VIEW POINT

VOICE INTERFACES

Abstract

A voice-user interface (VUI) makes human interaction with computers possible through a voice/speech platform in order to initiate an automated service or process. This Point of View explores the reasons behind the rise of voice interface, key challenges enterprises face in voice interface adoption and the solution to these.

Infosys®
Navigate your next

Are We Ready for Voice Interfaces?

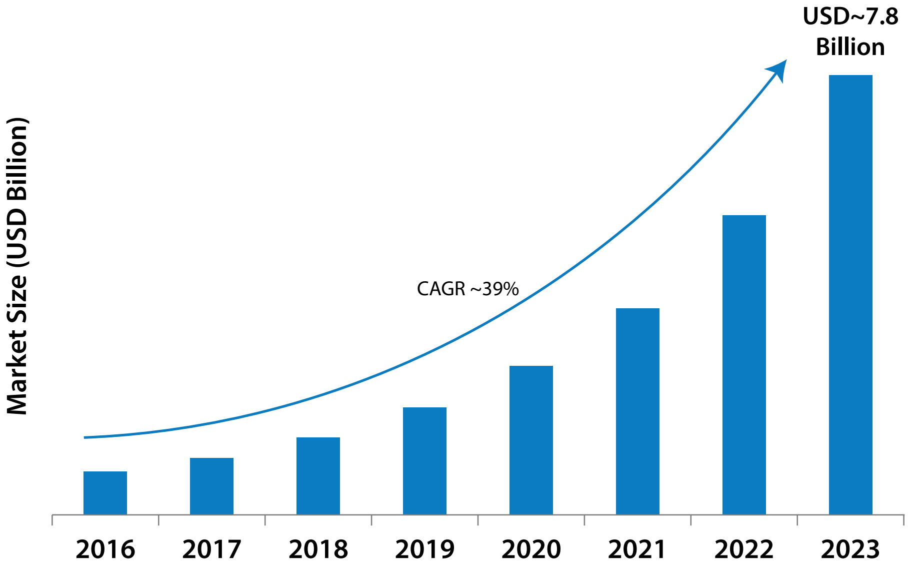
Let's get talking!

Since Apple integration with Siri, voice interfaces has significantly progressed. Echo and Google Home have demonstrated that we do not need a user interface to talk to computers and have opened-up a new channel for communication. Recent demos of voice based personal assistance at Google IO showed the new promise of voice interfaces.

Almost all the big players (Google, Apple, Microsoft) have 'office productivity' applications that are being adopted by businesses (Microsoft and their Office Suite already have a big advantage here, but things like Google Docs and Keynote are sneaking in), they have also started offering integrations with their Voice Assistants.

As per industry forecasts, over the next decade, 8 out of every 10 people in the world will own a device (a smartphone or some kind of assistant) which will support voice based conversations in native language. Just imagine the impact!

Voice Assistant Market



The Sudden Interest in Voice Interfaces

Although voice technology/assistants have been around in some shape or form for many years, the relentless growth of low-cost computational power—and breakthroughs in machine learning and deep learning—mean bots can now be built with self-learning capabilities.

Big breakthroughs in AI/ML space in the recent years have improved the voice recognition accuracy, adoption has also increased as it is convenient to use voice interfaces, as a result a lot of interest and investments are attracted.

