



EVOLUTION IN SPEECH AI SHAPING THE NEW WORLD OF INTERACTIONS

Abstract

Speech AI is a field of technology that bridges the gap between human speech and machines by allowing machines to understand, interpret, and even generate spoken language. These capabilities are transforming how we interact with devices and information, allowing for natural conversations with voice-activated assistants. But Speech AI has a wide range of applications across industries beyond just voice assistants. An example is call center automation, where Speech AI allows machines to handle simple inquiries and helps human agents resolve complex issues by providing proactive real-time assistance.

Discover how [Infosys Cortex](#) leverages Speech AI to offer innovative customer service use cases.

This two-part document delves into how the exciting interplay between speech AI research and its applications is promising to transform our lives in remarkable ways. In Part 1, we explore the recent technological advancements in speech AI while in Part 2, we will present our point

of view on emerging trends resulting from these technological advancements.

How Speech AI Works?

Speech AI typically consists of the following technologies:

Speech Recognition


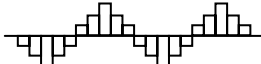
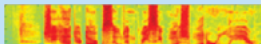
It converts spoken words into text that machines can understand. This technology powers voice assistants like Siri and

Alexa, allowing them to understand your instructions.

Speech Synthesis

It flips things around, where machines use AI to transform written text into spoken words, often in a way that sounds natural. This is the secret sauce of Siri and Alexa responding to you.

Let's break down speech recognition to understand the sequence of techniques involved.

| S. No. | Process | Commonly Done Through | Output | Example Output |
|--------|--------------------------------|---|--|---|
| 1 | Capture Voice | Microphone | Electrical Signal (Analog Voltage) |  |
| 2 | Digital Conversion | Sound Card | Digital format (series of 0s and 1s) |  |
| 3 | Speech Features Extraction | Mel-Frequency Cepstral Coefficients (MFCCs) | Speech signal features like pitch, volume, spectral energy distribution and formants |  |
| 4 | Acoustic Modeling | Acoustic Model | List of possible word sequences ranked by their likelihood | [i], [want], [to, two], [meet, meat], [you], [at], [ten], [today] |
| 5 | Decoding | Decoder | List of probable sentences based on acoustics | i want to meet you at ten today, i want to meat you at ten today, i want two meet you at ten today, i want two meat you at ten today |
| 6 | Language Modeling | Language Model | Most probable sentence based on grammar and context | i want to meet you at ten today |
| 7 | Language Specific Punctuations | Punctuation Model | Sentence with Punctuation | i want to meet you at ten today. |
| 8 | Inverse Text Normalization | ITN Model/Rules | Final Output Sentence | i want to meet you at 10:00 today. |



Speech AI is Evolving Rapidly

As per [Fortune Business Insights](#), the global speech and voice recognition market size in 2023 was \$12.62 billion and is projected to grow at a CAGR of 23.7% to reach \$84.97 billion by 2032.

Speech AI is undergoing a revolutionary transformation with deep learning powering new heights of accuracy, understanding even complex accents and background noise. This, coupled with the progress in natural language processing, is allowing AI to grasp the user's intent. The result? Voice assistants that hold natural conversations, real-time translation that shatters language barriers, and transcription tools that empower everyone. Here are a few recent developments in this field:

Accuracy

Speech AI is rapidly improving, thanks to large datasets of countless hours of spoken language, powerful GPU hardware for unprecedented speed, and efficiency and research on newer language modeling techniques to decipher the nuances of human language.

[500+ hours](#) of content is uploaded to YouTube every minute. Imagine the scale of data availability!

Language Inclusiveness

The availability of a large audio dataset and the ability to generate realistic synthetic speech samples for training and model fine-tuning allows pre-trained models to specialize in many indigenous languages, accents, and dialects bringing accessibility to a wide range of users.

[Two-third of popular videos on YouTube are not in English!](#)

Infosys Cortex [Language Neutralization](#) allows effective communication in a contact center between customers and agents who interact in different languages.

Domain Inclusiveness

Speech AI is gaining new levels of sophistication with the rise of domain-specific models. By training on massive datasets specific to a particular field, like medicine or legal proceedings, these models become accustomed to the unique vocabulary and sentence structures used within that domain. Consequently, Speech AI can assist in tasks like medical transcription or legal document analysis, making it a valuable tool for professionals across various industries.

