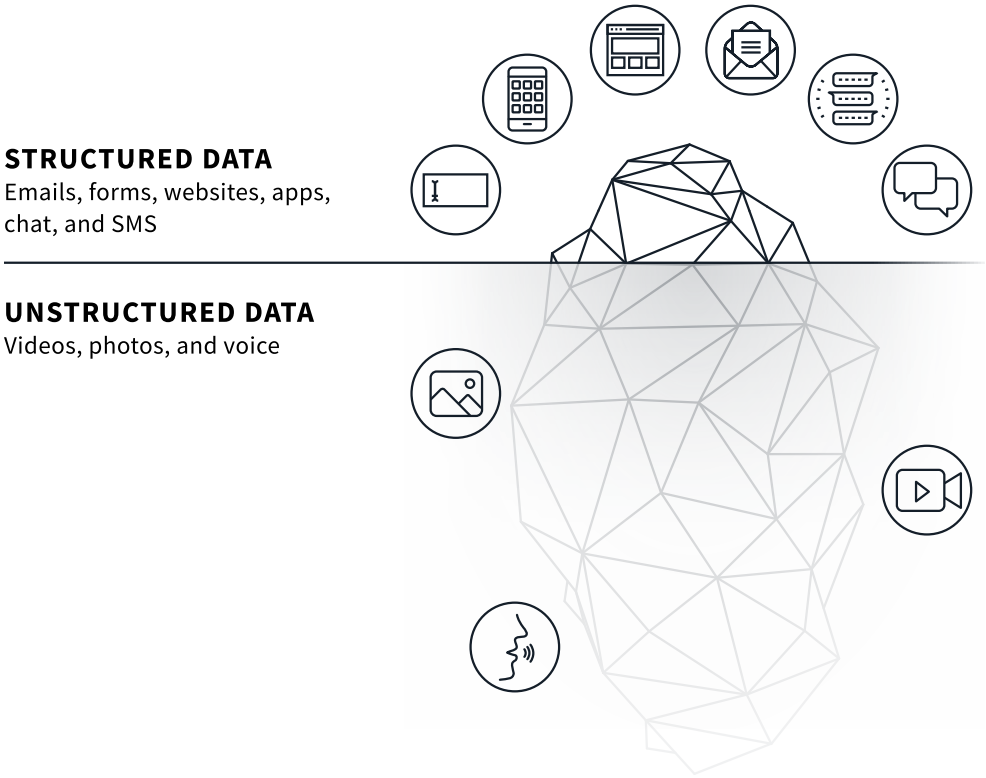
Speech recognition converts audio data into data formats that data scientists use to get actionable insights for business, industry, and academia. It is a method to change unstructured data (data not organized in a pre-defined manner) into structured data (organized, machine-readable, and searchable). Other names of speech recognition are speech-to-text (STT), computer speech recognition, or automatic speech recognition (ASR).

Some have also called it voice recognition but that term is defined differently. Voice recognition is defined as identifying a specific person from their voice patterns. Voice recognition is a feature of speech recognition. You can use speech recognition solutions in combination with artificial intelligence to identify a specific speaker and tie that voice pattern to a name.

Why is Speech Recognition Important?

When you look at all the data being generated in the world, only 10% of that data is structured data. That means 90% of the world's data is unstructured; unsearchable and unorganized, not yet being used for business insights. In addition, unstructured data is forecasted to increase by 60% per year. When you think about it, many organizations are making important decisions on only 10% of the data.

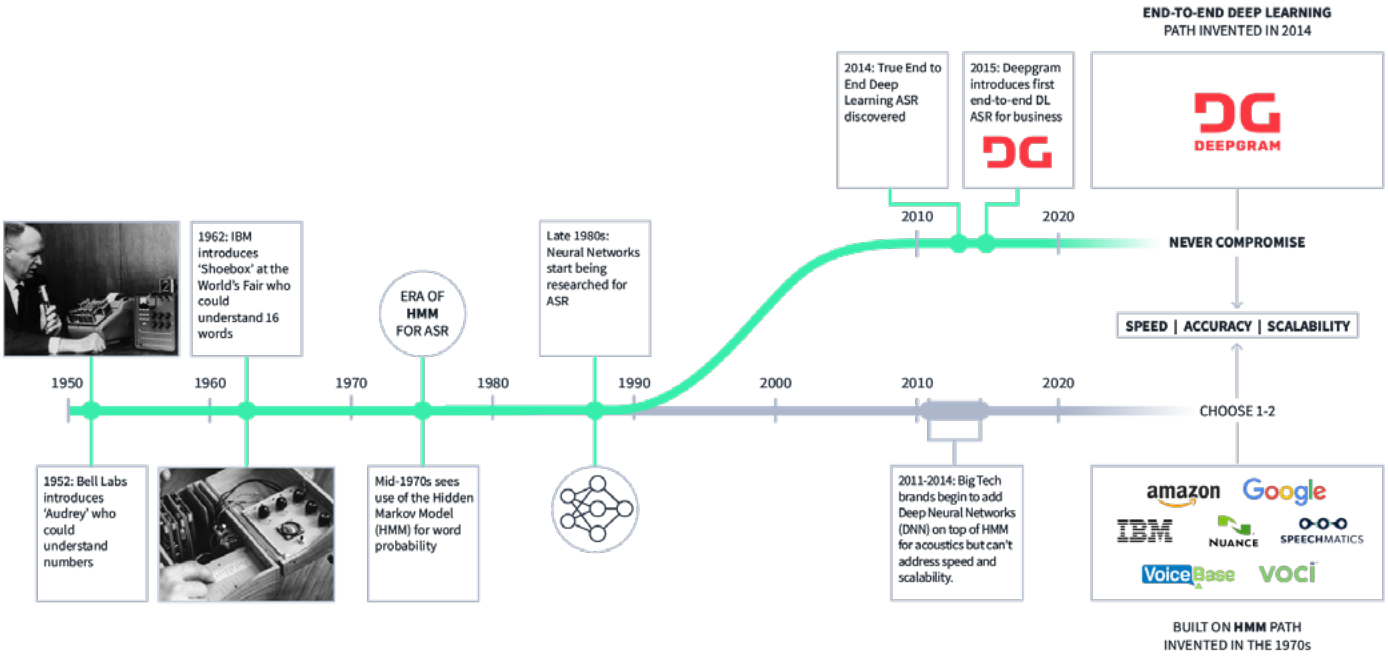
Most of this unstructured data is voice or video data that needs to be changed into machine-readable data to be used for decision-making. This is where ASR comes in and why it is important.