Voice Recognition Accuracy

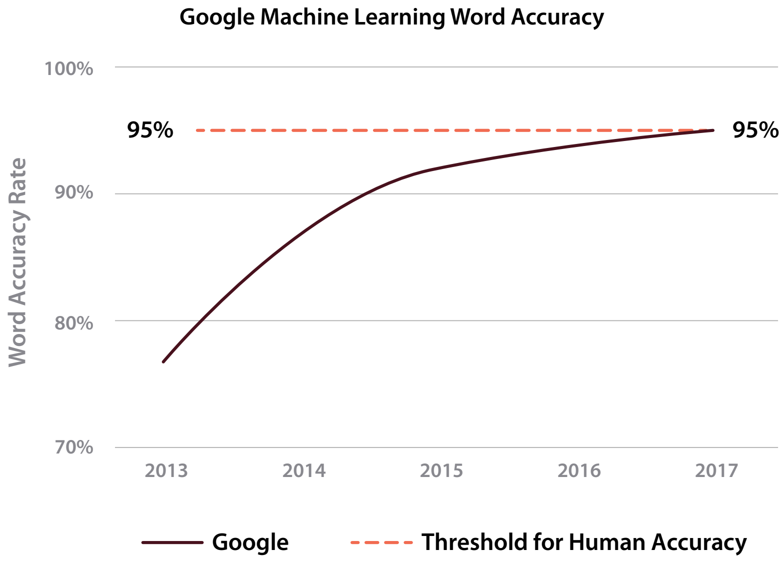
Voice Recognition accuracy continues to improve as we now have the capability to train the models using neural networks and large amount of relevant user data. Google has been able to achieve 95% machine learning word accuracy which is the same as human accuracy. It is also expected to receive 50% of all searches in voice by 2020 which will further improve recognition across languages as it will give access to variety of new data for training. A lot of recent advancements in the NLP (Natural Language Processing), NLG (Natural Language Generation) and Text to Speech(TTS) areas have also given a big push to voice based applications.

Convenience – speaking vs typing

Humans can speak 150 words per minute vs the typing speed of 40 words per minute. This makes voice interfaces an efficient way to interact with technology. With natural language support and ability to respond intelligently to user queries based upon user's context, behavior patterns and AI based learnings, voice based systems are increasingly becoming an exciting option for users as is evident from the ever increasing sales numbers of voice assistants (Amazon Echo, Google Home etc.).

Industry investments

Tech giants have entered the business of Voice Interfaces through either direct investments or acquisitions. For example, Google recently acquired Mobvoi (founded by Li Zhifei, a former Google Research Scientist) while Apple acquired Siri. Amazon's Bezos has openly said that they have more than 1,000 people working on Alexa and the Echo ecosystem for the past many years. Bezos, through his investment fund Bezos Expeditions, and the Alexa Fund have invested in a $2.5 million seed round for Pulse Labs led by the Seattle-based venture firm Madrona Venture Group.



Key Technologies Enabling Voice Interfaces

Currently there are four key enablers for voice interfaces:

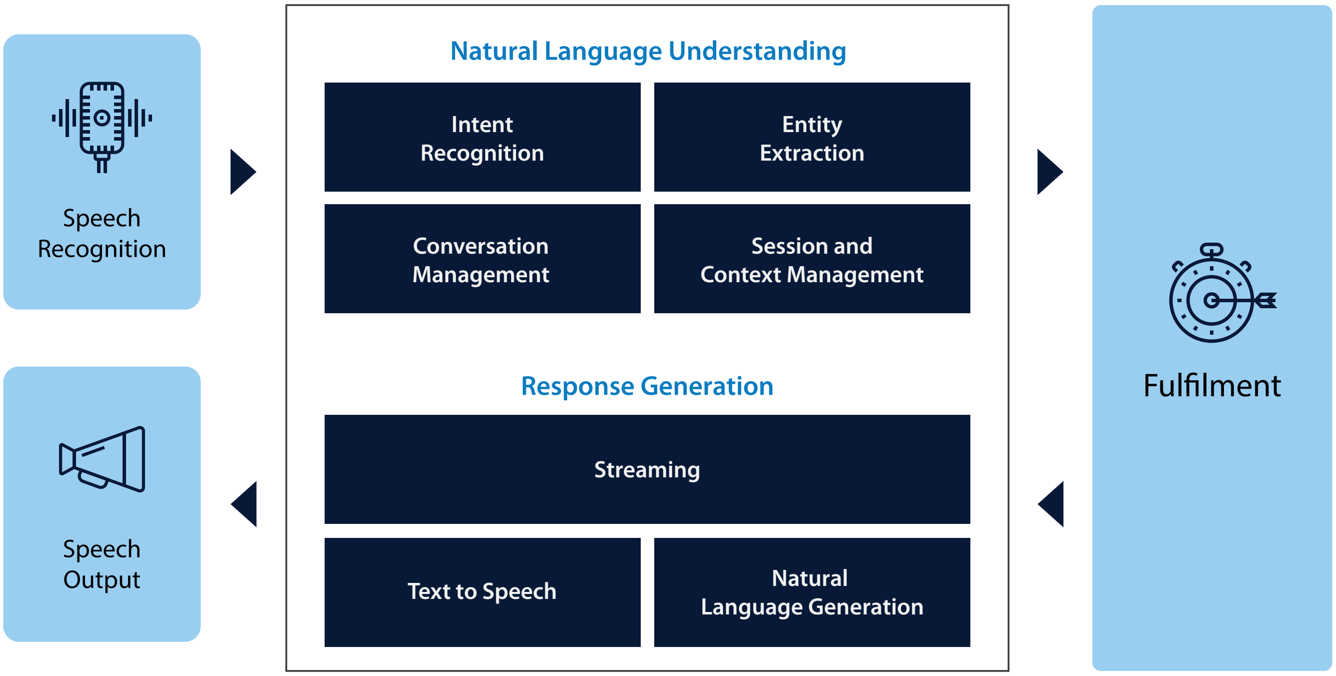
Text to Speech (TTS)

