VIEW POINT

SPEECH AI TRENDS SHAPING THE NEW WORLD OF INTERACTIONS

Abstract

The speech AI industry is rapidly transforming, driven by advancements in natural language processing (NLP), machine learning, and other emerging technologies. These advancements have significantly enhanced the accuracy and contextual understanding of speech recognition systems. In the first part of Speech AI - Recent Tech Advancements, we covered technological advancements in Speech AI and their applications in industries. This part will provide a point of view on the emerging trends seen due to these technological advancements in Speech AI.

The global voice recognition market, valued at approximately USD 12.62 billion in 2023, is expected to show an annual growth rate (CAGR 2024-2030) of 14.24%, resulting in a market volume of US$15.87 billion by 2030. This growth is fueled by the increasing adoption of voice-enabled devices across various domains including retail, healthcare, automotive, consumer electronics, and smart home automation systems.

Infosys®
Navigate your next

Emerging Trends

We have identified the following major trends, which reflect the latest developments in the Speech AI space.

Trend 1: Hyper Personalized Conversational Experience

Imagine a voice assistant that remembers your coffee order, favorite news source, and even your running pace. By leveraging data like past interactions, purchase history, and even calendar events, AI can tailor conversations to your unique needs. This level of personalization fosters a sense of familiarity and builds trust, making speech AI not just a tool but a companion that anticipates your needs and enhances your daily life.

Infosys Cortex brings a Generative AI-enabled conversational assistant to contact center agents and customers, to provide hyper-personalized service by accessing past and current interactions.

Google Gems allows customization in Gemini assistant to have varying personalities and to support seamless integration with other Google services to provide more contextual and data-driven responses.

Voice cloning is taking personalization to the next level by enabling conversations with cloned personas. By generating synthetic voices that are virtually indistinguishable from human voices, it's fostering personalized interactions in a variety of contexts.

Microsoft claims its VALL-E model can clone voice from a three-second audio clip.

Trend 2: Local Language and Domain-Specific AI Models Become Popular

The rise of local language and domain-specific models is shifting the focus away from English-centric speech AI. The digital divide is narrowing as more people are accessing the internet in their native languages. There's a surge in demand for speech AI that understands regional dialects and nuances. Secondly, industries like healthcare or finance require specialized vocabularies. Domain-specific models trained on relevant data can handle complex jargon and technical terms, leading to more accurate and secure interactions. This rise in local and domain-specific speech AI ensures technology becomes more inclusive and caters to the specific needs of diverse users and industries.

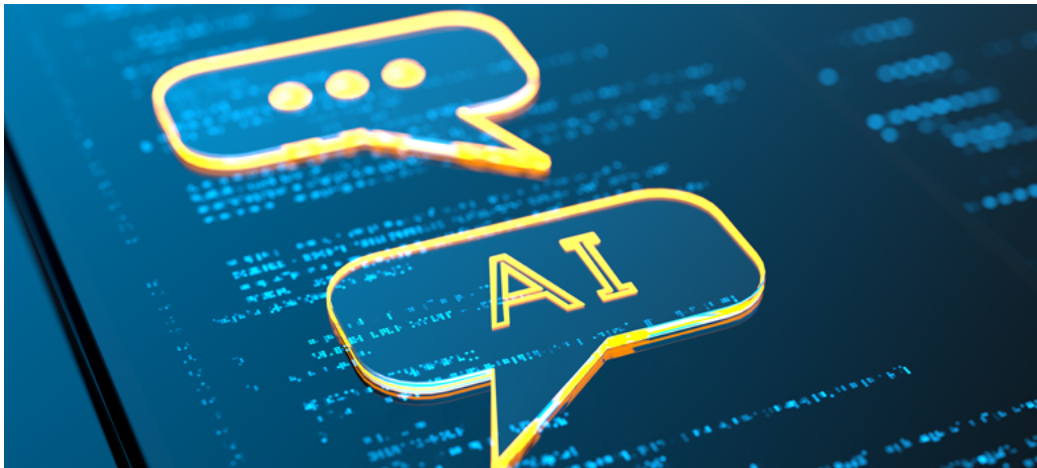
BHASHINI, launched by the Government of India, aims to transcend language barriers, ensuring that every Indian citizen can effortlessly access digital services in their language.