The History of Speech Recognition

Speech Recognition technologies began development in the 1950 and 1960s, when researchers made hard-wired (vacuum tubes, resistors, transistors and solder) systems that could recognize individual words, not sentences or phrases. That technology, as you might imagine, is essentially obsolete. The first known ASR was developed by Bell Labs and used a program called Audrey, which could transcribe simple numbers. The next breakthrough did not occur until mid-1970 when researchers started using **Hidden Markov Models** (HMM). HMM uses probability functions to determine the correct words to transcribe.

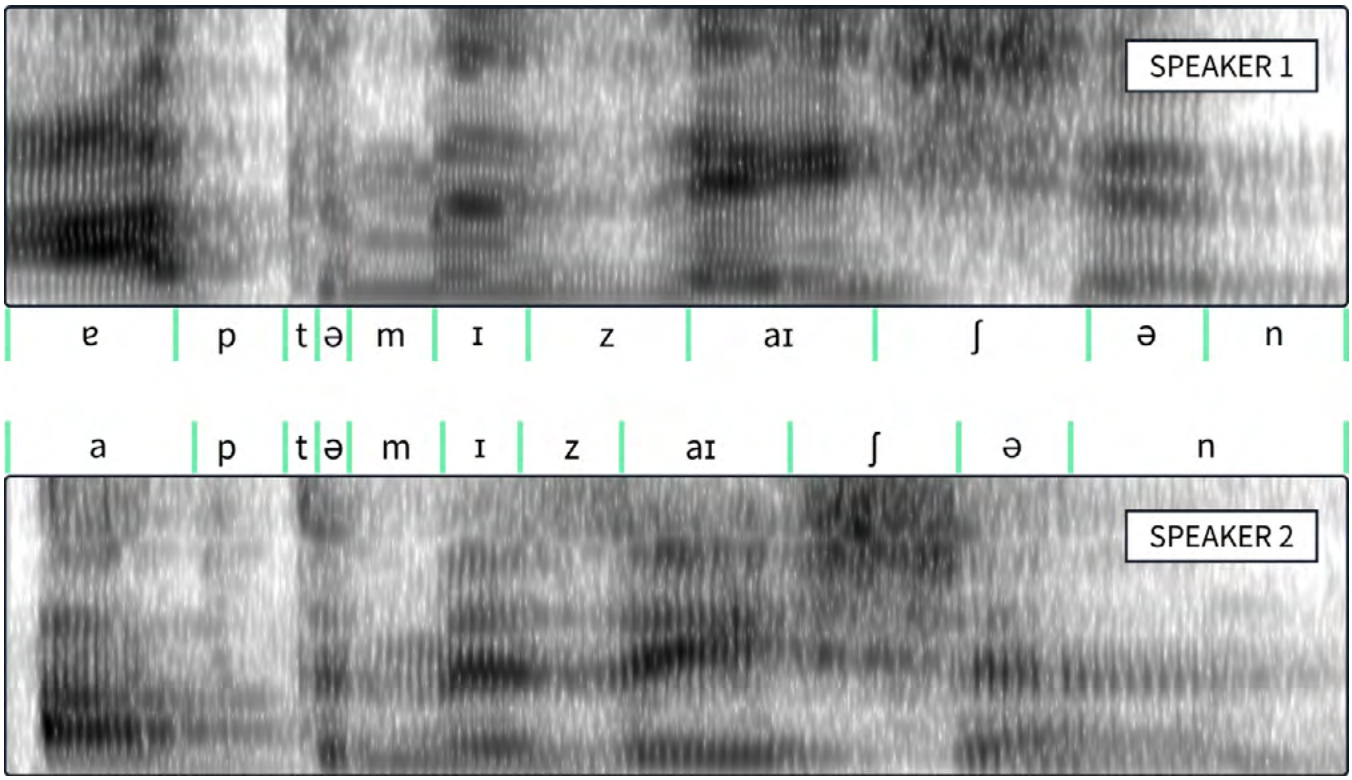
History of ASR →



Why is Speech Recognition Hard?

Between any two speakers there are variations in pronunciation, tone, word-choice, grammar choice, even amount of lung pressure, that from a mathematical perspective (computers speak math) make what they say completely different; even if it sounds the same to you and me. In fact, even if the same person utters a sentence twice, the sounds when recorded and measured are mathematically different. These are two spectrograms of two people saying the same word: optimization.