Automatic Speech Recognition (ASR)

Natural Language Understanding

Natural Language Generation

Each of these technologies have seen big breakthroughs in recent years providing enterprises with a multitude of solutions/options. Enterprises can also embrace either online (cloud services) or offline (on-premise) option depending on their business/use case requirements



Speech Recognition

Natural Language Understanding

Intent Recognition

Entity Extraction

Conversation Management

Session and Context Management

Response Generation

Streaming

Text to Speech

Natural Language Generation

Fulfilment

Speech Output

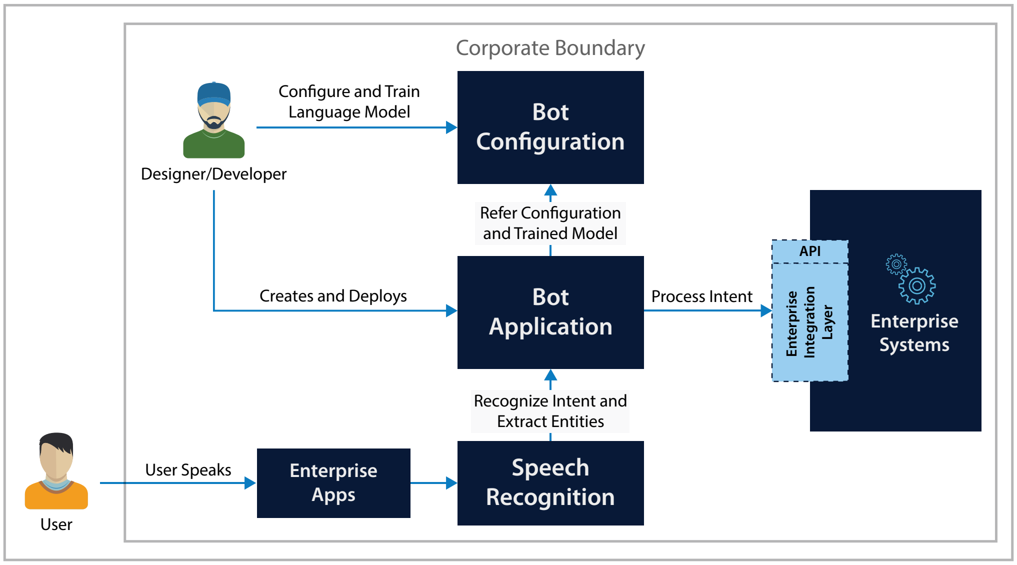
Key players and options for enterprises

The vendor landscape for Voice Interfaces has become quite crowded with the IT giants of the world having the majority share of the pie. Here is a quick look.

On Premise Deployment

- When speech recognition is brought on-premise, the issue of routing users and business data via third-party/cloud based services is taken care of
- Licensing cost, cost of maintaining/ scaling the infrastructure will vary depending upon the nature of use case/ volume of transactions
- While open-source on-premise options for speech recognition are evolving/ improving fast, there are commercial options available like Nuance, NICE etc.