Microsoft acquired Nuance Communications Inc., a pioneer in speech recognition and artificial intelligence technology used in industries including healthcare, in 2022 for $19.7 billion to expand its reach into these industries.

A great example of domain-specific implementation is in **Bioacoustics AI**, a new wave of speech AI that focuses on deciphering the language of the animal kingdom.

Google has introduced a groundbreaking bioacoustics model that can identify eight distinct whale species based solely on their vocalizations, including even the recently discovered "biotwang" sound of the elusive Bryde's whale. This advancement paves the way for a deeper understanding of whale communication and behavior.

Elephant Acoustics Project uses Speech AI on Asian elephant sounds (Rumble, roar, trumpet, and chirp) in the forests of Assam, a hotspot of human-elephant conflicts. AI helps to quickly interpret the vocal sounds of captive and wild elephants to reduce such conflicts. It also can separate trumpet calls based on whether the elephants are interacting with their mahouts or with other elephants. AI can figure out a hundred such calls in five minutes, while earlier, listening in to these calls, a researcher could take up to 30 minutes to decode the message.



Trend 3: Real-Time Productivity Improvement

Imagine a world where productivity thrives on the power of your voice, with a hands-free experience. Gone are the days of navigating through menus or typing out commands. With a simple voice prompt, you can delegate tasks, manage your schedule, and access information instantly. Need to research a competitor while writing a proposal? Speak your query, and relevant data pops up on your screen. Need to schedule a meeting with a colleague across time zones? Voice AI can find a mutually convenient slot and update everyone's calendars as well as automatically capture the meeting notes and actions. This real-time support streamlines workflows minimizes distractions and empowers you to focus on the high-impact aspects of your work, ultimately boosting your overall productivity.

Infosys Cortex extracts and converts microdata from contact center customer interactions into insights for real-time action which are delivered through proactive assistance on agent desktop.

BMW has launched Intelligent Personal Assistant to enhance the in-car experience with support for over 20 languages. It enables voice control of various car functions, adapts to the driver's emotional state, and improves safety and convenience through hands-free interaction.

By 2028, 90% of vehicles worldwide are projected to have voice assistants.

Struggling to find time for academic reading? Try Google Illuminate. It converts academic papers into podcast-style audio discussions, perfect for listening while you're on the go!

Trend 4: Privacy and Availability by Design with On-Device Speech AI Companion

The future of voice interaction lies in on-device Speech AI companions that prioritize privacy and availability. These companions process information directly on your device, eliminating the need for a constant internet connection. This ensures your voice data stays private, never leaving your device for analysis. Imagine seamlessly interacting with your AI companion while traveling or in remote locations, with all core AI functionalities still available.
There's an increasing demand for technology that prioritizes privacy from the outset, especially with stricter regulations in place like the European Union AI Act.

Apple is introducing Apple Intelligence, that combines the power of generative models with personal context, working in a completely private and secure way on apple devices.

With Local Voice Control feature, Amazon Echo can be used to control smart home devices even when Echo isn't connected to a internet.

Trend 5: Interactive Learning Assistants

Speech AI can simulate emotionally aware human-like interactions, allowing you to practice and refine your responses in a safe, controlled environment. Additionally, Speech AI can be a powerful tool for improvement. By analyzing your interactions, Speech AI can identify areas for growth. It can point out repetitive phrases, suggest more engaging conversational openers, and even flag overly critical language.