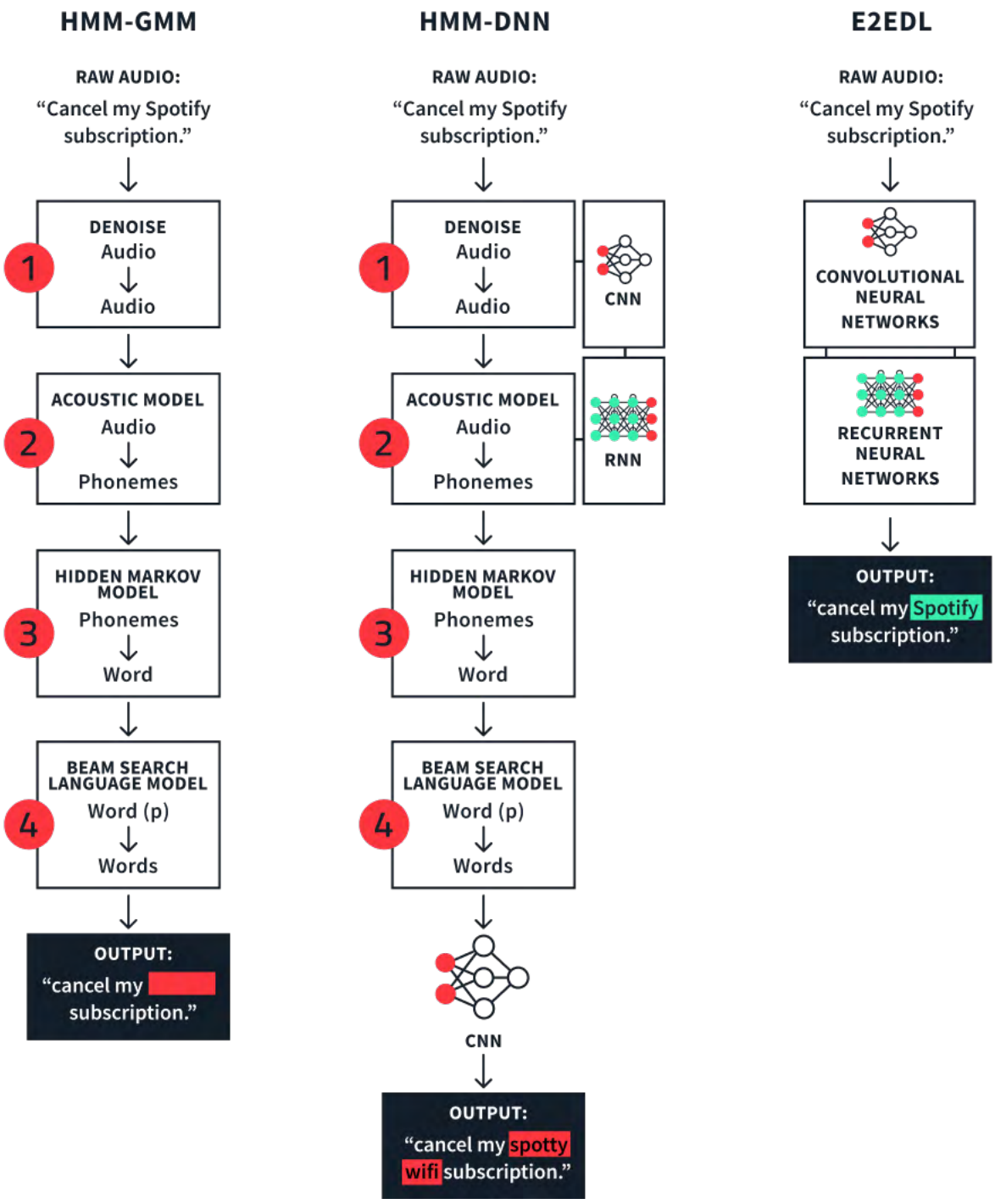
Spectrograms are one way to visualize audio data. As you can see, these two spectrograms are very different from one another. Pay special attention to the darker lines and their relative shapes. Same word to our human brains, but two mathematical realities for computer brains.



Speech Recognition Architecture

There are currently three main speech recognition architectures in existence today:

- **HMM-Guassian Mixed Model**, also called the Tri-gram model (HMM-GMM)
- **HMM-Deep Neural Network**, also called the Hybrid Model (HMM-DMM)
- **End to End Deep Learning** Speech Recognition (E2EDL)



HMM-GMM

The HMM-GMM architecture is a four to seven step process based on probability statistics. It is similar to an assembly line, where each step is independent, requiring the previous step's input to function. The four steps are:

1. **Denoise** - This step denoises the audio to make sure it is free of dogs barking or garbage trucks in the background. De-noising audio can be tricky because you don't want to accidentally remove the voice from the audio. There are a variety of approaches here, but primarily speech recognition software is trying to limit the audio to just what falls into the range of human voice. Once it has an idea of where the voice is in the audio, it cuts that out and moves it onto the next step in the process.

2. **Acoustic Model** - The acoustic model takes a representation of the waveform as input (usually in the form of a spectrogram) and tries to guess a phoneme probability distribution function as a function of time over windows of 10-80 ms throughout the entire recording. Think about how each letter makes a specific noise, and how some combinations, like "ph," are unique noises. These sounds that make up words are called phonemes. The output is a large lattice of possible phonemes as a function of time.