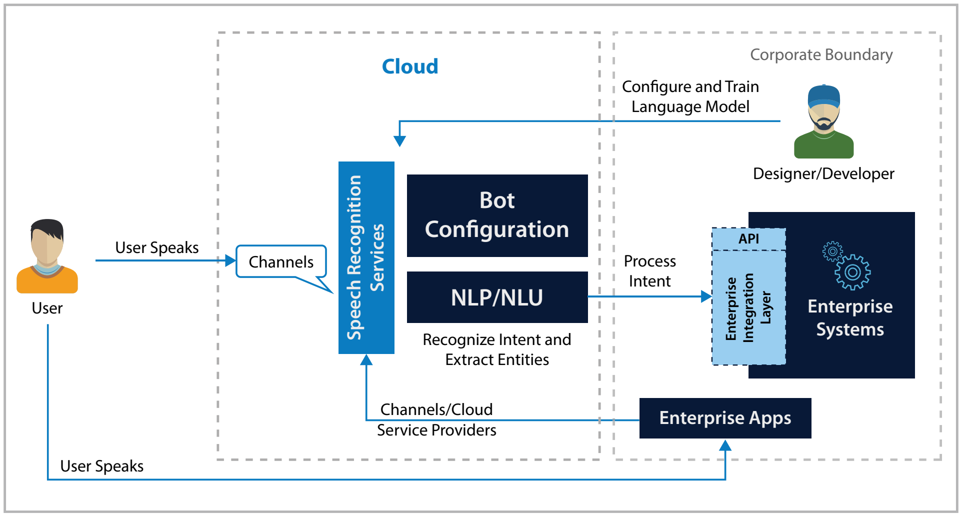
On-premise Deployment



Cloud Deployment

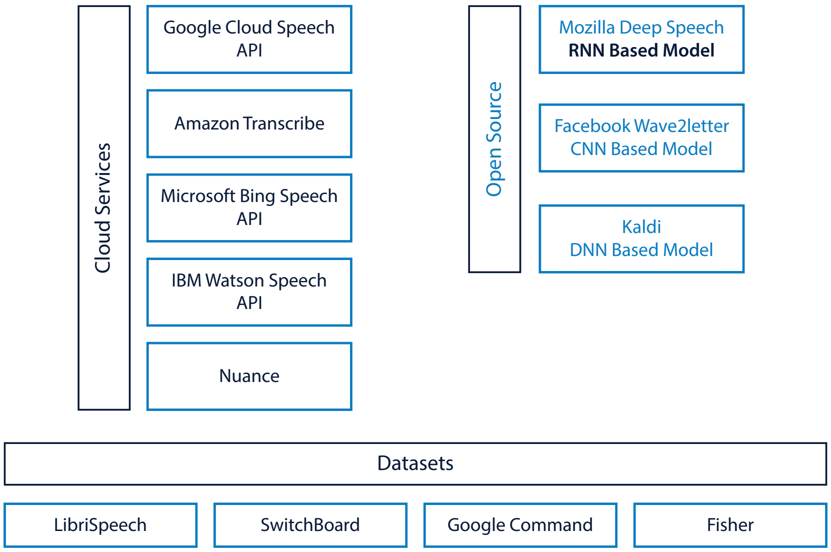
- Cloud service providers offer a promising option by allowing an enterprise to configure and deploy voice bots on cloud (server-less as lambda/cloud functions or server infrastructure) which can integrate with the enterprise systems
- There are advantages of pay as you use, ease of on demand scale, no infrastructure/platform maintenance overhead, inbuilt analytics etc.
- Leading online speech service providers are the giants like Google, Amazon and Microsoft. While Google provides ML powered Google Speech API for Speech to text conversion that is available for short or long-form audio with high accuracy recognition, Amazon has Transcribe and Microsoft comes with Microsoft Cognitive Services and Bing Speech API.

Cloud Deployment


Cloud
Corporate Boundary
Configure and Train Language Model
Designer/Developer
User
User Speaks
Channels
Speech Recognition Services
Bot Configuration
NLP/NLU
Recognize Intent and Extract Entities
Process Intent
API
Enterprise Integration Layer
Enterprise Systems
Channels/Cloud Service Providers
Enterprise Apps
User Speaks

The following diagram shows the various players in on-premise, cloud and data set services

Speech Recognition: A view of Online, Open Source and Dataset options



Enterprise Readiness for Voice as Their Strategic Priority

Redesigning/enhancing digital experience is by far the top strategic priority for most organizations while conversational AI with voice recognition capability is becoming key for businesses to enhance customers experience and attract new customers.

