Introduction to Speech AI

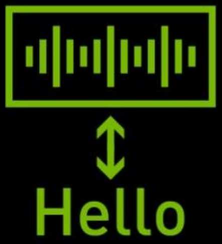

Table of Contents

Preface

Who is this e-book series for?

This e-book series is intended for business decision owners and developers within an enterprise who would like to understand the core concepts of speech AI and how to build and deploy applications for different use-cases.

The series consists of three parts:

Part 1: Introduction to Speech AI
Part 2: End-To-End Speech AI Pipelines
Part 3: Building Speech AI Applications

Over the course of this e-book series, you will learn about:

- Speech AI and its major building blocks.
- The role of speech AI in industries.
- Different components and technologies that are involved in building end-to-end speech AI pipelines
- How to get started with speech AI in your business with NVIDIA Riva.

Part 1: Introduction to Speech AI

Advances in artificial intelligence and high performance computing technologies have revolutionized numerous areas of human interactions. We have seen humans conversing with human-like robots in science fiction movies for many years. Today, thanks to what we call “Speech AI”, we effortlessly converse with “technology” - our smartphones, smart homes, websites, and even our cars.

Just as mobile devices transformed the way we live, so have these speech AI systems, which are becoming ubiquitous. Countless products are taking advantage of automatic speech recognition (a.k.a., voice recognition, speech-to-text) and speech synthesis (a.k.a., voice synthesis, text-to-speech). Thanks to new tools and technologies, developing speech AI applications is easier than ever, enabling a much broader range of applications, such as virtual assistants, real-time transcription, and many more.

In this e-book series, we provide an overview of the speech AI landscape - how it works and how important it is to various industries. We will glimpse into the evolution of speech AI, discuss the challenges in building speech AI applications, and more importantly, describe how to get started with integrating speech AI skills into your applications.

What is Speech AI

Speech AI is the use of machine learning to help humans converse with devices, machines, and computers via speech. An everyday example is when you drive and want to know where the nearest gas station is. Without an effective speech interface, many drivers would have to manually click on the phone map and search for the information while still driving. A much safer and more convenient way is to say, “Navigate to the nearest gas station”. Your smart map application then searches for a few options, reads them aloud, and asks you to confirm. You select your favorite option, and your map navigates you to your chosen destination. This reduces distracted driving, as the interactions just described do not require manual or eye contact with the mobile device.

Another popular example of speech AI is a home virtual assistant. A home virtual assistant is a smart device that communicates with users via a speech interface, and assists with various tasks, from smart home automation (e.g., voice commands like, turn on the TV, turn off lights in the living room, and so on) to fetching weather information, playing music, and answering trivia questions. Virtual assistants based on AI have come a long way in interacting with humans in a human-like way.

New deep learning-based speech AI applications are being introduced every day, from voice assistants and AI-powered chatbots, to question-answering systems that enhance customer service.

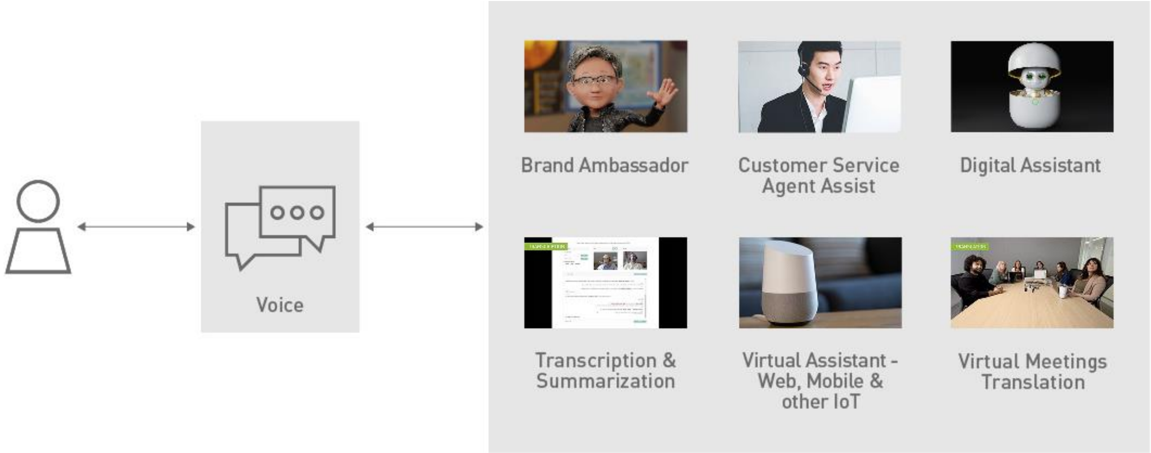
Voice
Brand Ambassador
Customer Service Agent Assist
Digital Assistant
Transcription & Summarization
Virtual Assistant - Web, Mobile & other IoT
Virtual Meetings Translation

Figure 1: Speech AI Applications

A speech AI application is a complex system that integrates multiple deep neural networks that must work in unison to deliver a delightful user experience with accurate, fast, and natural human-to-machine voice-based interaction. The ultimate goal of speech AI is to make conversing with machines indistinguishable from talking to a person.