3. **Hidden Markov Model** - The HMM or pronunciation model then takes the phoneme lattice as its input and tries to guess a word probability distribution function as a function of time over the time window. These guessed words are statistically based similar to the phonemes; i.e. how likely would these sounds make this word.The output is a huge lattice of possible words as a function of time. This step is also called the dictionary.

4. **Beam Search-Language Model** - The language model is used in conjunction with a beam search. "Pair" and "pear" sound the same, but are different words. How do you know which word is being said? You can use the context around the word to identify what it means. If I am asking you for a "pair of headphones," you understand from the context which "pair" is being used. ASR uses the same trick to identify which word is being spoken. As more of the text is output from the phonemes, an ASR process might go back and correct itself. It cuts down all the possibilities it thinks are less likely until it arrives at the final transcription. The "Beam" is how wide the search is for the word in context of the words surrounding it.

HMM had fallen out of use by the early 2000 as this architecture was slow and inaccurate especially in noisy, multi-speaker environments, like phone calls, meetings and conferences.

HMM-DNN

This architecture is based on the HMM but adds deep neural networks to assist with the denoise and

HMM-GMM

acoustic model steps. DNN helps to speed up and improve the accuracy of this acoustic step.

Each DNN is modeled and trained independently for each step so that each step can be optimized. However, global optimization of the entire process is very difficult. In addition, each step has set parameters built into the model that carries through the entire process, which leads to additional errors and inaccuracy. Because it is still using the assembly-line process, speed remains an issue, especially for real-time streaming or transcribing large amounts of audio data.

Currently, HMM-DMM is the dominant architecture in speech recognition today with Amazon, Google, Microsoft, and Nuance all using this architecture.

E2EDL

End-to-End Speech Recognition or End-to-End Deep Learning Speech Recognition is the third and newest technology in production. E2EDL takes the DNN architecture and uses it exclusively for the entire speech-to-text process. Audio goes into the deep learning neural network and accurate text comes out. It is truly a one-step process. Under the covers, however, E2EDL is a multi-layered neural network that can self-optimize and can be trained with audio and text data to be more accurate. Unlike HMM-DMM, global optimization is easily achieved to get the most accurate transcriptions out. In addition, tailoring E2EDL for different languages, use cases, dialects, accents, and industry terminology can more quickly and easily be achieved without recoding the entire process. E2EDL can “learn” and improve.

Industry and academic comparisons show that E2EDL outperforms the other speech recognition technologies in speed with only millisecond transcription lag and trained accuracy exceeding 90%+.

NASA, after reviewing all technologies available, has chosen E2EDL as the most accurate STT technology for their mission communications. Speech technology experts have also said that End-to-End Deep Learning technology may be the holy grail of speech recognition.

Speech Recognition Features

AUDIO INPUT TYPES
Type of audio format can be accepted; i.e. WAV, MP3, FLAC, and ACC.

AUDIO TIMESTAMP
Add timestamps to each word or utterance with specific start and end times for easier search.

BATCH OR PRE-RECORDED TRANSCRIPTIONS
Ability to take pre-recorded audio and transcribe them.

CAPITALIZATION AND PUNCTUATION Type
Add capitalization and punctuation into transcriptions to improve readability.

CONFIDENCES
Each word and entire transcript is rated on confidence that the word or transcript is correct.

CUSTOMIZED/TAILORED SPEECH MODELS
Ability to customize the speech model to your use case, dialect, accent, noise, and terminology.

DEEP SEARCH
Search audio for words or phrases. Unlike text search, deep search looks for similar audio wave patterns that may be transcribed incorrectly.

DEPLOYMENT
Options for speech recognition solution deployment, on-premise, in the cloud, or on a private cloud.

DIARIZATION
Identify and separate different speakers by their utterances to ease readability and determine who said what.

INTERIM TRANSCRIPTION
During streaming, transcriptions are sent immediately and transcriptions are updated as more audio becomes available for analysis.

KEYWORD BOOSTING/ CUSTOM VOCABULARY
Add words or phrases such as industry terms, unique product names, and jargon that will be boosted to increase transcription accuracy.

LANGUAGE DETECTION
Ability to detect the language and switch transcriptions to the correct language.

LANGUAGES
Languages that can be transcribed

MULTI-CHANNEL
Transcribe multiple channels of audio; i.e. phone call is two channels and a 5 person conference call is 5 channels.

NAMED ENTITY RECOGNITION
Recognizes alphanumerics in audio and removes whitespace between the characters in the transcript

NOISE REDUCTION
Identify and reduce background noise to improve transcription accuracy.

NUMERICAL FORMATTING
Numbers can be formatted in the transcript as words or digits.

PROFANITY FILTERING
Filter any profanity from transcriptions.

REAL-TIME OR STREAMING TRANSCRIPTION
Ability to instantly provide captions or transcripts in real-time while the audio is being streamed.

REDACTION
Automatically remove sensitive data such as private health information, credit card information, banking account numbers, and social security numbers from transcripts

SENTIMENT
Determine the sentiment of the speaker either by audio or by text; i.e. is the speaker happy, sad, mad, neutral, etc.

UTTERANCES
Segments audio speech into meaningful semantic units as on speaker may pause and then resume talking, that will be two utterances.

Speech Recognition Metrics

As you compare different speech recognition solution providers and different architectures, these are the metrics that should be considered in your evaluation, depending on your use case.