Challenges

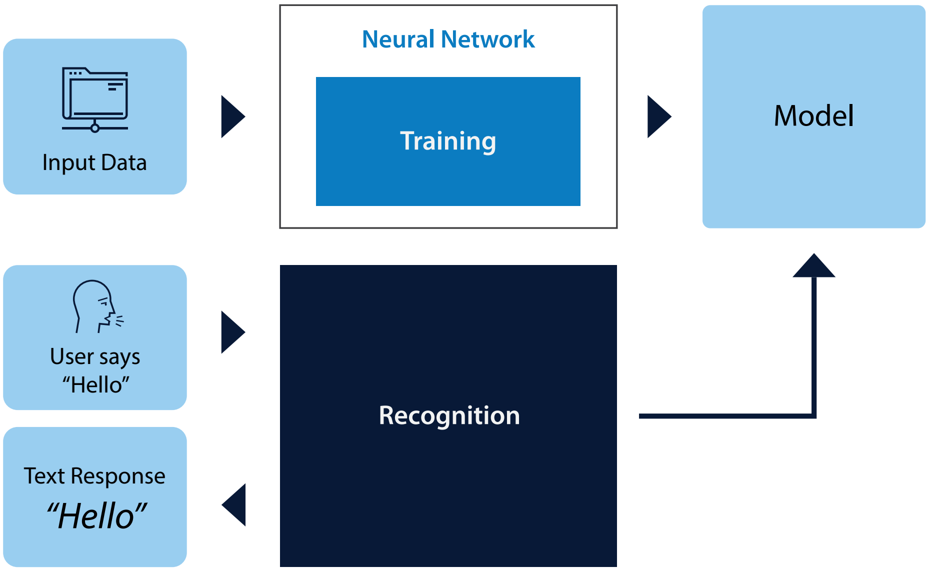
Speech Recognition owing to its complexity has taken years of effort to evolve to this level of maturity. There are multiple challenges with users' real life audio streams including low quality microphones, background noise, reverb and echo, accent variations etc. Training data should contain samples with all these problems in order to make sure the neural network can deal with them. Other challenges that enterprises face include:

Data set for training

Accuracy of machine learning solutions is completely driven by the amount of available data sets for training. As for the voice recognition accuracy, it is difficult for enterprises to create their own approach and solution because of their inability to gather variety and volume of data required to build the right models.

IT giants like Google, Amazon or Baidu do possess huge datasets with hours and hours of audio recordings in real-life situations. It is the single biggest factor that separates their world-class speech recognition system from any other speech recognition system created elsewhere. Alexa, Google Home, Apple Home Pod, Siri, Google Now are providing convenience to the customer at minimal one-time or no cost but most importantly they are getting real life users' audio data for further training to improve their experience. This drives enterprises to depend on such cloud players in developing voice interfaces for enterprise applications.

Speech Recognition



Lack of mature open source options

While there is a dependency on hosted players like Google, Amazon etc. for building voice interfaces, open source stack for voice processing is an alternate option. There are some promising beginnings in open source world which have the potential to reach a good level of maturity. Mozilla has launched Project Common Voice, an initiative to help make voice recognition open to everyone.