Imagine rehearsing a job interview or a difficult conversation with an AI companion in a simulated environment, receiving real-time feedback too. Speech AI can be a valuable coach, helping you become a more articulate, empathetic, and effective communicator.

Infosys Cortex simulated learning powered by Generative AI & Speech AI helps to train contact center agents on behavioral and conversational aspects by letting AI play the role of a customer during conversation.

Language learning apps like Duolingo employ speech AI to help users practice pronunciation and conversation in different languages.

Khan Academy has launched Khanmigo powered by OpenAI models to provide on-demand AI-powered support for education which can provide audio explanations too.

OpenAI during the launch of GPT-4o in 2024 showcased a voice interactive demo of solving a math problem by collaborating with Sal Khan, founder of Khan Academy

Trend 6: Bridging the Language Divide through Speech Neutralization

Speech AI is poised to bridge the language divide through a powerful tool called "Speech Neutralization." This technology goes beyond simple translation. It analyzes the content and intent behind spoken words, then neutralizes accents and speech patterns, essentially creating a universally understandable version of the message. Imagine attending a multilingual conference where everyone speaks in a neutral, clear voice, removing the barrier of accents and allowing ideas to flow freely. Speech Neutralization can empower real-time conversations across cultures, fostering collaboration and understanding in a globalized world. However, ethical considerations regarding cultural identity and preserving linguistic diversity remain important aspects of this evolving technology.

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

Samsung has introduced Galaxy AI in Galaxy S24, which has a feature that allows you to converse freely with people who don't speak your language as well as to translate live calls to your language.

Trend 7: Secure Access with Voice-Based Identity Verification

The future of authentication is looking for increasingly voice-activated, thanks to advancements in Speech AI and voice biometrics. Unlike passwords, which are vulnerable to hacking and easily forgotten, voice biometrics leverage unique vocal characteristics like pitch, intonation, and even speaking patterns to create a personal and un-replicable "voiceprint." Imagine seamlessly accessing your bank account or logging into your work computer with a simple spoken phrase, which is a more secure and frictionless authentication experience.

As more and more households are using smart-home devices in their homes, voice shopping becomes integral to these connected environments. Customers can instruct their smart-home devices to add a product to a cart and place orders using their voice. As voice commerce is becoming popular, more attention is being paid to implementing strict security measures in place to safeguard user data and transactions. Biometric authentication and advanced voice recognition technologies are some useful technologies to ensure the privacy and security of users' personal and financial data.

By combining voice biometrics with other biometric factors like facial, retina or fingerprint recognition, we can create a more robust and secure multi-factor authentication system.

Dutch bank ING offering voice support in mobile banking application, Capital One partnering with Amazon Alexa to conduct voice-activated banking, Barclays using Siri to allow payments using voice commands are recent examples of Speech AI in banking.

30 percent of US adults checked their account balances using a smart device as per Forrester's "The State of Digital Banking, 2022" report.

American Express in 2023 became the first card issuer to offer facial and fingerprint recognition to prevent the fraudulent use of Amex Cards. It won't be surprising if speech AI is incorporated as an extra layer of defense soon.

Trend 8: Conversation Mining with Actionable Insights

Speech AI is transforming interactions by enabling conversation mining, a process that extracts actionable insights from a goldmine of voice data. Speech AI can pinpoint specific keywords, analyze sentiment through voice inflections, and categorize conversations by topic.

Infosys Cortex provides analytics on contact center conversations by deriving values, KPIs, and insights using the power of Generative AI and Speech AI. This empowers businesses to understand customer needs and pain points at an unprecedented level. The actionable insights can be used to improve customer service processes, personalize product offerings, and even train agents to address customer concerns more effectively.

Emotion recognition using speech AI is an emerging field with great potential for improving mental health monitoring. Speech AI can analyze speech patterns to detect emotional stress. By examining factors like volume, pitch, and pauses, AI models can identify emotional cues, potentially aiding mental health monitoring. This system can be applied to various cases such as early detection of mental health, personalized treatment plans, real-time support, and therapy augmentation.