Word Error Rate (WER) - This is the standard measure of speech recognition systems. WER is defined as:

$$WER = \frac{S + D + I}{N}$$

Where S is the number of word substitutions, D is the number of word deletions, I is the number of word insertions, and N is the number of total words in the ground truth transcript. The ground truth transcript is a human transcribed passage. The lower the WER the better accuracy of the overall transcription.

WER is a widely-used industry standard for assessing the quality of speech models, but it too has its limitations. A lower WER score often corresponds to increased readability of transcriptions by a human, as the human brain is better able to "fill in the gaps" created by missing words, incorrect words, or missing keywords. However, AI systems are much weaker than humans at filling in the gaps caused by errors, even with the help of a robust language model by today's standards.

Furthermore, not all errors are created equal, and this is where WER begins to fail. WER treats all words with the same weight; i.e. "a" "the" "of" is measured the same as "Salesforce" "Big Mac" "BMW X3", but the latter words are much more important for analysis and machine response.

Word Recall Rate/Word Recognition Rate (WRR) - WRR measures the percentage of words in the truth text that were correctly predicted, or matched (i.e. true positives). In other words, this does not include insertions (where there isn't a word in the truth transcript).

$$WRR = \frac{\#\ word\ matches}{\#\ of\ words}$$

The higher the WRR the better the transcription in terms of word accuracy. WRR is again not the perfect metric as it does not include incorrect insertion of words.

The above two main metrics are fairly good to measure human readability, where humans can fill in missing words, automatically correct words, or understand context; i.e. bite is different from byte when

you are speaking about data. There are additional more specific metrics that can be used to pinpoint accuracy for data analysis and Artificial Intelligence (AI) use. Some examples of these new metrics are shown below.

Alphanumeric Word Error Rate (anWER) - Suppose you use a lot of alphanumeric terms like booking ID, flight IDs, or driver licenses. Because these terms need to be highly accurate or the whole term is useless, anWER is used as a measure of accuracy for these types of terms. anWER is defined as:

$$\text{anWER} = \frac{\text{\# of alphanumeric terms (S + D + I)}}{\text{Total \# of alphanumeric terms}}$$

anWER is a word matching metric. In the sentence, "My confirmation code is N2NPCK," does the STT correctly transcribe the alphanumeric code to be "N2NPCK"?

Email Address Error Rate (emER) - This is similar to anWER but focused on email addresses. Transcribing email addresses is another use case-specific metric. Email addresses are notoriously difficult to transcribe and get correct, as most are not actual words. And if you get one digit or alpha character incorrect, the email address is useless, so this is an all or nothing term in usefulness. Because of this, instead of using term matching, we used pattern matching to get our metric. emER is looking for a specific pattern that has "@" symbol and "XXX.XXX" The formula can be defined as:

$$\text{emER} = \frac{\text{\# of email addresses incorrect}}{\text{\# of email address terms}}$$

The smaller the emER the better the STT in correctly transcribing email addresses. emER is very important for political campaigns signing up donors or voters, recruiting new employees, or automated email support.

Along the same line as emER and anWER, you can also look at other specific error rates.

Capitalization Error Rate (CapER) - Measure of how many terms are capitalized incorrectly over all terms needing to be capitalized

Punctuation Error Rate (PER) - Measure of punctuation errors over all punctuation tokens in the transcript.

There are many more term-specific error rates which may or may not be important to you depending on your use case. In addition to measuring accuracy, other metrics to consider when researching speech recognition solutions are the following:

Real-Time/Streaming Transcription Lag - How long does it take for the last word to appear in the transcription after the speaker has completed his utterance. This metric is important for voicebots, captions, and real-time analytics. Speaker recognition systems streaming lag range from 200 milliseconds to 4 seconds.

Transcription Speed Up - Transcription speed up of pre-recorded audio; i.e. one hour of audio takes 30 minutes to transcribe. Speaker recognition systems speed up ranges from 1X (one hour audio transcribed in one hour) to 120X (one hour audio transcribed in 30 seconds).

Scaleup Optimization - As you scale up your speech recognition, how much computing power will be needed per real-time stream or hours of audio per day. HMM-GMM runs on CPUs, HMM-DNN runs on both CPUs and GPUs and E2ESR runs on GPUs.

Costs - What are the costs per hour of using a speech recognition service. Taking into account free credits, volume costs, up charging to 15 second utterances, and computing power if deployed on-premises.

Speech Recognition Use Cases

Speech recognition has progressed from simple dictation to voicebots and continues to expand with innovative companies finding additional uses. Here is a list of the current use cases.

Call Analytics

Captioning

Command and Control Applications

Compliance

Conversational AI/Voicebots

Conversation Intelligence and Analysis

Customer Experience Analysis

Dictation

Electronic Health Record Input

Earnings Call Analysis

Gaming Profanity Alerts

Intelligence Gathering

Interactive Voice Response

Language Fluency Testing

Media Management

Meeting Summary and Management

Police Bodycam Analysis

Product Ideation from Customer Input

Recruiting Interview Analysis

Retail Employee Coaching

Sales Coaching

Sales Onboarding

Sales or Support Enablement

Shopping

Social Media Profanity Alerts

Telehealth Transcriptions

Video and Audio Transcriptions

Video and Podcast Attribution

Voice-Enabled Machines

Voice of the Customer Analysis

Workforce Optimization

Work Simulations

Learn more about how innovative companies are using speech recognition:

Call Analytics

Case Study:

Conversational AI/Voicebots

Case Study:

Sales Onboarding

Case Study:

Recruiting

Video: Case Study:

Voice of the Customer

Case Study:

Call Analytics

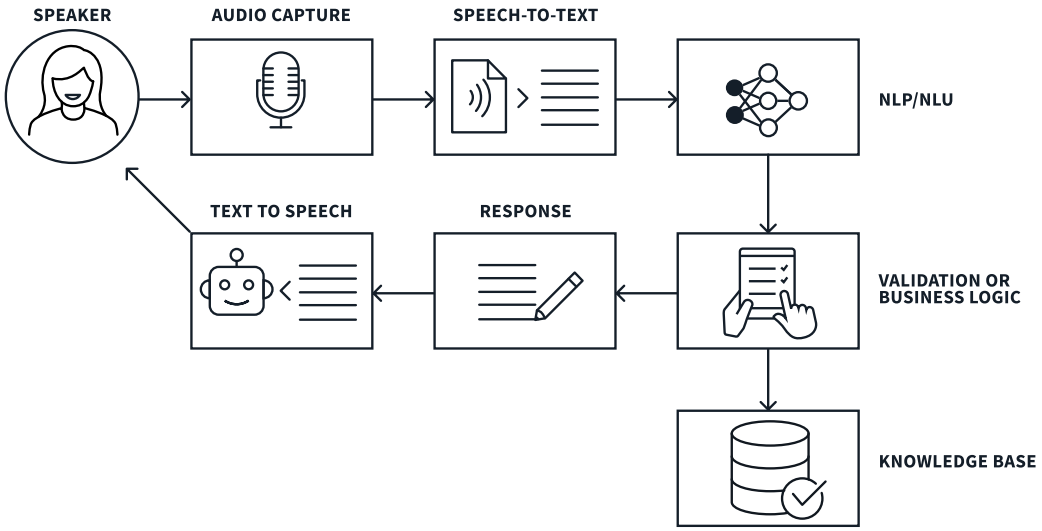
Case Study:

Voice Technology Development Path

Of the various use cases for speech recognition, these use the latest speech recognition technology and fully optimized the capabilities. The reference architecture and tips on starting development are linked below.