How a Speech AI System Works

A speech AI system includes two main components:

- **An automatic speech recognition (ASR) system**, also known as speech-to-text, speech recognition, or voice recognition system.
  It converts the raw speech audio signal into text for processing by subsequent components.
- **A text-to-speech (TTS) system**, also known as speech synthesis.
  It turns the text into audio.

The technology behind speech AI is complex. It involves a multi-step process requiring a massive amount of computing power and several deep learning models that must run in tens of milliseconds to deliver human-like responses.

Speech AI components typically form part of a larger voice-based conversational AI system, which combines various technologies such as Automatic Speech Recognition, Natural Language Processing (NLP), Text-To-Speech and Dialog Manager to understand and respond to different interactions.

In most cases, a voice-based conversational AI pipeline consists of three stages:

- Automatic Speech Recognition
- Natural Language Processing and Dialog Manager
- Text-To-Speech

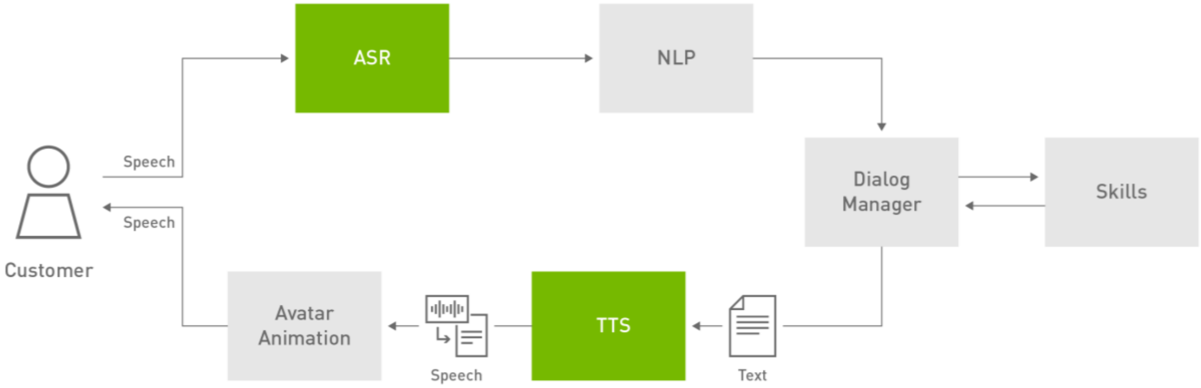

Figure 2: Voice-Based Digital Avatar Pipeline

First, the raw audio is an input to the ASR system. With ASR, the audio is processed and transcribed to text.

Second, the application needs to understand what the text means and act accordingly. This is the job of an NLP and dialog system which manages the conversation with the user while interacting with external fulfillment systems, also known as skills, to correctly respond to the user. Next, the output from the ASR phase is interpreted by the NLP system and acted upon by the Dialog Manager with the assistance of a relevant skill.

In the final step, TTS converts the text response into speech. It is in this step thatAI is used to produce human-like speech from text.

Modern state-of-the-art speech AI systems make extensive use of deep neural network models trained on massive datasets. Over time, the size (i.e., number of parameters) of speech AI models has grown. Training such models can take weeks of intensive compute time and is usually performed using deep learning frameworks, such as PyTorch, TensorFlow, and MXNet on high-performance GPUs. There are several large and freely available speech datasets, such as Mozilla Common Voice, to train models for speech AI tasks. However, models trained on public datasets rarely meet the quality and performance expectations of enterprise apps as they often lack context about the industry, domain, company, and products.

A proven approach to address these challenges is to use transfer learning: You start from a model pre-trained on a generic dataset and apply transfer learning and fine-tuning for your use case using a small amount of domain-specific data. Fine-tuning is far less compute-intensive and data-demanding than training a model from scratch.