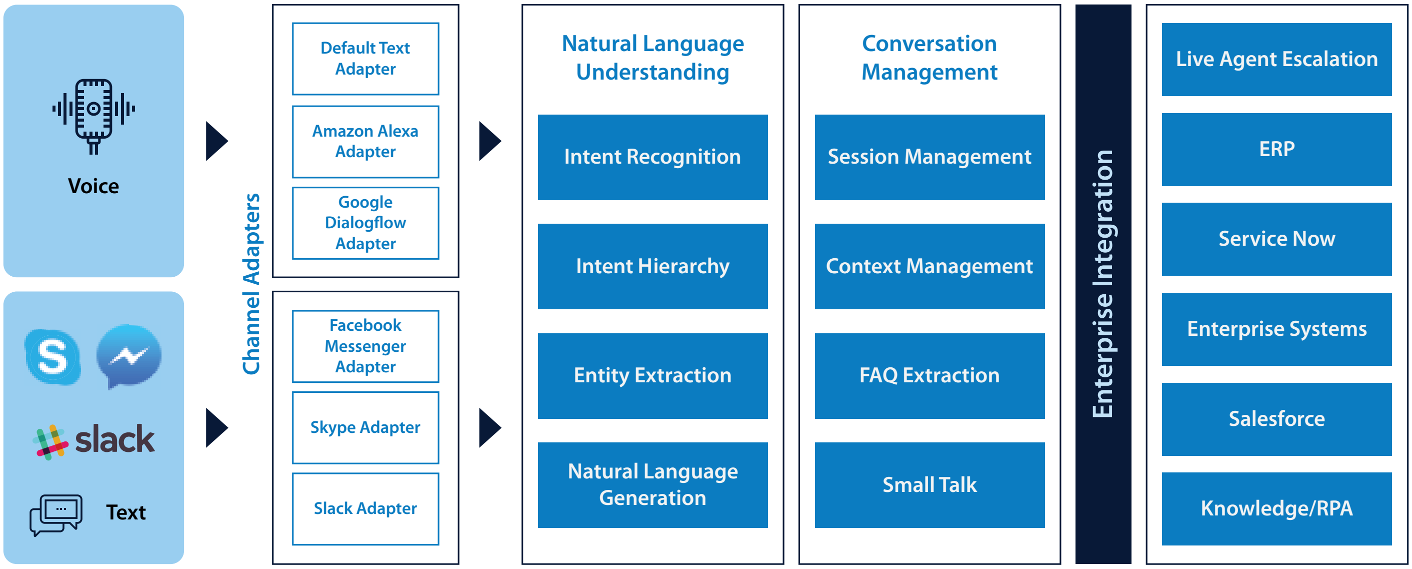
Data privacy concerns with cloud platforms

Businesses have their own inhibitions about user's interactions (involving users' data and other critical business information) being routed via cloud providers for voice recognition. Some industries especially banking and finance prefer equivalent/mature open source and on premise options in order to protect their customers' data within their own datacenters.

Enterprise Readiness for Voice as Their Strategic Priority

Voice interfaces being critical for our customers we lay special emphasis on a focused approach to overcome the various challenges they face. Enabling voice support is one of the key feature in Infosys Chatbot Roadmap. As a systems integrator, we engage with our customers in early stages, help them design thinking and enable them on the best solution in their context. We believe that, when it comes to voice interfaces, our customers do need a comprehensive approach with solid small steps instead of jumping on to some visibly exciting option.