Ellipsis Health uses speech AI on recorded speech to identify signs of depression and anxiety in the senior population. They are also evaluating the feasibility and accessibility of Speech AI-enabled distress screening mobile applications for adolescents and young adults diagnosed with cancer.

The Future: Addressing the Challenges

The advent of AI-powered conversational agents is also leading to **cultural shifts** like the normalization of machine-human interaction with a shift in communication patterns from traditional politeness with humans to direct and informal with machines, a deeper emotional connection and relationship with AI systems and the evolution of spoken language with AI-specific terms and jargons, e.g. Alexa, Siri, etc. are perceived as part of vocabulary now. However, excessive reliance on technology can have detrimental effects.

A study published in the journal Archives of Disease in Childhood in 2022 shows long-term negative effects on children's social and cognitive development of voice assistants

OpenAI highlighted risk of emotional attachment to its GPT-4o human-like voice.

A study in 2023 reveals that using algorithmic responses (e.g. smart replies), are perceived as less cooperative, less affiliative, and more dominant.

Deepfakes and misinformation too pose a significant threat. Malicious individuals can use voice cloning to create highly convincing fake audio recordings, spread disinformation, frame innocent people, cause social unrest, or commit financial fraud.

Audio deepfakes have emerged as a weapon of choice in election disinformation.

An "artificially generated" voice in the likeness of US President Joe Biden was reported to robocall voters encouraging them not to vote in New Hampshire's 2024 presidential primary.

A deepfake of UK opposition leader Sir Keir Starmer allegedly berating a staffer got released before the 2023 general elections.

A Journalist in 2023 broke into his Lloyds bank account voice biometrics by using an AI-synthesized clone of his voice .
The Deloitte Center for Financial Services expects $23 billion losses by 2030 due to frauds related to AI cloning.

Voice biometrics, while valuable for security, can be vulnerable to misuse through voice cloning.

Like other AI technologies, speech AI also raises responsible usage concerns related to bias, fairness, privacy, security, transparency, inclusivity, etc.

Addressing these challenges is critical for the responsible development, deployment, and usage of speech AI technology.
It requires a collaborative effort from researchers, developers, policymakers, and society.

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

Google Research (September 2024). Recognizing whale vocalizations with AI.
Retrieved from - https://research.google/blog/whistles-songs-boings-and-biotwangs-recognizing-whale-vocalizations-with-ai/

Nature India (June 2024) . Interpreting the call of the wild with AI.
Retrieved from - https://www.nature.com/articles/d44151-024-00096-6

Automotive World (September 2019). Digital voice assistants are the future of in-vehicle control.
Retrieved from - https://www.automotiveworld.com/articles/digital-voice-assistants-are-the-future-of-in-vehicle-control/

Forrester (February 2022). The State Of Digital Banking.
Retrieved from - https://www.forrester.com/report/the-state-of-digital-banking-2022/RES177049?ref_search=2881951_1644591383698

BMJ Journals (August 2022). Effects of smart voice control devices on children: current challenges and future perspectives.
Retrieved from - https://adc.bmj.com/content/early/2022/08/22/archdischild-2022-323888

Nature Scientific Reports (April 2023). Artificial intelligence in communication impacts language and social relationships.
Retrieved from - https://www.nature.com/articles/s41598-023-30938-9

Deloitte (July 2023). Using biometrics to fight back against rising synthetic identity fraud.
Retrieved from - https://www2.deloitte.com/us/en/insights/industry/financial-services/financial-services-industry-predictions/2023/financial-institutions-synthetic-identity-fraud.html

Statista (March 2024). Speech Recognition – Worldwide.
Retrieved from - https://www.statista.com/outlook/tmo/artificial-intelligence/computer-vision/speech-recognition/worldwide

For more information, contact askus@infosys.com

Infosys®
Navigate your next