During inference, several models need to work in tandem to generate a response for a single query, requiring the latency for a single model to be only a few milliseconds. Latency is the elapsed time between the computed output and the available input. GPUs are used both to train deep learning models and to perform inference because they can deliver 10X higher performance than CPU-only platforms. Given the latency requirements, GPUs make it practical to use the most advanced speech AI models in production for real-time apps.

Evolution of Speech AI - From Statistical to Deep Learning Approaches

Humans recognized the significance of speech-to-text conversion thousands of years ago. Over 5000 years ago, humans converted spoken language into written form in order to transfer knowledge to the future generations. Similarly, computer-generated voices play an important role in knowledge transfer. An early example of speech synthesis technology is the use of a speech synthesizer by the iconic theoretical physicist Stephen Hawking, who after being diagnosed with motor neuron disease (MND) which causes difficulty in speech, used a speech synthesizer to continue communicating and teaching.

Automatic Speech Recognition

Research in speech recognition started as early as in the 1960s, but practical speech recognition systems only appeared in the 1990s. Automatic Speech Recognition (ASR) systems from this era draw techniques from signal processing and statistics. Hidden Markov models (HMMs) were widely used in early ASR experiments and were far from smooth and pleasant due to their relatively lower accuracy.

The latest round of innovation in the ASR field occurred around 2014 when researchers attempted deep learning techniques. Deep learning involves using deep neural networks consisting of millions of "neurons" and hundreds of millions of "links" between these neurons, each associated with a tunable parameter. The organization of the neurons into multiple sequential layers is the origin of the term "deep". Neural networks are known to be universal approximators, meaning that they are capable of approximating any arbitrarily complex function given sufficient parameters. For a given parameter budget, when organized as a deep network of successive processing steps, i.e., going "deep", neural networks have more expressive power than ones having fewer layers which consist of more neurons per layer, i.e., going "wide". In order to prevent overfitting, in which the model memorizes the data rather than discovering the pattern governing it, deep neural networks are trained on massive amounts of speech data instead.

Deep neural networks allow "end-to-end" ASR, meaning that they can jointly model the pronunciation and acoustic aspects of spoken languages, whereas HMM-based approaches require separate components for each. Deep learning implementation in ASR significantly lowered the word error rate down to a level matching or better than human error rate (yes, humans also make mistakes when doing speech-to-text at an estimated 4-5% error rate), making ASR a much more effective and pleasant experience.

Text To Speech

Similar to ASR, text-to-speech (TTS) or speech synthesis has a long history of research and development. The simplest and earliest systems used pre-recorded phones or diphones that were then concatenated to make longer phrases and sentences. These early approaches produced speech that lacked both clarity and naturalness.

In later statistical approaches, such as Hidden Markov models, the frequency spectrum (vocal tract), fundamental frequency (voice source), and duration (prosody) of speech were modeled simultaneously by HMMs. Speech waveforms were generated from HMMs based on the maximum likelihood criterion and have better fluency than the concatenation method.

Recent innovation in TTS revolves around the use of deep neural networks. Deep neural networks are trained using a large amount of recorded speech to approach the naturalness of the human voice. Advances in transfer learning in recent times allow an existing speech synthesis model to be adapted to a new voice actor/actress within as little as 30 minutes of recorded speech data, making it easy to create a unique, high-quality voice personality.

Challenges in Building Speech AI Applications

Speech AI is highly beneficial, however, there are several challenges to building universal, real-time and high-quality speech AI applications. In this section, we will cover some of the key challenges that should be addressed when developing speech AI applications, such as achieving high performance and scalability, high accuracy, support for multiple languages, and ensuring data security and privacy.

High Performance and Scalability

For a quality and engaging conversation between a human and a machine, responses have to be quick, intelligent, and natural. These requirements pose the following challenges:

- Computations of the full pipeline consisting of several deep learning models, each with millions of parameters, must complete under 300 ms, which empirically was found to be the limit for a natural experience.

- Trade-offs between speed and quality of the response must be considered.
  More complex models usually result in better ASR and TTS quality but require more computation time and power.
- Millions of concurrent users causing high latency must be managed.

High computing power and highly optimized speech AI software and libraries ensure high performance and low latency, which are essential to speech AI applications.

High Accuracy

Speech-to-text systems should be highly accurate to use. Even a low percentage of word error rate in dictation or a voice command system might cause a significant inconvenience and result in unnecessary corrective actions. For example, the wrong recognized location in the voice-based navigation system might add a few hours to getting to your destination. Some users would stick to a traditional interface, such as a keyboard, mouse, or touch pad, to avoid such occasional inconveniences.