Infosys NIA Chatbot Platform



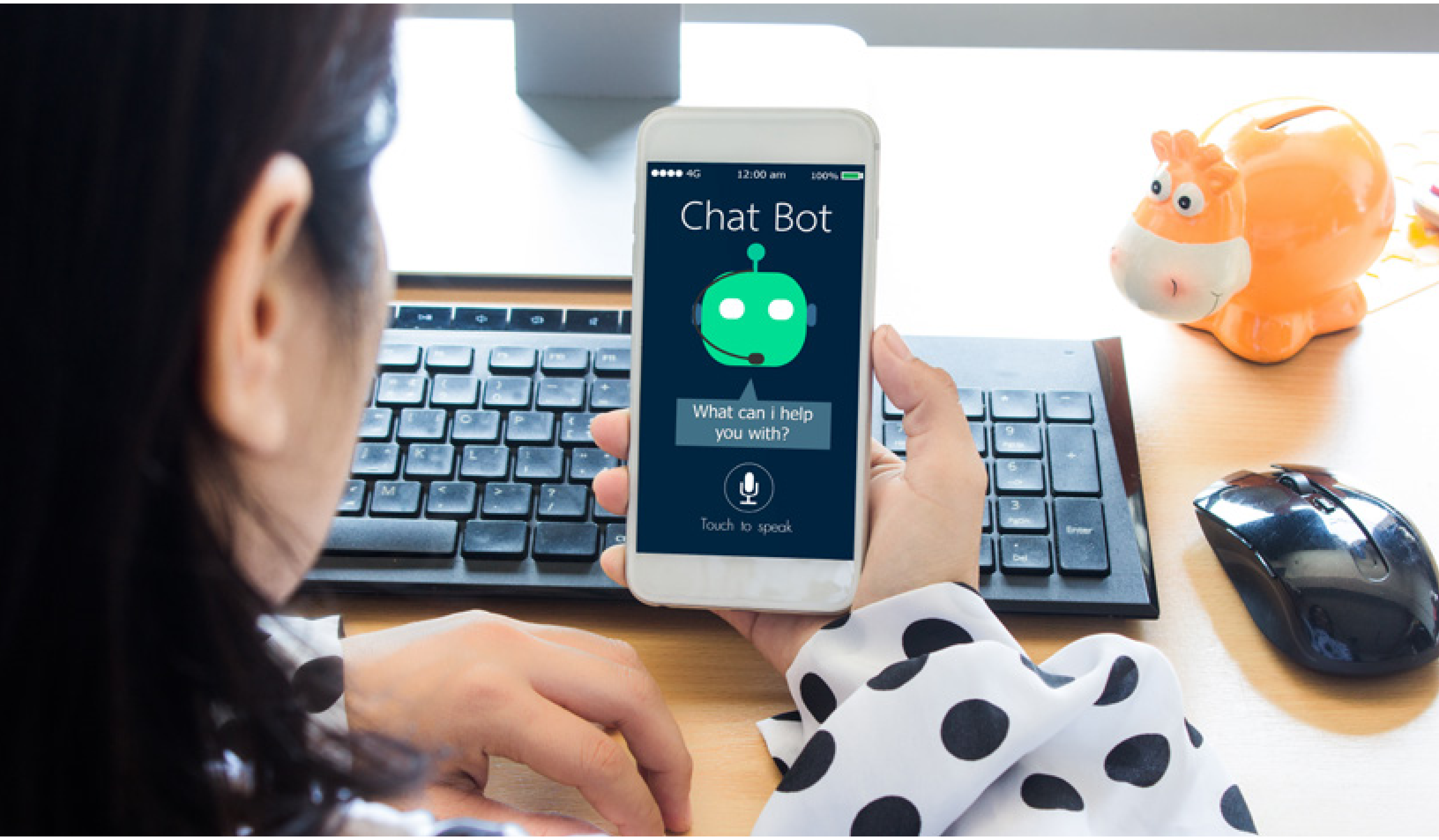
We have made early inroads in this capability. Infosys Chatbot is architected to leverage the best of the breed technologies to create conversational channels like Skype, Facebook messenger, Slack etc. while voice is one of the channels which can be realized through popular last mile solutions like Amazon Echo, Google Home, Siri, Google Assistant, etc. Chatbot architecture is focused on enterprise integration, conversation management and addressing the non-functional requirements related to voice support while being modularized to support the preferred channel of voice.

Platform provides customized integrations with leading commercial speech recognition options like NICE and Nuance Dragon, which claim to be the world's bestselling speech recognition software

Summary

Conversation as an interface is the most natural way for humans to interact with technology.

Technology is evolving rapidly in order to allow people to interact with technology as naturally as they interact with each other. These technology advances will definitely bring more and more meaningful improvements in people's experience in interactions with voice interfaces. Businesses are seeking to leverage this and rapidly extend reach across segments and improve convenience of use.



Reference:

- https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html
- https://www.nasdaq.com/article/voice-control-platforms-a-key-technology-trend-in-2017-cm745944
- https://www.kleinerperkins.com/perspectives/internet-trends-report-2018
- https://www.recode.net/2018/2/1/16958432/amazon-alexa-pulse-labs-jeff-bezos-madrona-skills-voice-apps-google
- https://github.com/mozilla/DeepSpeech
- https://cloud.google.com/speech-to-text/

About the Author

Vishal Manchanda is a Principal Technology Architect with Infosys Center for Emerging Technology Solutions and has over 19 years of experience in the IT Industry. Vishal is involved in developing, architecting and incubating various IT solutions in the area of personalized intelligent interfaces involving hyper-contextual personalized videos and engineering of the platform for conversational interfaces. Vishal works towards understanding how enterprises can adopt voice interfaces given a plethora of new voice channels and big cloud players in the fray.

For more information, contact askus@infosys.com

Infosys®
Navigate your next