[Kaggle](#), [LibriSpeech](#), etc. are hosting vast collection of domain datasets.

Realtime Processing

While CPUs were inefficient for very complex deep learning computations, GPUs have significantly sped up real-time speech processing. Their parallel processing architecture enables them to handle complex calculations efficiently, allowing for near-instantaneous speech AI, and fostering realism in conversations.

CIFAR-100 image classification dataset [training](#) on Azure Standard NC4as T4 v3 CPU takes 17:55 minutes compared to 5:43 minutes on Nvidia Tesla T4 GPU showing a stark difference in performance.

The latest Nvidia H100 GPU is around 8 times faster compared to the T4 GPU on which this benchmarking was done.

Model Optimization

Model optimization and compression techniques are enabling Speech AI on resource-constrained devices like smartphones, allowing offline functionality, and expanding the reach to areas with limited connectivity or privacy-sensitive scenarios, making it a more versatile tool for everyday life.

[TensorFlow](#), Google's open-source machine learning framework, also provides a specialized version of [LiteRT](#) (formerly TensorFlow Lite), optimized for on-device machine learning which is suitable for mobile and embedded devices with limited resources.

Overcoming Noisy Environments

Speech AI is conquering the challenge of noisy environments with a multi-pronged approach. Noise reduction algorithms can filter out unwanted background sounds, while training with noise-corrupted data helps the models learn to differentiate between speech and interference. Voice Activity Detection (VAD) identifies periods of silence or non-speech sounds and allows the model to focus only on the speech. Additionally, noise-canceling microphones play a crucial role by physically attenuating background noise before it even reaches the AI for processing. This combined effort ensures that AI can accurately understand speech even in noisy conditions, improving the reliability of voice interactions.

[Nvidia Maxine](#) provides high-quality audio, video, and augmented reality effects, including background noise reduction, through which [Nvidia is pushing to transform the \\$10 Billion Video Conferencing Industry](#).

Audio to Audio Generation

Audio-to-audio generation is emerging as a powerful tool for boosting Speech AI capabilities. This technology allows AI models to manipulate audio directly without the need to first convert it to text, offering exciting advancements in several areas. For instance, converting spoken language into another language's audio representation, retaining tone and emotions, for real-time natural-sounding translations.

OpenAI GPT-4o voice mode appears in [demos](#) to support voice response to queries without any text transcription.

Open Source speech language model [Llama Omni](#) combines [Llama Text](#) and [Whisper large](#) models to simultaneously generate speech and text with less than 250 ms latency.

Emotionally Resonant Speech

Speech AI is becoming more human-like. By analyzing context and sentiment, AI can modulate speech with emotions, creating a more engaging and natural interaction experience. This emotional intelligence builds trust and enhances user satisfaction, transforming how we interact with technology.

[OpenAI GPT-4o](#) supports very high-quality human-like voice variation.

Voice Cloning

Speech AI offers a fascinating solution for preserving speaker identity: voice cloning. By analyzing a speaker's voice recordings, AI can create a synthetic replica that captures the unique characteristics of their voice, including pitch, timbre, and even emotional inflections. This cloning

can recreate the speaker's voice for various purposes, even after they are no longer able to speak themselves. Imagine authors having their audiobooks narrated in their voice even after they've passed away or historical figures delivering their speeches using their cloned voices for educational purposes.

Spotify piloted [AI Voice Translation](#) in 2023, for translating podcasts into additional languages in the podcaster's voice.

Speaker Diarization

By analyzing audio characteristics, AI can effectively identify and label individual voices within a recording. This makes transcripts more readable and usable for applications like meeting transcriptions and call center analysis.

Nvidia Riva automated speech recognition (ASR) supports an on-prem [speaker diarization](#) feature to get each word transcript tagged with the ID of the speaker who has spoken that word.

Speaker Verification

By analyzing unique vocal characteristics against a pre-enrolled voiceprint, speech AI can confirm a speaker's identity. This secure and convenient method can be used for applications like authorizing financial transactions and granting access to restricted areas.

[Microsoft Teams identifies in-room participants in a meeting by using digital voice profiles of participants.](#)

AI can identify the language by analyzing acoustic properties and phonetic patterns. This unlocks possibilities like routing callers to appropriate contact center agents and adjusting real-time captioning systems.