Similarly, text-to-speech applications must be natural as well as accurate. Wrong pronunciation of acronyms or uncommon terminologies, may be misunderstood or convey incorrect information.

For building highly accurate speech-to-text systems and natural text-to-speech systems, a large quantity of high-quality data is critical, and can be challenging to curate.

Multilingualism

There are over 6500 spoken languages in the world. Language diversity and factors such as accent, dialects, pronunciation, and slang must be integrated into the speech AI system. Users trust applications when they are able to communicate in their mother language. Therefore, gathering diverse training data from different regions and considering cultural differences is critical in developing a universal application.

Security and Privacy

Ensuring that data are securely processed and stored is crucial when building speech AI applications. Organizations must define and ensure high-security standards and privacy when planning to build speech AI applications to gain customers’ trust. There should be transparency in how the organization uses the data to address privacy concerns.

In summary, building a real-time and high-quality speech AI application with little upfront investment comes with several challenges and ensuring that these challenges are addressed is the key to a successful speech AI application.

Role of Speech AI in Industries: Today and Beyond

Speech AI is central to the future of many industries, as applications gain the ability to understand and communicate naturally using a voice-based interface. The range of industries adopting speech AI into their solutions is wide and growing, with diverse domains extending from finance, telecommunications, and retail to healthcare. Given the constraints and challenges in building speech AI, employing GPU-optimized speech AI models is a logical solution. Speech AI can power advanced digital voice assistants in smart speakers and customer service lines, enabling a wide range of businesses to attain an unprecedented standard of personalized customer service.

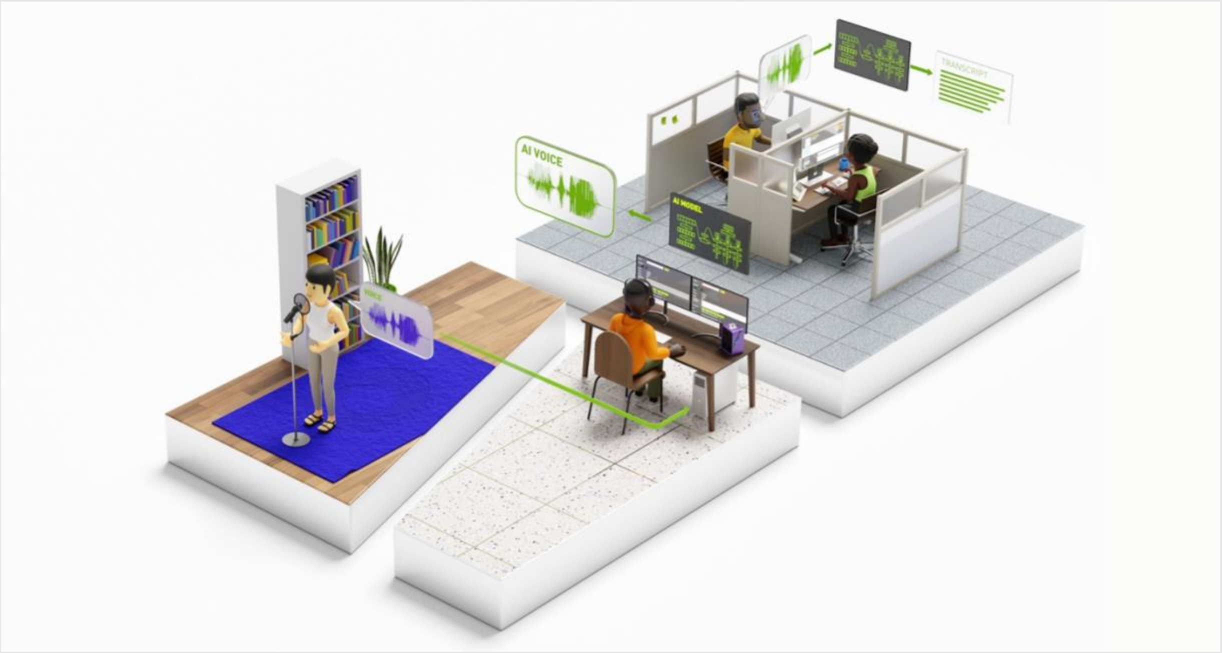

Figure 3: Speech AI Industry Applications

Telecommunications

Delivering high-quality and seamless customer service is key to the success of the telecommunication industry. Approximately 10 million call center agents worldwide answer about 2 billion phone calls daily. Spikes in call volume can lead to bottlenecks, increased customer waiting time, and ultimately lead to low customer satisfaction and churn. Speech AI systems, especially with GPU-accelerated applications, can speed up resolution of customer issues by transcribing calls and enabling automated analysis to provide timely and intelligent responses and solutions. Speech AI transcribed calls can also analyzed for identifying any decline in customer satisfaction, threatening behavior, and fraud attempts.