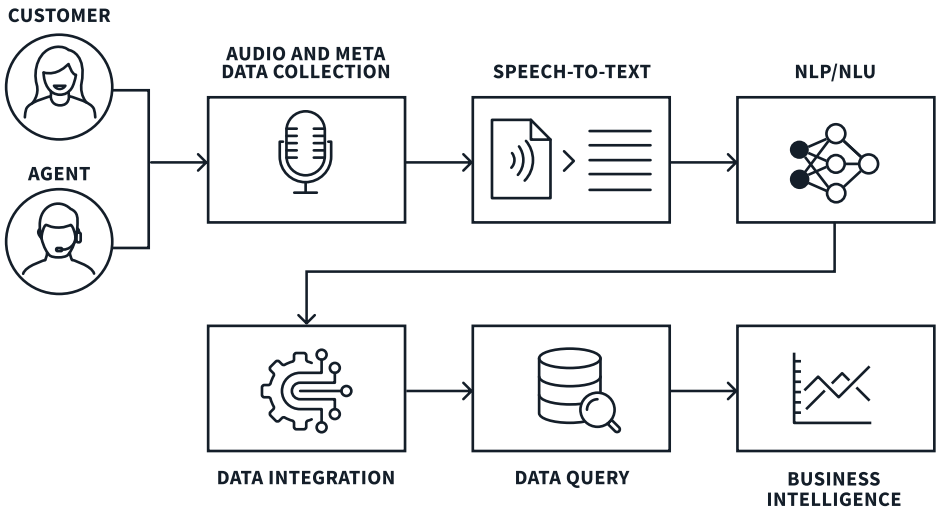
Conversational AI/Voicebots Development

Learn More →



Call Analytics Development

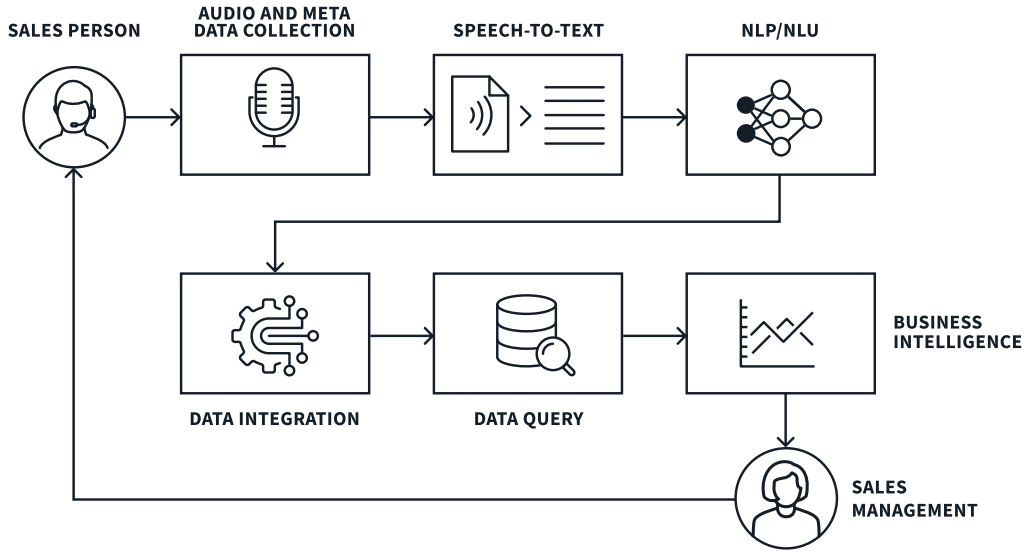
Learn More →



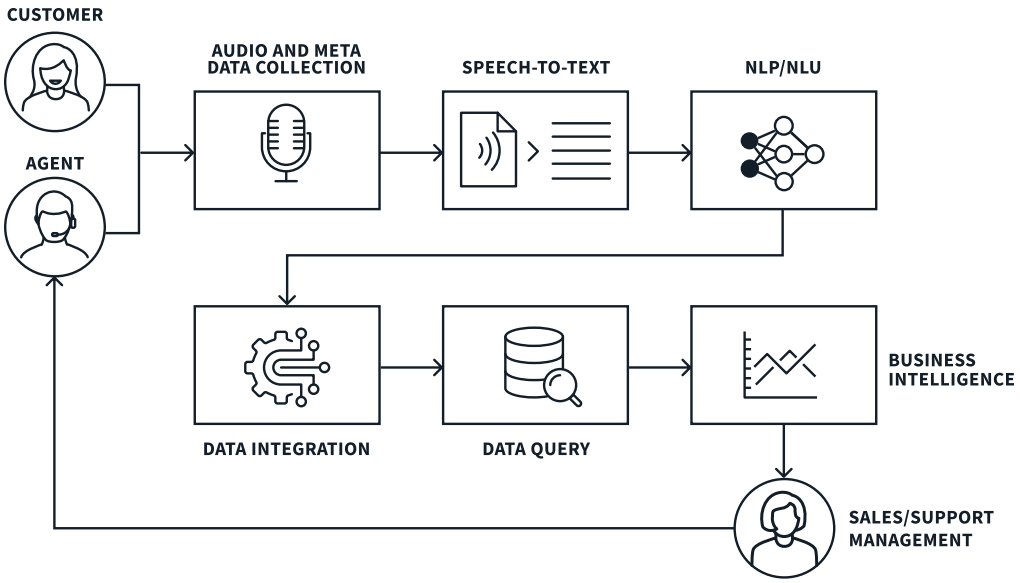
Sales and Support Enablement Development

Learn More →

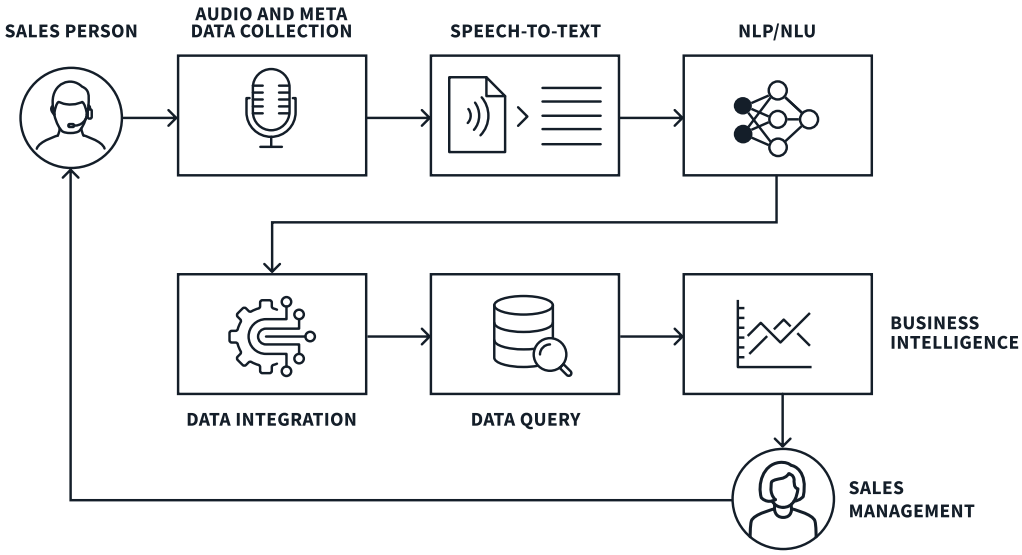
ONBOARDING REFERENCE ARCHITECTURE



COACHING REFERENCE ARCHITECTURE



REAL-TIME SALES SUPPORT REFERENCE ARCHITECTURE:



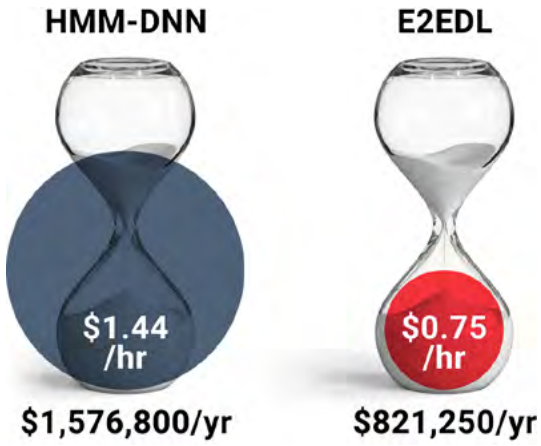
Follow us on LinkedIn for additional development path cases.

The Associated Costs of Speech Recognition

The cost of Speech Recognition isn't always just the cost of transcription. Though that's a key figure to consider. Here is a list of costs to be aware of when you are evaluating speech recognition solutions.