Google search automatically identifies the language of your spoken query to provide search results in your language.

Audio Classification

By analyzing acoustic features, AI can distinguish between music, speech, and noise. This has applications like predictive maintenance where audio classification on sounds emitted by machinery can help detect early signs of failure to prevent costly downtime, security systems taking appropriate actions by differentiating between a car alarm and a breaking window, etc.

When you say "Alexa", "Hey Siri" or "Ok Google", your voice assistant uses audio classification to determine that you're summoning it.

Audio Fingerprinting

AI can identify unique characteristics within audio, allowing it to be uniquely recognized even when it is altered or embedded within other audio content. It has practical applications like audio forensics to compare audio recordings to identify if they originate from the same source, which is often used in criminal investigations.

[Shazam](#) uses audio fingerprinting on a voice sample and compares it to a vast database of song fingerprints to identify any song in seconds.

Audio Watermarking

Speech AI is emerging as a powerful tool for audio watermarking, offering a new layer of security and attribution for spoken content. By embedding imperceptible signals into audio recordings, speech AI can encode information like copyright ownership or source identification. These watermarks are inaudible to the human ear but can be readily detected by specialized algorithms. This technology holds promise for combating plagiarism of spoken word content, such as audiobooks or podcasts.

Netflix and other similar service providers protect their content by using technologies such as Digital Rights Management (DRM), based on encryption-decryption and digital watermarking in audio and video.