Financial Services

According to Juniper Research, banks have the potential to automate up to 90% of their customer interactions using chatbots by 2022.

An intelligent agent is capable of understanding complex problems and is able to provide information can resolve customer call inquiries quickly. Call center traffic and waiting time can be greatly reduced while improving customer satisfaction by employing an intelligent agent.

Healthcare

One of the challenges in healthcare today is access to health services. Calling your doctor's office and waiting on hold is a common occurrence. Connecting with a claims representative can be challenging; there are an estimated 30 million consultations daily. These audio conversations can be analyzed with ASR technology.

A healthcare call center virtual assistant can augment health workers' efficiency using a database of medical information that can be provided quickly to address patients' enquiries—reducing the wait time for a medical on-call support. In addition, speech recognition can help physicians to voice record clinical notes instead of typing, giving them more time for other patients. On average, according to "Better Health Outcomes with AI-powered Virtual Assistants", patient handling time can be reduced by 20 percent, resulting in cost benefits of hundreds of thousands of dollars.

Unified Communication as a Service

Unified Communications as a Service (UCaaS) is a cloud-based communications model that offers integrated enterprise communications such as audio/video/web conferencing, instant messaging, dedicated communication channels, and emails. Every single day, 70 million hours of web meetings are held on different conferencing platforms. The increase of these virtual meetings require the development of applications that can enhance the quality of these meetings and provide the users with high quality service.

Speech AI is an answer to what every UCaaS platform needs. ASR algorithms in Speech AI make it possible to automatically transcribe meetings, lectures, and social conversations while simultaneously identifying speakers and labeling their contributions, reducing manual work. This includes multispeaker transcription and generating meeting notes.

Retail

Consumers visit approximately 12 million retail stores every day. Voice-based assistants serving customers are a key resources for contactless commerce; they provide virtual assistance for curbside pickup and in-store assistance. These assistants can connect with customers and give them a memorable shopping experiences associated with the brand.

Conclusion

In this part of the e-book series, we introduced Speech AI and described how a generic modern Speech AI system works. An overview of speech synthesis technology can be found in, "An Overview of Speech Synthesis Technology" [1]. An extended and accessible introduction to speech recognition technology, for the general audience can be found in "The Voice in the Machine. Building Computers That Understand Speech" [2].

Next, we provided a brief overview of the historical evolution of Speech AI approaches, a topic that deserves a book on its own. Therefore, we limited our description to focus on practical and useful information. To learn more about developments in ASR technology development, please refer to articles and papers [3], [4], and [5]. For details on TTS refer to “History of speech synthesis” [6].

We have outlined the challenges in building AI Applications.

The role of Speech AI is highlighted using a few areas of industry. Other use-cases, which are beyond the scope of this book, that deserve to be explored are:

automatic subtitling, translation, education, home automation, robotics, video games, assistive technology for people with disabilities, and so on.

Speech AI is a vital ingredient in building modern human-like systems and is here to stay. Once a voice-based interface becomes effortless, efficient, and natural, it will be hard to reverse this experience, just like smart phones revolutionized our telecommunications, and laptops our personal computing experience.

In the next part, Part 2 of this e-book series, we will explain the different technologies that are involved in building end-to-end speech AI pipelines.

References

- [1] Z. Yin, "An Overview of Speech Synthesis Technology", 2018 Eighth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC), 2018, pp. 522-526, DOI: 10.1109/IMCCC.2018.00116.
- [2] Roberto Pieraccini (2012), "The Voice in the Machine. Building Computers That Understand Speech".
- [3] Juang, B. H., Rabiner, Lawrence R., "Automatic speech recognition–a brief history of the technology development"
- [4] Melanie Pinola (2 November 2011), "Speech Recognition Through the Decades: How We Ended Up With Siri". PC World. Retrieved 22 October 2018.
- [5] Benesty, Jacob; Sondhi, M. M.; Huang, Yiteng (2008), Springer Handbook of Speech Processing. Springer Science & Business Media.
- [6] Story, Brad. (2019). History of speech synthesis. DOI: 10.4324/9780429056253-2.