Transcription Fees

Most **HMM-DNN** providers tend to offer cloud transcription services at a rate of $1.44/hour. Deepgram — currently the only **E2EDL** provider on the market — offers transcription services at $0.75/hour. A $0.69 difference may not seem dramatic, but it can be an issue for users who transcribe a significant amount of audio. For example, at a rate of 3,000 hours of pre-recorded audio transcription per day, the annual cost for HMM-DNN vs. E2EDL speech recognition is a difference of hundreds of thousands of dollars.

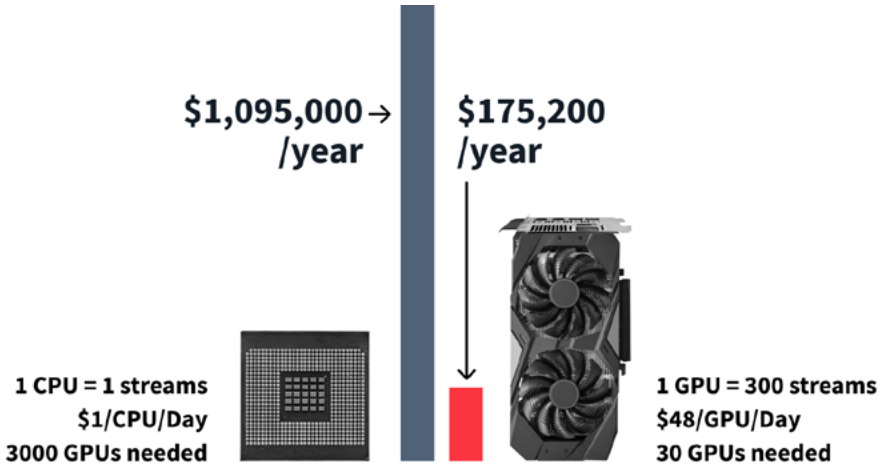


It's also worth noting, that some HMM-DNN services round up their fees to the nearest 15 seconds, which the image above does not account for. Again, while that may seem insignificant in small quantities, when you're transcribing thousands of hours of audio per month, those round-up fees can drive up costs dramatically.

On-Premises Operations Costs

One of the key differences between HMM-DNN and E2EDL is their speed. That can be attributed to the hardware used in each process. HMM-DNN uses CPUs which requires one CPU per transcription stream. E2EDL uses GPUs. One GPU can process 300 streams at a time. While the cost of one CPU ($1/day) vs. one GPU ($48/day) is a significant difference, the capabilities of the latter offset the cost. For example, if you're processesing 3,000 streams of audio concurrently per day, with HMM-DNN you need 3,000 CPUs, or $3,000/day. With E2EDL, because each GPU can process 300 streams, you only need 10 GPUs, or $480/day.

Let’s take a look at how that plays out over the course of a year:



Investing in Customization

Another area you may want to account for in budgeting your Speech Recognition is in customization. Whether you have jargon particular to your business or have a unique use case, in may be worth improving your transcription with keyword boosting or data labeling and model training. Just note that there is a greater time/cost investment to do this with the HMM-DNN model vs E2EDL as the latter is a one-step process vs. a multi-step process and hence easier to optimize in a shorter period of time.

Future of Speech Recognition

Speech recognition is already part of many consumers' lives through Alexa, Cortana, Google, and Siri. We see this voice technology trend moving swiftly into your business lives. The innovation from start-ups to larger corporations will change the way we run businesses. At a point in the future, you will see voice technology be in every type of business touching many functions from sales, marketing, production, operations, finance, fulfillment, customer support, and product development.

For speech recognition specifically, we will see increased accuracy rivaling human transcriptionists, truly instantaneous transcription speed, and self-learning speech models. E2ESR will continually optimize itself as it is fed more and more data.

Speech models will be more and more tailored to the use case, industry, intent, and language. One voicebot may have hundreds of speech models that can handle multiple countries for technical support, account support, the recommendation for upsells, and conversations about the weather and your children. It will be personalized to your specific customer quickly and easily after engaging with that customer for a few minutes.

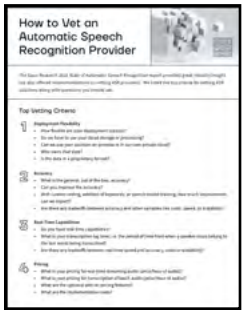
These speech models along with natural language understanding may be a conversational companion remembering all your conversations, understanding metaphors, colloquialisms, and jargon.

What future can you imagine in a voice-enabled business?

Speech Recognition Evaluation Assets

How do you choose the right speech recognition architecture and features? What is the path for testing? Ultimately, how do you choose the right ASR partner? Here are a few helpful resources to guide your process.

Tips to get started:



HOW TO VET AN ASR PROVIDER

10 STEPS TO CALCULATE WER

HOW TO EVALUATE A DEEP LEARNING ASR PLATFORM

HOW TO MAKE YOUR APPLICATION VOICE-READY

AUTOMATIC SPEECH RECOGNITION: PAST, PRESENT, AND FUTURE

2021 STATE OF AUTOMATIC SPEECH RECOGNITION REPORT

Webinars and Assessments:

HEAD-TO-HEAD: DEEPGRAM VS. GOOGLE SPEECH-TO-TEXT

HEAD-TO-HEAD: DEEPGRAM VS. AMAZON TRANSCRIBE

SELF-ASSESSMENT FOR BUSINESS LEADERS