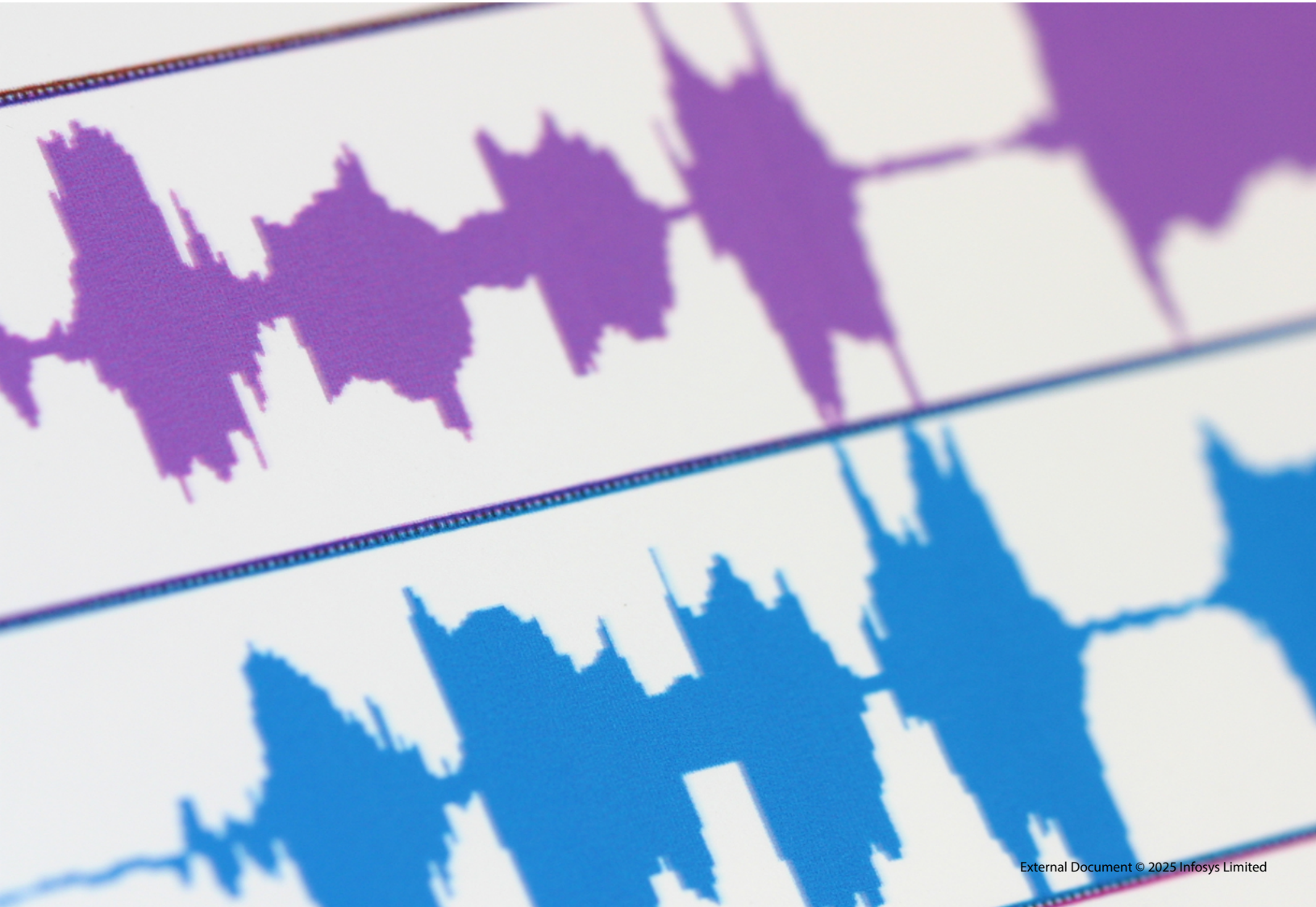
Audio Mining

Speech AI acts as a digital pickaxe in the goldmine of audio data, allowing various analyses like identifying trends and sentiment in customer service calls or gauging public opinion from social media audio clips, etc.

Spotify uses AI to analyze and understand the music content it hosts. By analyzing these features, Spotify recommends similar songs, creates curated playlists like Discover Weekly, and improves search results.

Emerging Trends Through These Advancements

Speech AI is having a big impact on industries providing unprecedented convenience to consumers. Please refer to Part 2 of this document, where we delve into emerging trends reflecting the latest development in Speech AI technology as well as the associated challenges.



About the Authors

Samit Sawal

Samit Sawal is a Senior Architect with 17 years of experience which includes incubating emerging tech, building IP, accelerators, platforms, and product engineering with a strong understanding of technologies such as Conversational AI, Generative AI, and domains like Customer Service and Core Banking.

Amit Kumar

Seasoned AI leader with 17+ years of experience, 3 patents, and expertise in generative AI, classical AI, discriminative AI, and hybrid architectures. Mastery of LLMs, multi-agent AI, RAG, and SLMs. Thought leader with publications and patents. Deep technical proficiency in AI, MLOps, and responsible AI. Experienced in designing and implementing enterprise-grade AI solutions.

Pankaj Negi

Pankaj Negi is a Principal Consultant at Infosys Center for Emerging Technology Solutions. He brings 18+ years of experience as an emerging technology incubator, innovator, digital transformation consultant, strategist, and a product manager. Pankaj holds a bachelor's degree in electronics engineering and an M.B.A. from SP Jain Institute of Management and Research, Mumbai.

Srushti Kadam

Srushti Kadam is a Senior Associate Consultant at Infosys Centre for Emerging Technologies Solutions with more 2.5 years of work experience in consulting, go-to-market activities and pre-sales for emerging technologies like Conversational AI, Personalized Videos.





References

Fortune Business Insights. Speech and Voice Recognition Market Size, Share & Industry Analysis.

Retrieved from - <https://www.fortunebusinessinsights.com/industry-reports/speech-and-voice-recognition-market-101382>

Pew Research Center (July 2019). YouTube Research.

Retrieved from - <https://www.pewresearch.org/internet/2019/07/25/popular-youtube-channels-produced-a-vast-amount-of-content-much-of-it-in-languages-other-than-english/>

Microsoft (December 2023). Exploring CPU vs GPU Speed in AI Training.

Retrieved from - <https://techcommunity.microsoft.com/t5/azure-high-performance-computing/exploring-cpu-vs-gpu-speed-in-ai-training-a-demonstration-with/ba-p/4014242>

Statista (January 2024). Voice technology - statistics & facts

Retrieved from - <https://www.statista.com/topics/6760/voice-technology/>

For more information, contact askus@infosys.com

Infosys[®]
Navigate your next

© 2025 Infosys Limited, Bengaluru, India. All Rights Reserved. Infosys believes the information in this document is accurate as of its publication date; such information is subject to change without notice. Infosys acknowledges the proprietary rights of other companies to the trademarks, product names and such other intellectual property rights mentioned in this document. Except as expressly permitted, neither this documentation nor any part of it may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, printing, photocopying, recording or otherwise, without the prior permission of Infosys Limited and/ or any named intellectual property rights holders under this document.